Statistics You Need to Know for Machine Learning

Slides and Notes

Statistics You Need to Know for Machine Learning was developed by Sharad Saxena, Manoj Singh, and Tarek Elnaccash. Additional contributions were made by Dee McKoy, Kirstin Morrison, Catherine Truxillo, Andy Ravenna, and Linda Sheppard. Instructional design was provided by Julia Johnson.

Statistics You Need to Know for Machine Learning Slides and Notes
Content based on SAS® Viya® Release LTS 2025.03

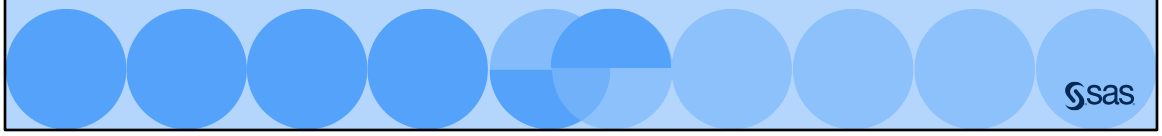
Course code EVST151, prepared date November 04, 2025.
http://support.sas.com/training/

For information about other courses in the curriculum, contact the SAS Education Division at 1-800-333-7660, or send e-mail to training@sas.com. You can also find this information on the web at http://support.sas.com/training/ as well as in the Training Course Catalog.

For a list of SAS books (including e-books) that relate to the topics covered in this course notes, visit https://www.sas.com/sas/books.html or call 1-800-727-0025. US customers receive free shipping to US addresses.

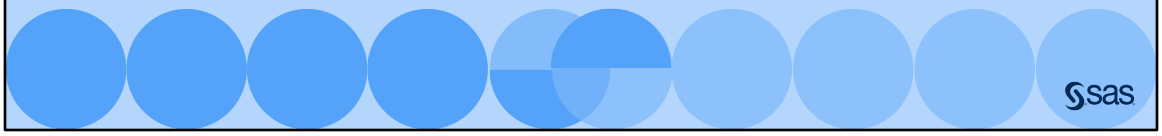
Welcome to the Course

To explore this learning resource in greater detail, please visit the course page on learn.sas.com. There, you'll find comprehensive course materials, tutorials, and updates to support your learning journey.



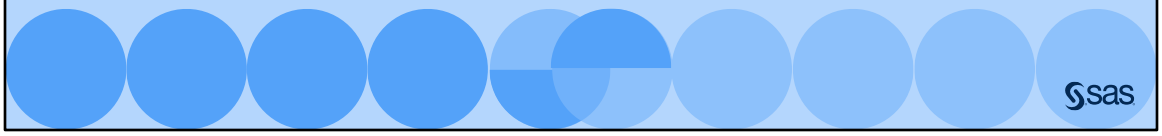
Meet Your Instructors

To explore this learning resource in greater detail, please visit the course page on learn.sas.com. There, you'll find comprehensive course materials, tutorials, and updates to support your learning journey.



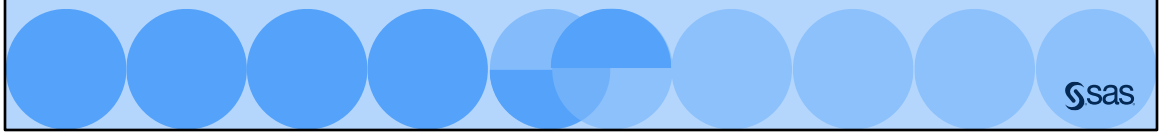
Practicing in This Course (REQUIRED)

To explore this learning resource in greater detail, please visit the course page on learn.sas.com. There, you'll find comprehensive course materials, tutorials, and updates to support your learning journey.



Data Dictionary

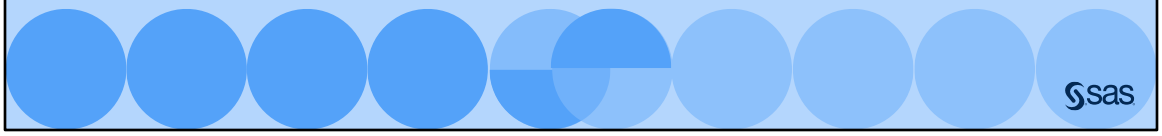
To explore this learning resource in greater detail, please visit the course page on learn.sas.com. There, you'll find comprehensive course materials, tutorials, and updates to support your learning journey.



Lesson 1: Statistics and Machine Learning

Lesson Overview

To explore this learning resource in greater detail, please visit the course page on learn.sas.com. There, you'll find comprehensive course materials, tutorials, and updates to support your learning journey.



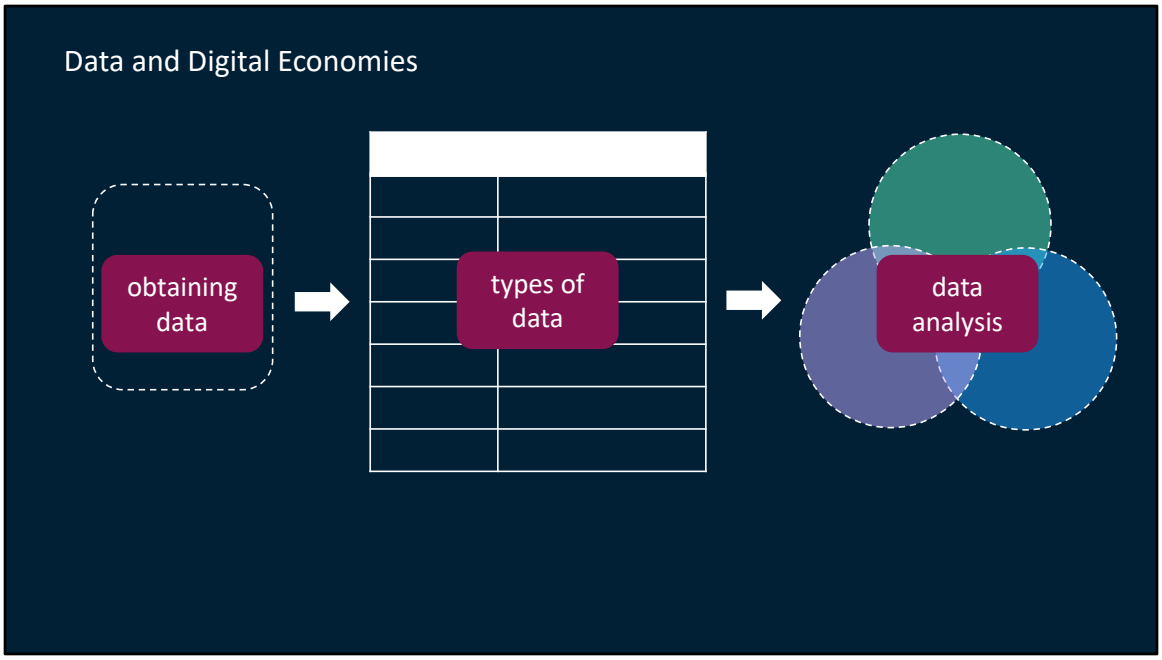
1.1 Relevance of Statistics in Big Data and Machine Learning

Data and Digital Economies

> Statistical thinking will one day be as necessary for efficient citizenship as the ability to read and write.

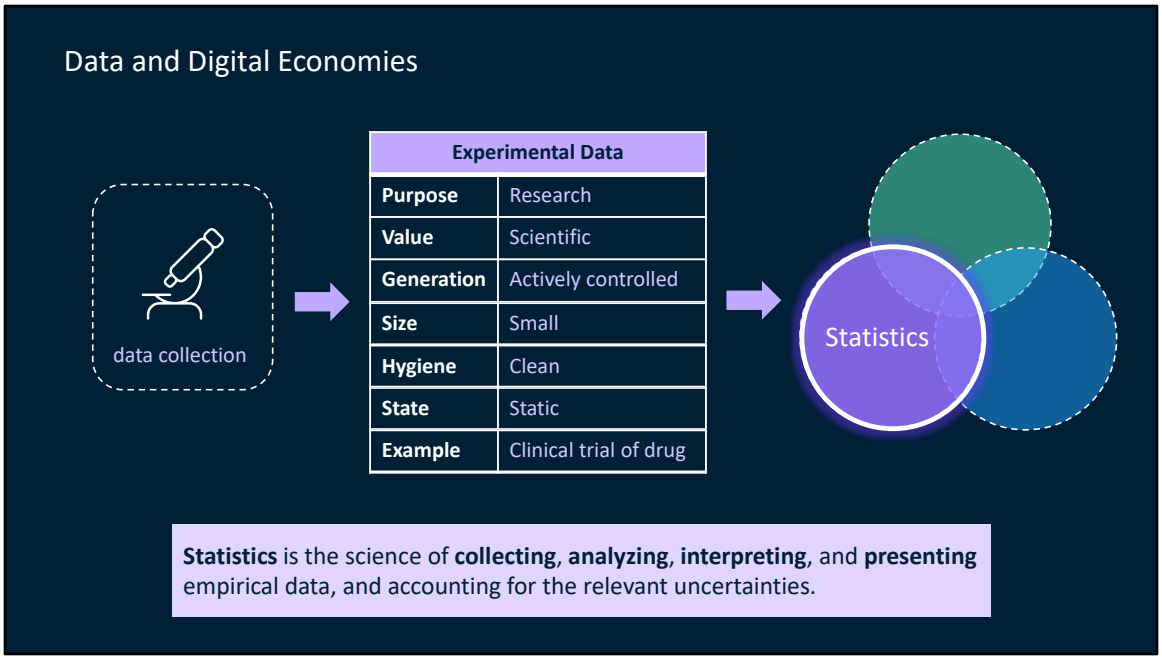
– Samuel S Wilks (1951) *paraphrased from* Herbert G Wells (1903)

Statistical thinking will one day be as necessary for efficient citizenship as the ability to read and write! This is the quotation from the presidential address in 1951 of mathematical statistician Samuel S. Wilks to the American Statistical Association. Wilks was paraphrasing Herbert G. Wells from his book "Mankind in the Making“ published in 1903.



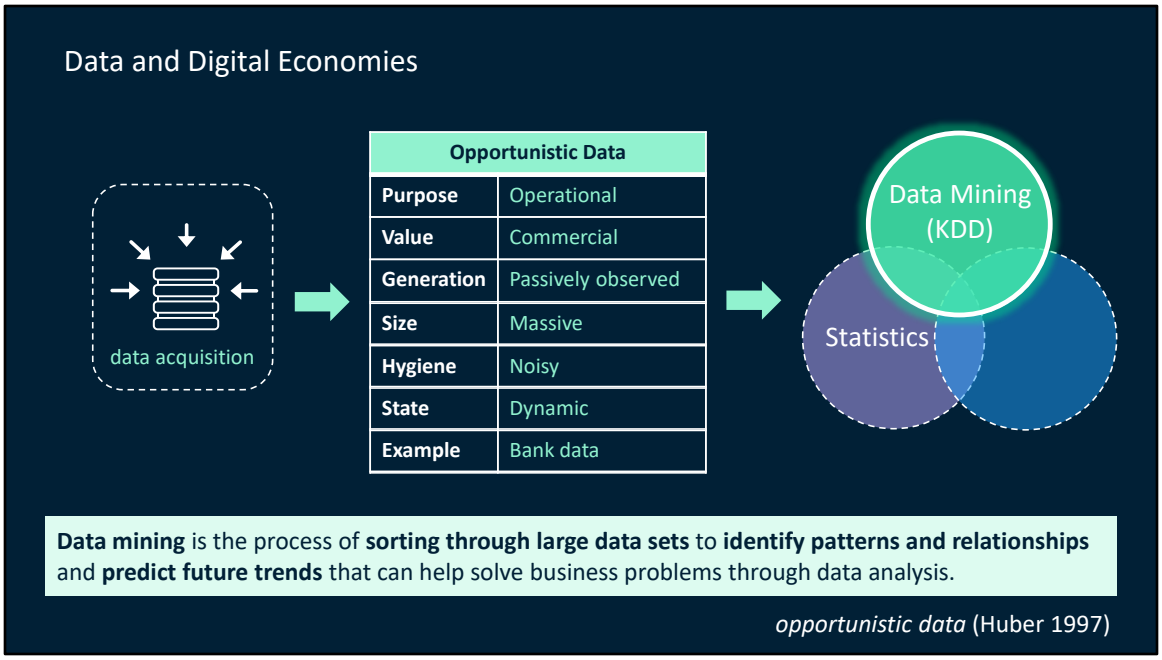
So, what's the relevance of statistics in the big data machine learning world? Before taking on this question, it's helpful to understand data and digital economies. Data are the foundation of everything you do.

First, let's look at the ways we obtain data, the different types of data, and how each type of data is analyzed.



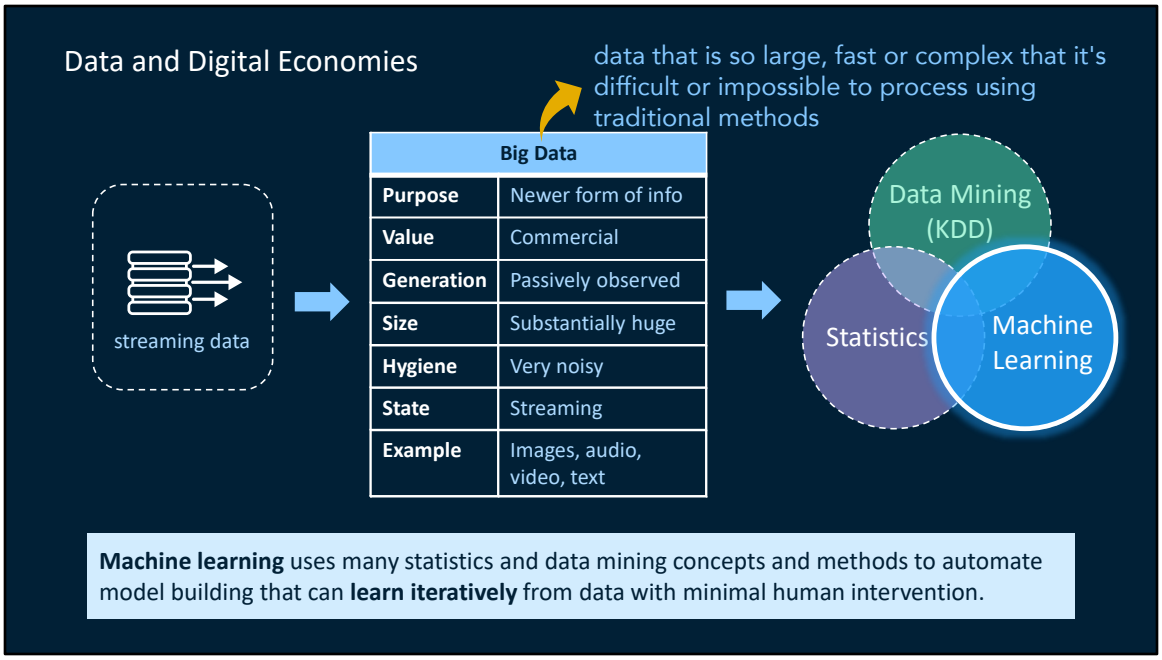
Historically, most data were generated or collected for research purposes. Therefore, their value is scientific. This type of data is commonly called *experimental data*. Experimental data are actively controlled, so they are clean, small, and static. Examples of experimental data include a clinical trial of a drug and the effect of fertilizers on agriculture yield.

Such data analysis is performed using statistics. *Statistics* is the science of collecting, analyzing, interpreting, and presenting empirical data, and accounting for the relevant uncertainties. Traditional statistics relied heavily on small samples and rigid assumptions about data and its distributions. Large and complex data sets were rare - and those few that did exist were rarely analyzed.



Over time, businesses accumulated massive amounts of operational data. This operational data was not generated with data analysis in mind. It is aptly characterized as *opportunistic data*. This contrasts with experimental data where factors are controlled and varied in order to answer specific questions. This type of data are collected in a *database* - an organized collection of structured information, or data, typically stored electronically in a computer system.

Opportunistic data are passively observed and are noisy, massive, and dynamic. An example is data coming from bank operations. Such data is analyzed using *data mining*. Data mining is the process of sorting through large data sets to identify patterns and relationships and predict future trends that can help solve business problems through data analysis. *Knowledge discovery in databases*, abbreviated as KDD, is often used synonymously with data mining. *KDD* is a multidisciplinary research area that uses statistics, in the form of exploratory data analysis and predictive models, to reveal patterns and trends in very large data sets.



Now, newer forms of data are coming in. We've moved *away* from systems of inconvenient data acquisition - requiring days, weeks, or months - and we've moved *to* streaming data. *Streaming data* is data that is continuously generated by different sources, such as social media, censors, and cameras. Such data arrives at exceptional speed, in a variety of forms, and in humongous volumes. Everyone might have access but might not have the skills to make sense out of the data.

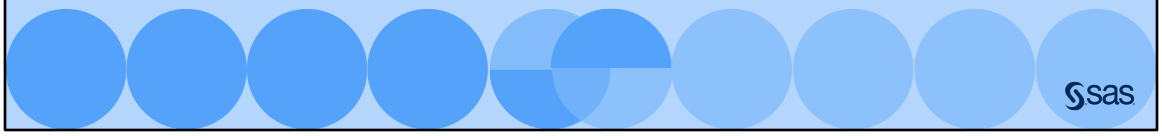
This large volume of data – both structured and unstructured – inundates a business on a day-to-day basis. Big data refers to data that is so large, fast, or complex that it's difficult or impossible to process using traditional methods.

To analyze this type of data, we must use and develop computer systems that can learn and adapt without following explicit instructions, by using algorithms and statistical models to analyze and draw inferences from patterns in data.

Machine learning uses many statistics and data mining concepts and methods to automate model building that can learn iteratively from data, identify patterns, and predict future results – with minimal human intervention.

What Is Big Data?

To explore this learning resource in greater detail, please visit the course page on learn.sas.com. There, you'll find comprehensive course materials, tutorials, and updates to support your learning journey.



Question: Big Data

Which of the following are examples of big data? (Select all that apply.)

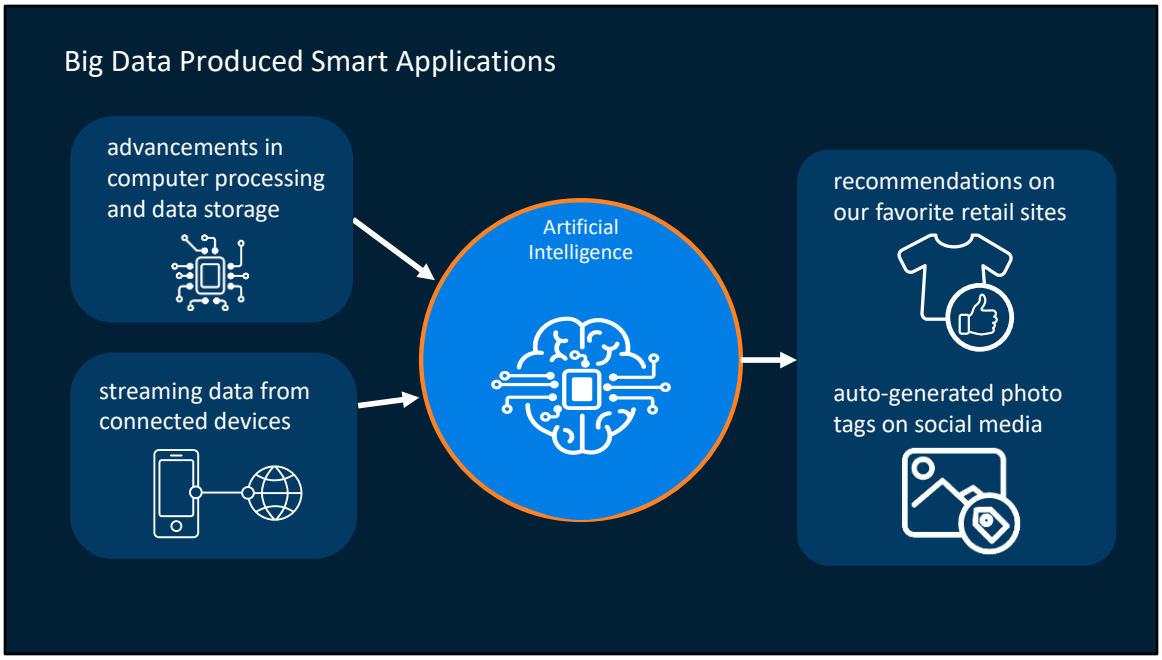
a. Sample of controlled clinical drug trials
b. Social media feeds
c. Feeds coming in from cameras installed on city roads
d. Taking two samples of the same plant and exposing one of them to sunlight, and the other is kept away from sunlight

Answer: Big Data

Which of the following are examples of big data? (Select all that apply.)

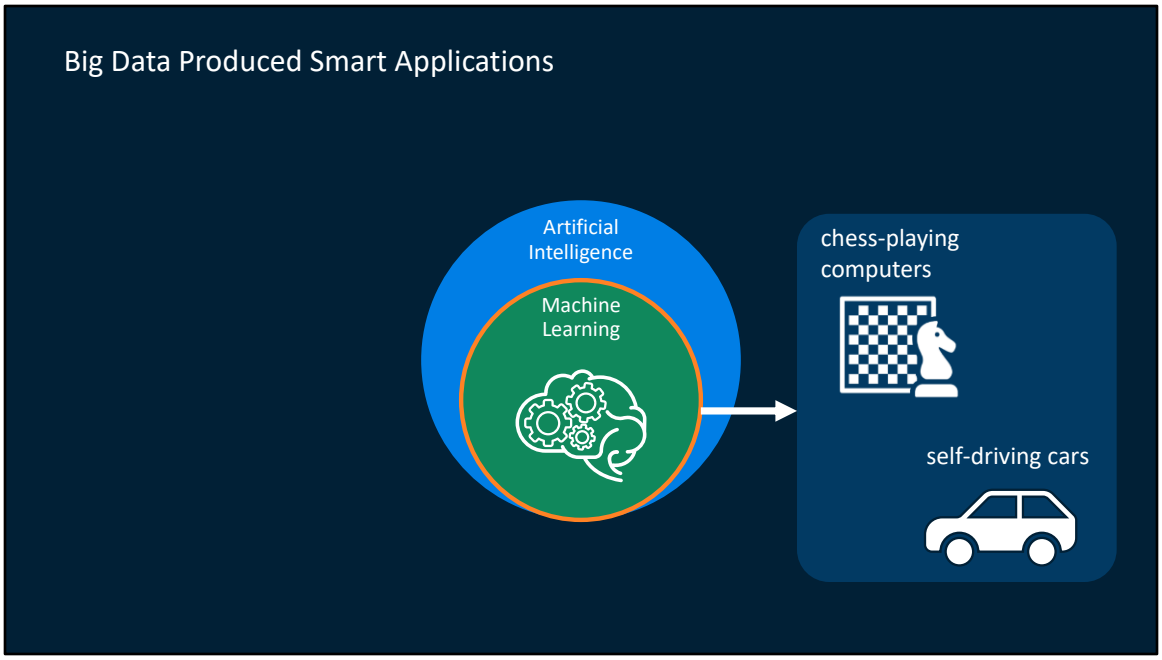
a. Sample of controlled clinical drug trials
b. Social media feeds
c. Feeds coming in from cameras installed on city roads
d. Taking two samples of the same plant and exposing one of them to sunlight, and the other is kept away from sunlight

Correct answer: b and c.
Note that **a** and **d** are examples of experimental data.

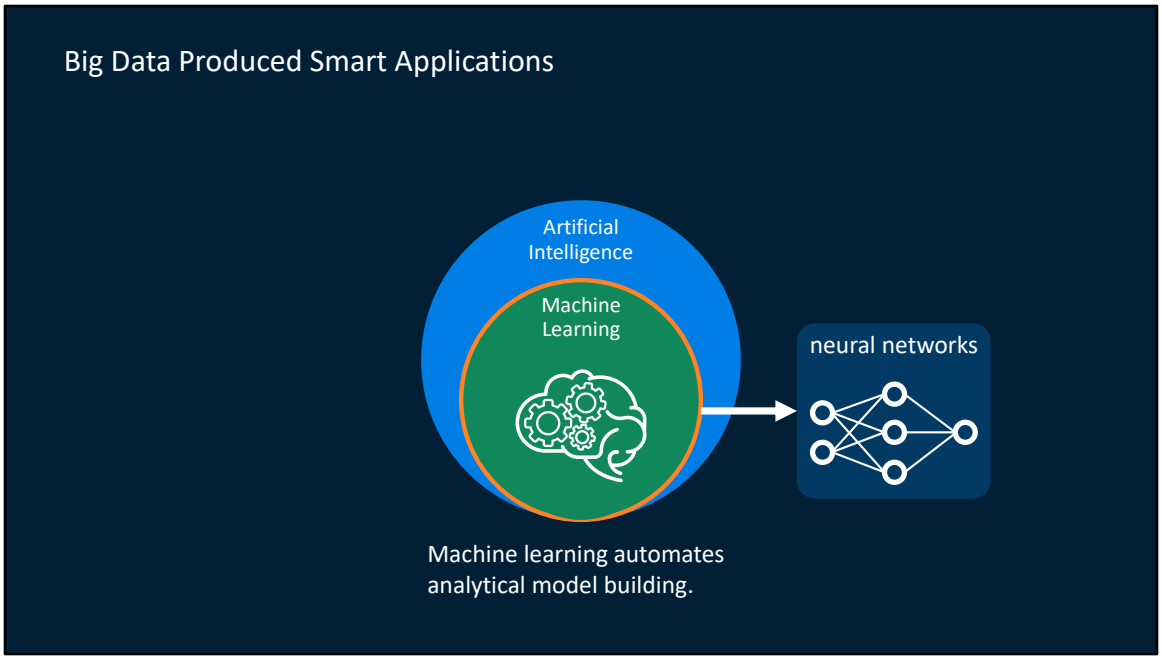


Advancements in computer processing and data storage made it possible to ingest and analyze more data than ever before. Around the same time, we started producing more and more data by connecting more devices and machines to the internet and streaming large amounts of data from those devices.

All these advancements brought artificial intelligence closer to its original goal of creating intelligent machines, which we're starting to see more and more in our everyday lives. From recommendations on our favorite retail sites to auto-generated photo tags on social media, many common online conveniences are powered by artificial intelligence.

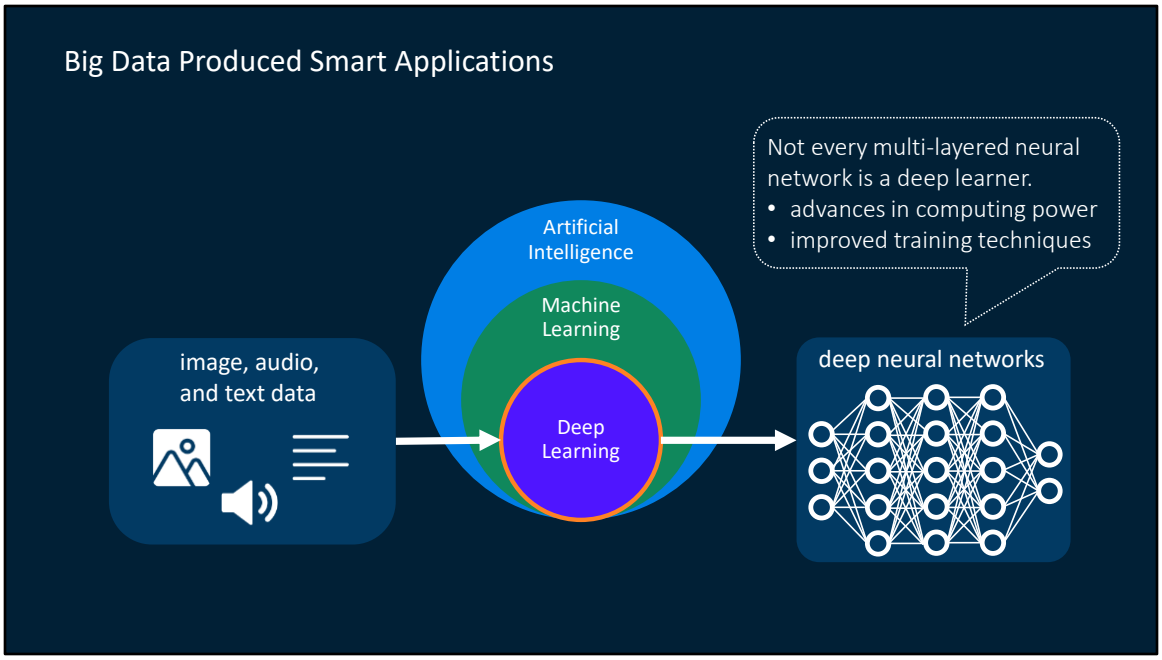


Most artificial intelligence examples that you hear about today – from chess-playing computers to self-driving cars – rely heavily on machine learning and deep learning.



Artificial intelligence is the broad science of mimicking human abilities, and machine learning is a specific subset of artificial intelligence that trains a machine how to learn. Machine learning automates analytical model building. It uses methods from neural networks, statistics, operations research, and physics to find hidden insights in data without being explicitly programmed where to look or what to conclude.

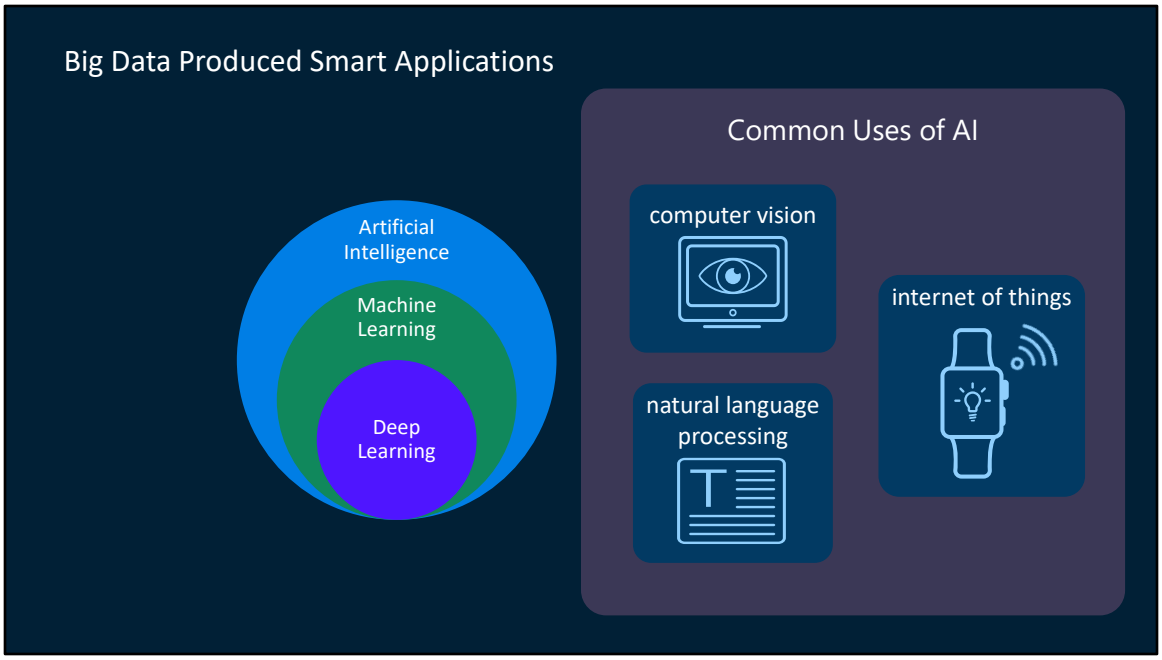
Early work with neural networks during 1950s-1970s stirred excitement for "thinking machines." A neural network is a kind of machine learning inspired by the workings of the human brain. It's a computing system made up of interconnected units (like neurons) that processes information by responding to external inputs, relaying information between each unit. The process requires multiple passes at the data to find connections and derive meaning from undefined data. Machine learning became popular in 1980s-2010s.



Machine learning and deep learning are subfields of artificial intelligence, and deep learning is a specific field of machine learning. While machine learning mostly dealt with huge tabular data, deep learning generally dealt with image, audio and text data that results in even larger tables.

Recent deep learning breakthroughs are driving an AI boom. *Deep learning* uses huge neural networks with many layers of processing units to learn complex patterns in large amounts of data.

Does this mean that every multi-layered neural network is a deep learner? No. Although the term deep learning refers to the numerous hidden layers used in a neural network, they take advantage of advances in computing power such as GPUs and improved training techniques such as better performing activation functions, randomly initialized weights, advanced regularization techniques, and efficient optimization algorithms.



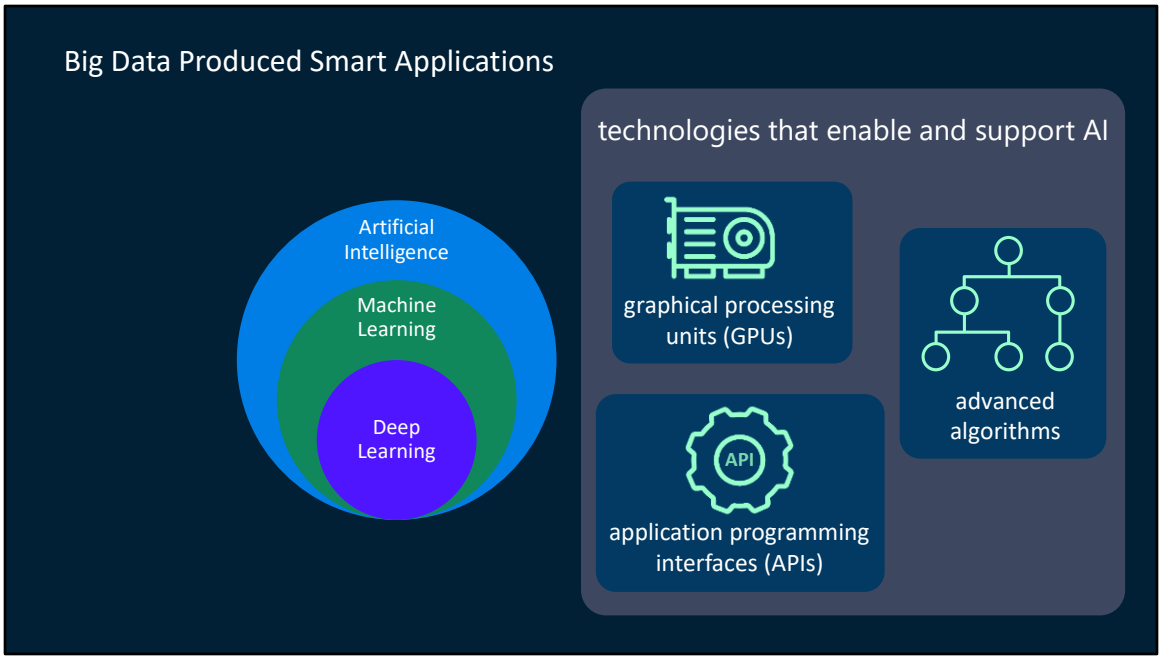
Computer vision is a field of artificial intelligence that trains computers to interpret and understand the visual world. Using digital images from cameras and videos and deep learning models, machines can accurately identify and classify objects — and then react to what they "see." When machines can process, analyze, and understand images, they can capture images or videos in real time and interpret their surroundings. From recognizing faces to processing the live action of a football game, computer vision rivals and surpasses human visual abilities in many areas.

Natural language processing, or *NLP*, is a branch of artificial intelligence that helps computers understand, interpret, and manipulate and generate human language, including speech.

NLP helps computers communicate with humans in their own language, making it possible for computers to read text, hear speech, interpret it, measure sentiment, and determine which parts are important. The next stage of NLP is natural language interaction, which allows humans to communicate with computers using normal, everyday language to perform tasks.

The Internet of Things, or *IoT*, refers to a vast number of "things" that are connected to the internet so that they can share data with other things – IoT applications, connected devices, industrial machines, and more. Internet-connected devices use built-in sensors to collect data and, in some cases, act on it. IoT connected devices and machines can improve how we work and live. Real-world Internet of Things examples range from a smart home that automatically adjusts heating and lighting to a smart factory that monitors industrial machines to look for problems, and then automatically adjusts to avoid failures. The Internet of Things generates massive amounts of data from connected devices, most of it unanalyzed. Automating models with AI will allow us to use more of it.

Machine learning is based on the idea that machines should be able to learn and adapt through experience, and AI refers to a broader idea where machines can execute tasks "smartly." Artificial Intelligence applies machine learning, deep learning, and other techniques to solve actual problems.



In addition, several technologies enable and support AI**.**

Graphical processing units, or *GPUs*, are key to AI because they provide the heavy compute power that's required for iterative processing. Training neural networks requires big data plus compute power.

Advanced algorithms are being developed and combined in new ways to analyze more data faster and at multiple levels. This intelligent processing is key to identifying and predicting rare events, understanding complex systems, and optimizing unique scenarios.

Application programming interfaces, or *APIs*, are portable packages of code that make it possible to add AI functionality to existing products and software packages. They can add image recognition capabilities to home security systems and Q&A capabilities that describe data, create captions and headlines, or call out interesting patterns and insights in data.

In summary, the goal of AI is to provide software that can reason on input and explain on output. AI will provide human-like interactions with software and offer decision support for specific tasks, but it's not a replacement for humans – and won't be anytime soon.

Question: Computer Vision

Which of the following is an example of computer vision?

a. Facial recognition-based security access in your phone
b. Smart watch connected to your phone
c. Siri or Google Assistant in your phone
d. Text autocorrect feature in your phone

Answer: Computer Vision

Which of the following is an example of computer vision?

a. Facial recognition-based security access in your phone
b. Smart watch connected to your phone
c. Siri or Google Assistant in your phone
d. Text autocorrect feature in your phone

Correct answer: a.
Note that **b** is an example of Internet of Things (IoT) and **c** and **d** are examples of natural language processing (NLP).

Knowing Statistics (1)

What do you think…. To do machine learning on big data, do you need to know statistics?

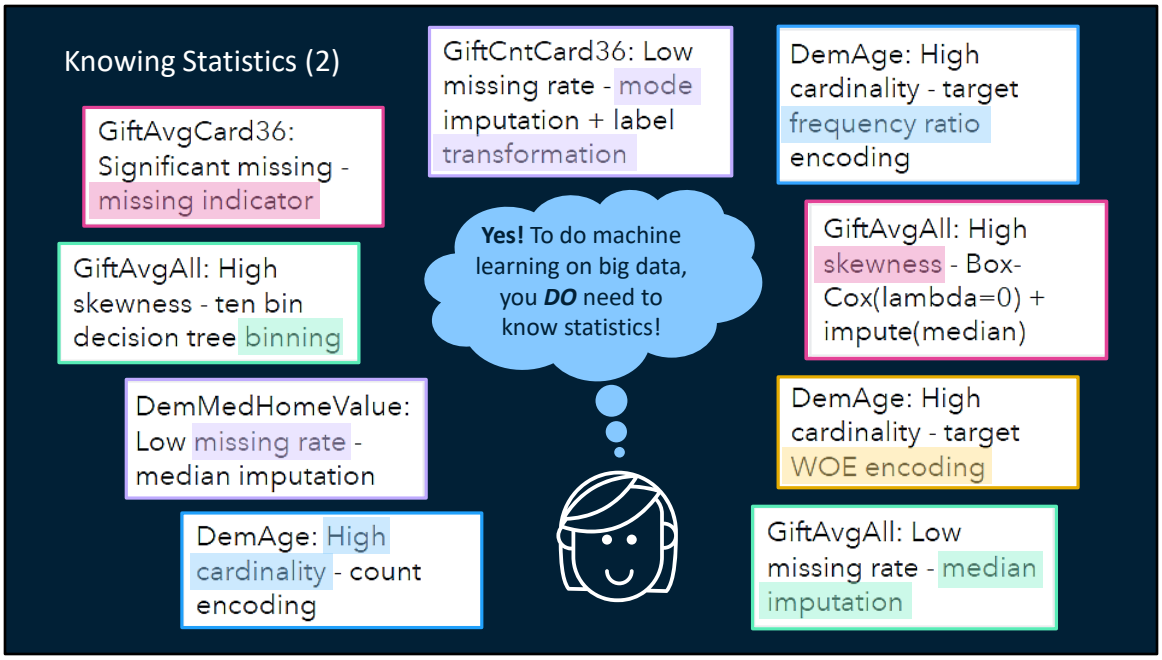
- ❑ No, not at all.
- ❑ Might be, quite unsure.
- ❑ Yes, definitely.

Let's explore running an automated machine learning task.

To do machine learning on big data, do you need to know statistics? At this point, you might not be sure. To answer this question, let's explore running an automated machine learning task.

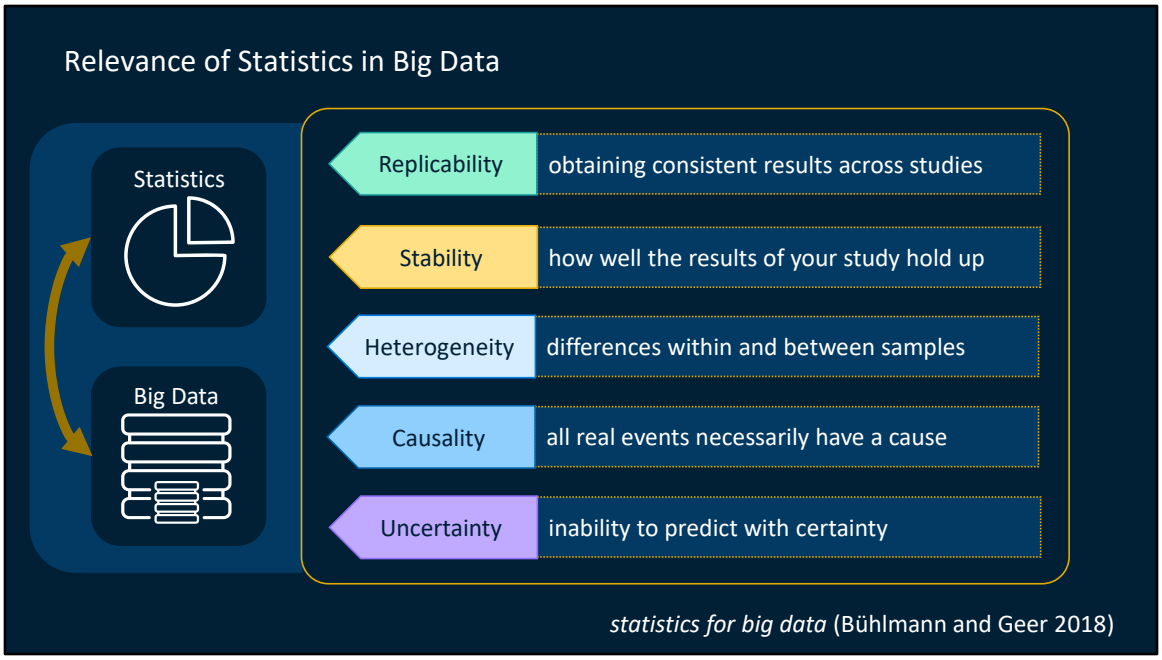
Demo: Automated Machine Learning in SAS Studio

To explore this learning resource in greater detail, please visit the course page on learn.sas.com. There, you'll find comprehensive course materials, tutorials, and updates to support your learning journey.



In the previous demonstration, even though the machine learning capability in the tool can fulfill your objective of creating new features automatically, you still need to have the knowledge of basic statistical concepts to understand and explain the output better. You encountered some of the statistical terms highlighted here, for example, missing indicator, mode, transformation, frequency ratio, binning, skewness, missing rate, WOE encoding, high cardinality, median imputation, and so on. All of these and more will be discussed in this course. Yes! To do machine learning on big data, you do need to know statistics!

Machine learning tools use **data-driven algorithms** and **statistical methods** to prepare and analyze data sets and then **draw inferences** from identified patterns or **make predictions** based on them.



Statistics remains highly relevant irrespective of the "bigness" of data. The role of statistics remains what it has always been, and is even more important now. Perhaps the core statistical task in (traditional) statistics is inductive inference from data to models and scientific conclusions. This core task is still very relevant in the advent of massive data sets.

Replicability, stability, heterogeneity, causality, and uncertainty are the five basic principles of statistics, and they all hold equally well with big data.

Replicability is obtaining consistent results across studies aimed at answering the same scientific question, each of which has obtained its own data.

Related to replicability is the notion of stability. Stability is how well the results of your study or experiment consistently hold up - in other words, the strength of support for your results subject to small perturbations in input data. It's a useful concept to address whether an alternative, “appropriate” analysis would generate similar knowledge. More specifically, it's a measure of how well you control for random errors in your study.

Heterogeneity is a vital concept that appears in various contexts, and its definition varies accordingly. Heterogeneity can indicate differences within individual samples, between samples, and between experimental results in a meta-analysis.

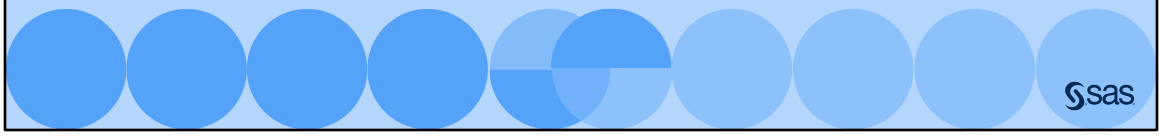
For a simple causation definition, statistics describes a relationship between two events or two variables. Causation is present when the value of one variable or event increases or decreases as a direct result of the presence or lack of another variable or event.

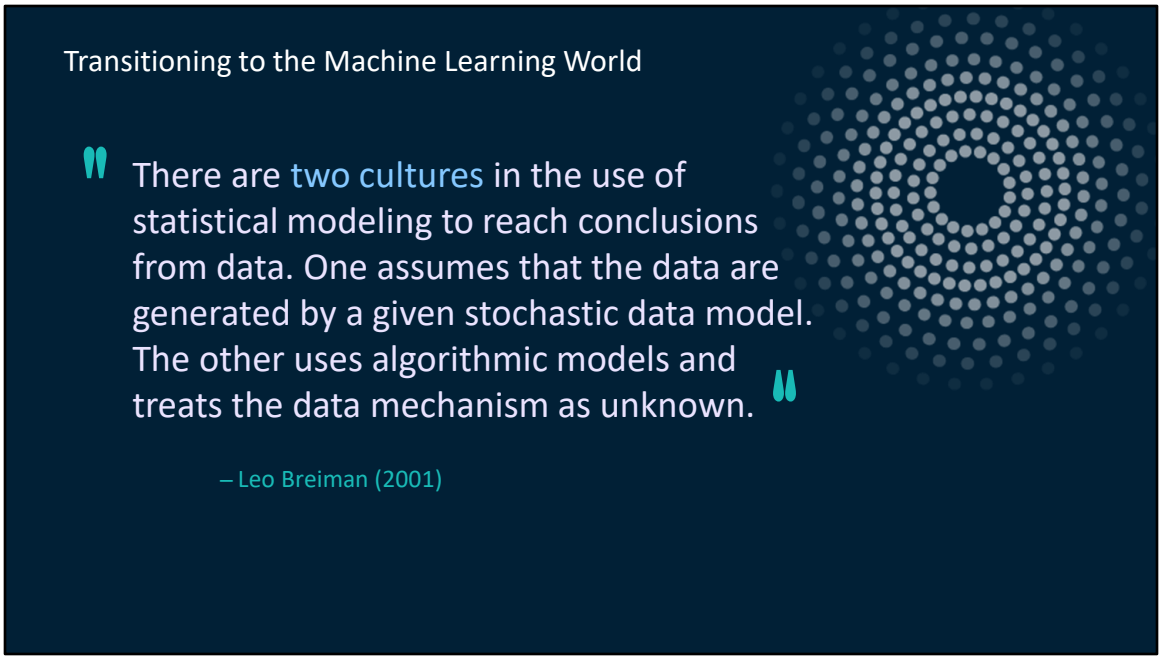
The uncertainty principle implies that it is in general not possible to predict the value of a quantity with arbitrary certainty, even if all initial conditions are specified.

Ideally, in big data scenario too, the conclusions and findings are ***replicable*** and generalizable. If you imagine running the analysis again, now on a new data set, would the outcome be similar, meaning that the model is ***stable***? How would you find out what similarity in outcomes means and how to evaluate accuracy, to quantify ***uncertainty***. Understanding ***heterogeneity*** in large-scale data sets is more crucial, and comprehending ***causality*** and its connection to robust prediction is still interesting.

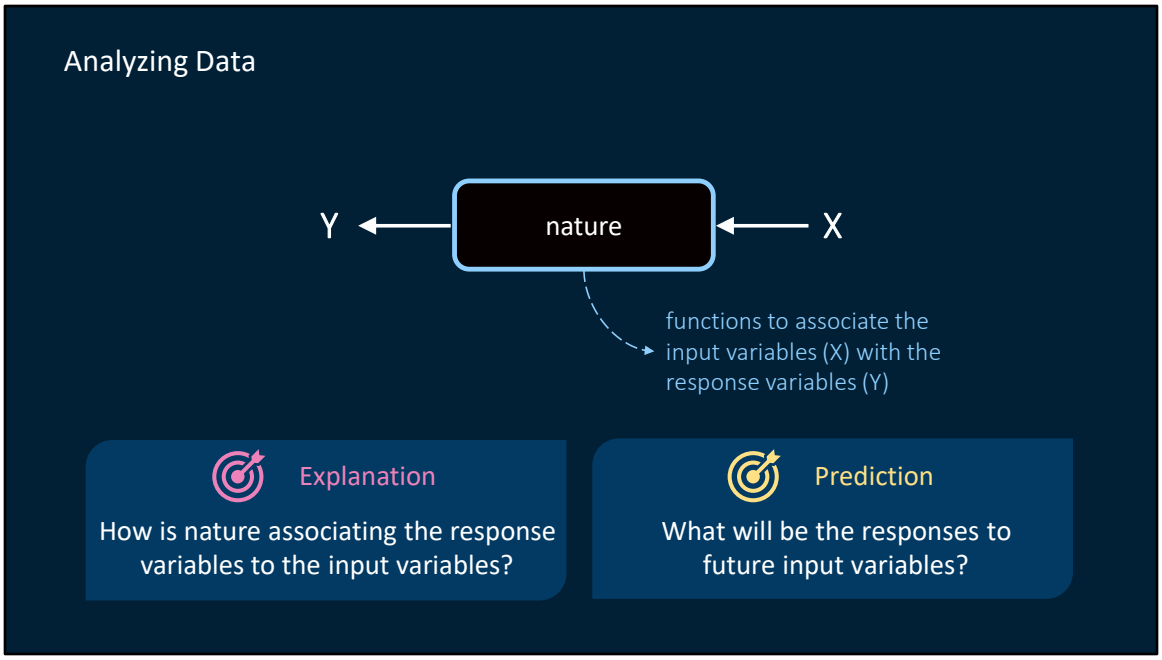
More about the Relevance of Statistics in Big Data

To explore this learning resource in greater detail, please visit the course page on learn.sas.com. There, you'll find comprehensive course materials, tutorials, and updates to support your learning journey.





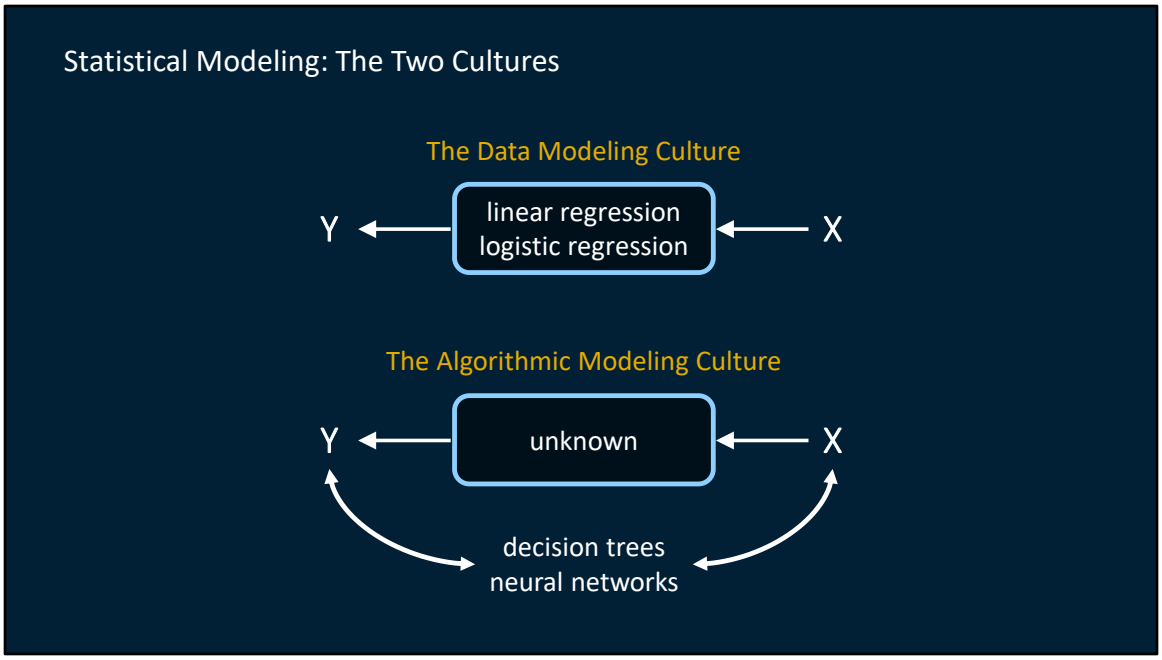
According to Leo Breiman, "There are two cultures in the use of statistical modeling to reach conclusions from data. One assumes that the data are generated by a given stochastic data model. The other uses algorithmic models and treats the data mechanism as unknown." In other words, there is a need to transition from traditional statistical modeling to the machine learning world.



Statistics starts with data. Think of the data as being generated by a black box in which a vector of input variables X go in one side, and on the other side, the response variables Y come out. Inside the black box, nature functions to associate the input variables with the response variables.

There are two goals in analyzing the data: explanation and prediction.

Explanation (or explanatory analysis) is the extraction of some information about how nature associates the response variables to the input variables. Prediction (or predictive modeling) refers to predicting the responses to future input variables. You will learn about both of these data analysis goals in the upcoming lessons.



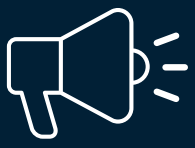
There are two different approaches toward these goals: the data modeling culture and the algorithmic modeling culture.

The analysis in the *data modeling culture* starts with assuming a stochastic data model for the inside of the black box, for example, a linear or a logistic regression model. The values of the parameters are estimated from the data and the model then used for information and/or prediction. Thus, the black box is filled in like the one shown at the top. Here the model is evaluated like a yes-no goodness-of-fit tests and residual examination.

The analysis in the *algorithmic modeling culture* considers the inside of the box complex and unknown. Their approach is to find a framework or mathematical function that operates on X to predict the responses Y. Their black box looks like the one shown at the bottom. In this culture, model is evaluated by predictive accuracy. Most of the data mining and machine learning models falls in this culture, for example, decision trees and neural networks. The difference with classical statistics is that they are less model based, but rather algorithm based, and data driven.

Statistical Modeling: The Two Cultures

marketing campaign example



The Data Modeling Culture

What is the psychology of response behavior? Or, why a person with certain characteristics is not a responder?

The Algorithmic Modeling Culture

Whether a person is going to respond to a marketing campaign or not?

Thus, there are several things you might aim at. For example, consider a marketing campaign scenario. If your aim is to know whether a person is going to respond to a marketing campaign, you run an algorithm on his or her data, and out comes an answer. The main point is that the answer should typically be correct. You might not be so much interested in how the algorithm works. You even might be not interested in the psychology of behavior and why a person with certain characteristics is not a responder. This is a distinctive characteristic of a machine learning algorithm.

Question: Modeling Cultures

Which of the two modeling cultures do you think is used more often in the machine learning world?

a. The data modeling culture
b. The algorithmic modeling culture

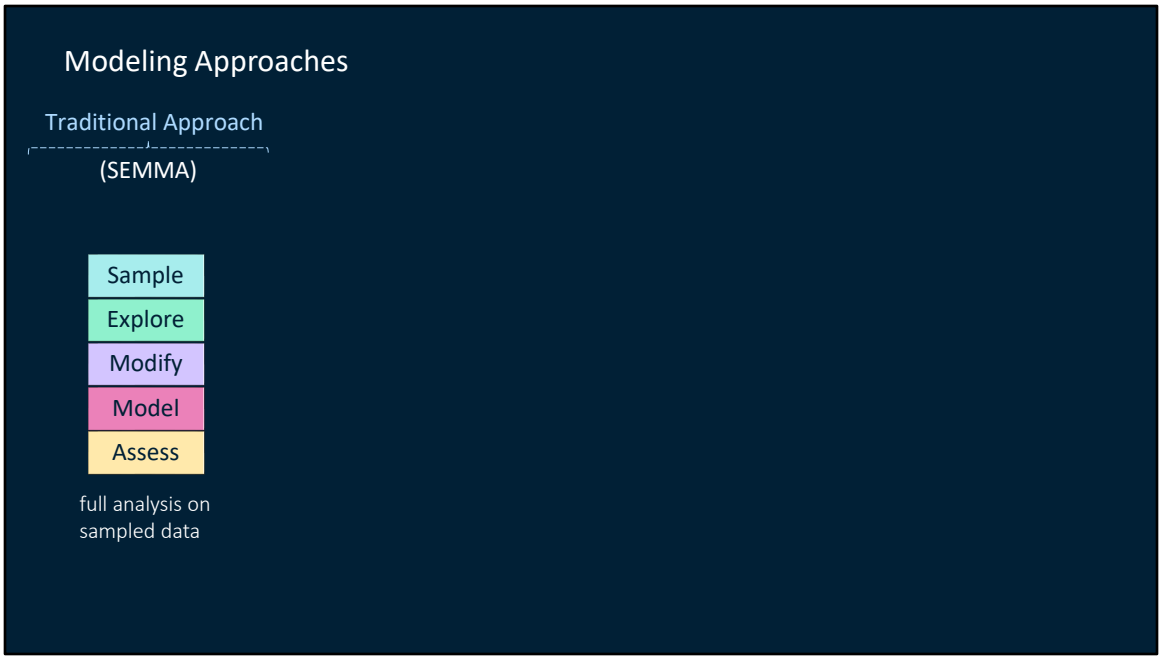
Answer: Modeling Cultures

Which of the two modeling cultures do you think is used more often in the machine learning world?

a. The data modeling culture

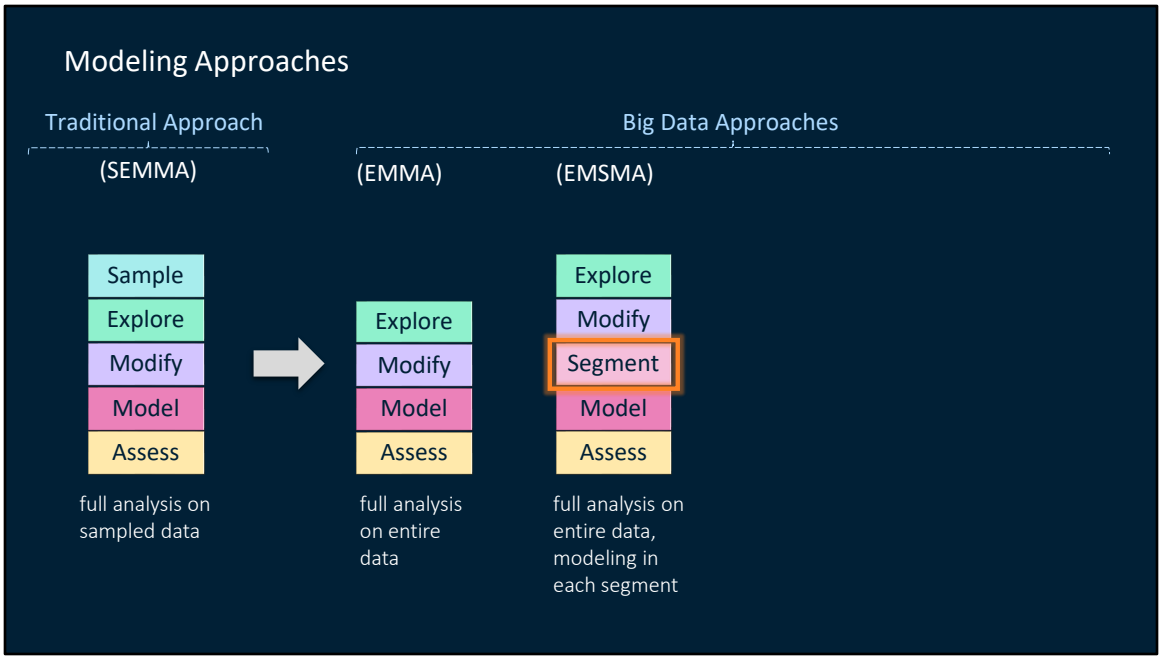
b. The algorithmic modeling culture

Correct answer: b. In the algorithmic modeling culture, the model is evaluated by predictive accuracy. You might not be so interested in how the algorithm works. Most of the data mining and machine learning models fall in this culture.



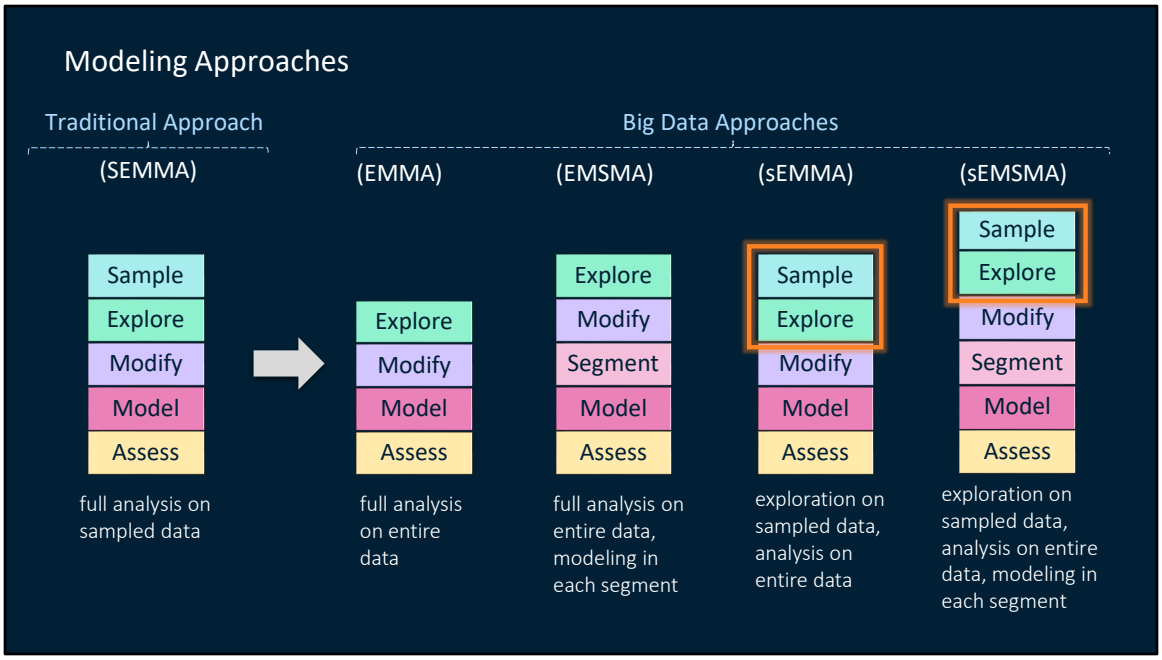
Over time, SAS has followed a number of approaches for data mining and machine learning. The traditional approach is *SEMMA*, which stands for Sample, Explore, Modify, Model, and Assess.

First, you **sample** the data by creating one or more data tables. The samples should be large enough to contain the significant information, but small enough to process. Next, you **explore** the data by searching for anticipated relationships, unanticipated trends, and anomalies in order to gain understanding and ideas. Then you **modify** the data by creating, selecting, and transforming the variables to focus the model selection process. Next, you **model** the data by using the analytical tools to search for a combination of the variables that reliably predicts a desired outcome. Finally, you **assess** competing predictive models.



With the advent of big data, this framework can be slightly modified to many frameworks. For example, *EMMA*: Explore, Modify, Model, and Assess. In EMMA, complete analysis is performed on the entire data, so sampling is not required.

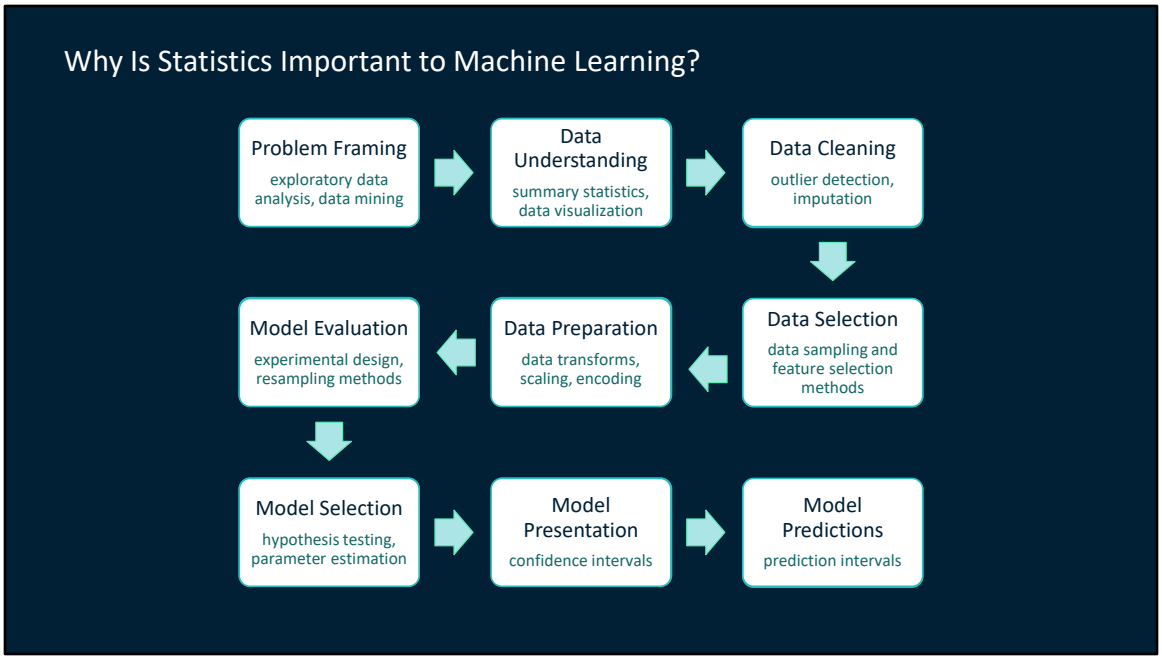
Another variation, *EMSMA*, adds a Segment piece: Explore, Modify, *Segment*, Model, and Assess. Why would you add Segment? Big data has greater diversity in the sense that it represents two or more populations. Different relationships between an input and the target might exist in different populations. For this purpose, you might want to first segment the data and then run a separate predictive model in each segment. However, this modification is not universally accepted and exact.



Sampling in SEMMA is optional, but generally used in case of big data. SAS Viya can handle big data, but it's not entirely correct to say that we don't need to sample now.

Why sample? Real life data is usually dirty and noisy because of inconsistencies, incompleteness, duplication, and merging problems. Preparing your data for analysis starts with exploring your data. Exploring the data can be time-consuming if processed using entire data. So, sampling is inevitable.

After gaining intimate knowledge of the variables by using both graphical and numerical methods, then maybe open the flood gates… let the full data flow… and develop your final, "deeply trained" model. In a big data world, sampling is, in fact, a **best practice** when it comes to rapidly learning and gaining insights, and this leads to modeling approaches where sampling is optional: *sEMMA* (sample, Explore, Modify, Model, Assess) and *sEMSMA* (sample, Explore, Modify, Segment, Model, Assess).



Statistics is a pillar of machine learning. You cannot develop a deep understanding and application of machine learning without statistics. Let's discuss some typical steps where statistical methods are used in an applied machine learning project.

Problem Framing requires the use of exploratory data analysis and data mining.

Data Understanding requires the use of summary statistics and data visualization.

Data Cleaning requires the use of outlier detection, imputation and more.

Data Selection requires the use of data sampling and feature selection methods.

Data Preparation requires the use of data transforms, scaling, encoding, and much more.

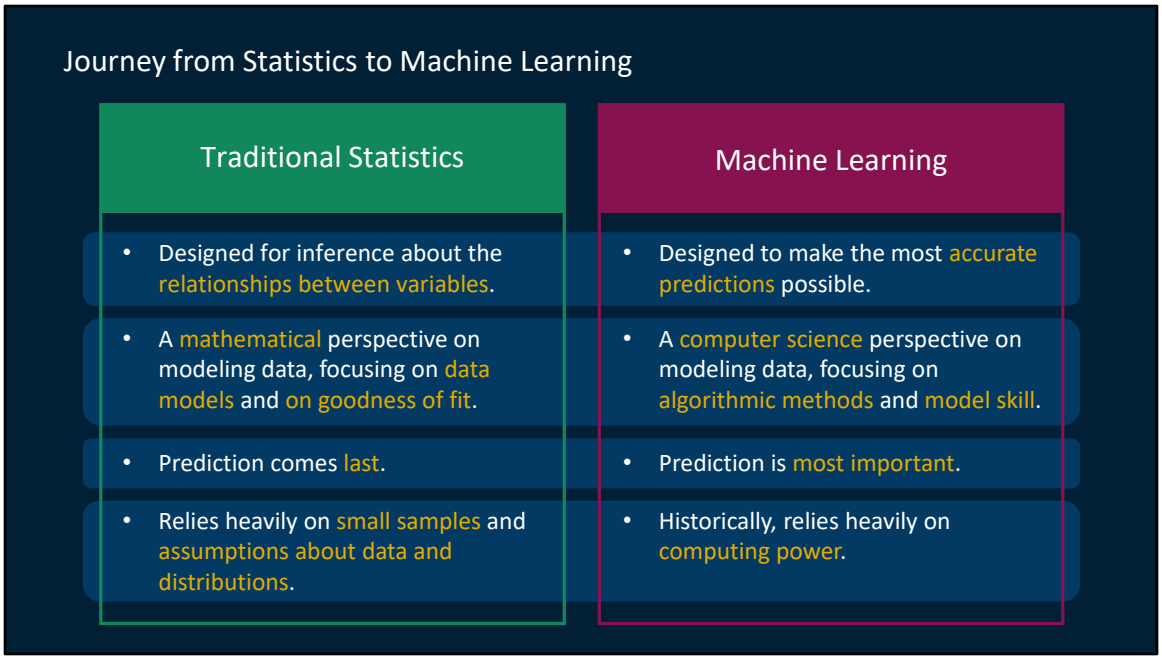
Model Evaluation requires experimental design and resampling methods so that you have the right data to effectively answer your analytical questions.

Model Selection requires the use of statistical hypothesis tests and estimation statistics.

Model Presentation requires the use of estimation statistics such as confidence intervals.

Model Predictions requires the use of estimation statistics such as prediction intervals.

Jason Brownlee discusses these topics in detail in his book **Statistical Methods for Machine Learning**.



When it comes to using data, there are two main camps: traditional statistics and machine learning. The major difference between them is their purpose. Statistical models are designed for inference about the relationships between variables. Machine learning models are designed to make the most accurate predictions possible.

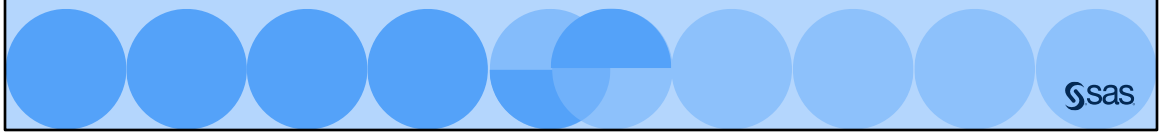
You already know that these two complement each other when it comes to analyzing data. *Statistics* and *statistical learning* are a mathematical perspective on modeling data with a focus on data models and on goodness of fit. *Machine learning* and *predictive modeling* are a computer science perspective on modeling data with a focus on algorithmic methods and model skill.

Much traditional statistics is motivated primarily by showing that one factor causes another, for example, clinical trials. Understanding comes next, prediction last. In machine learning and data mining, the order is usually reversed -- prediction is most important.

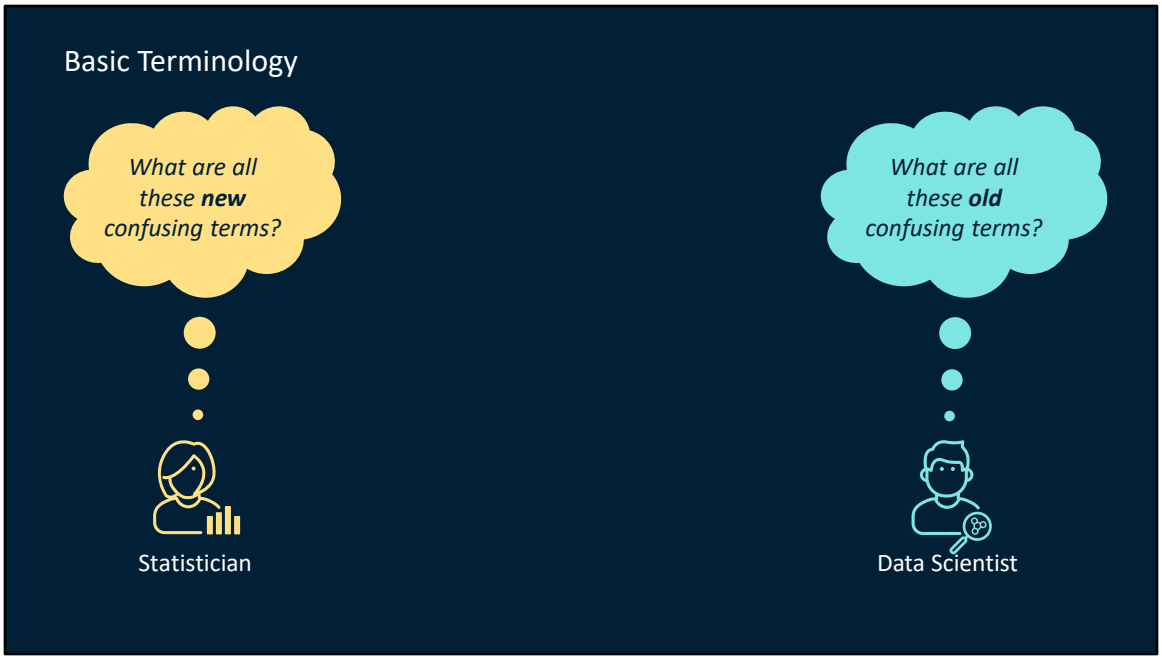
Traditional statistical techniques were mostly developed where computing power was not an option. As a result, traditional statistics relies heavily on small samples and heavy assumptions about data and its distributions. In contrast, historically, machine learning techniques and approach heavily rely on computing power.

More about the Journey from Statistics to Machine Learning

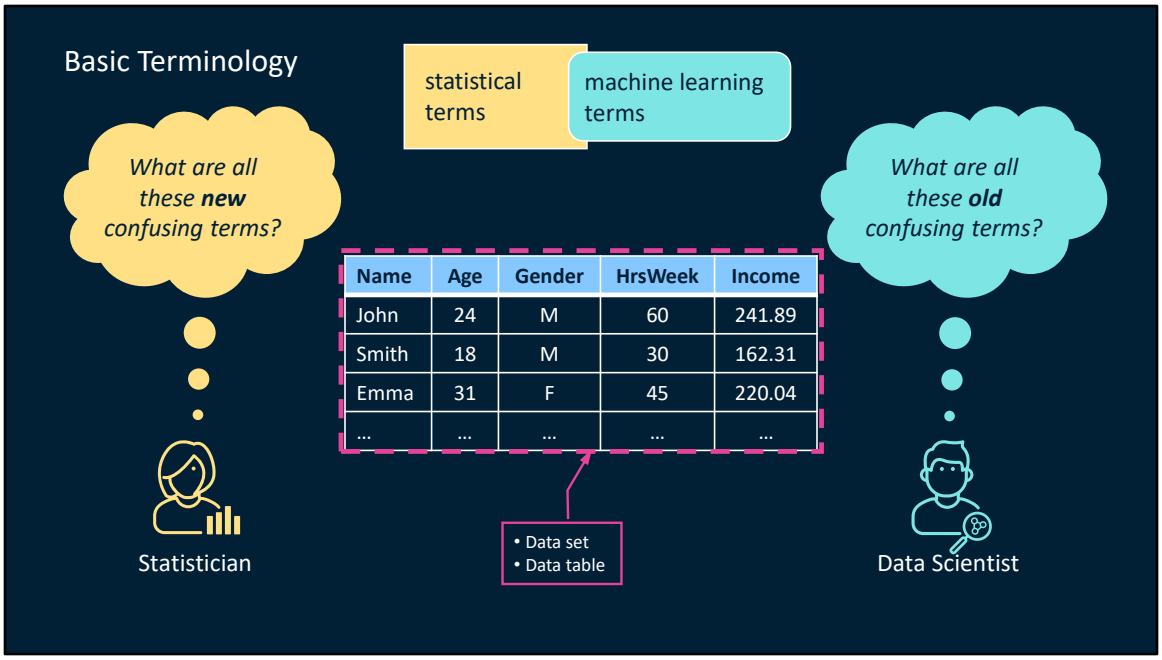
To explore this learning resource in greater detail, please visit the course page on learn.sas.com. There, you'll find comprehensive course materials, tutorials, and updates to support your learning journey.



1.2 Terminology and Vocabulary



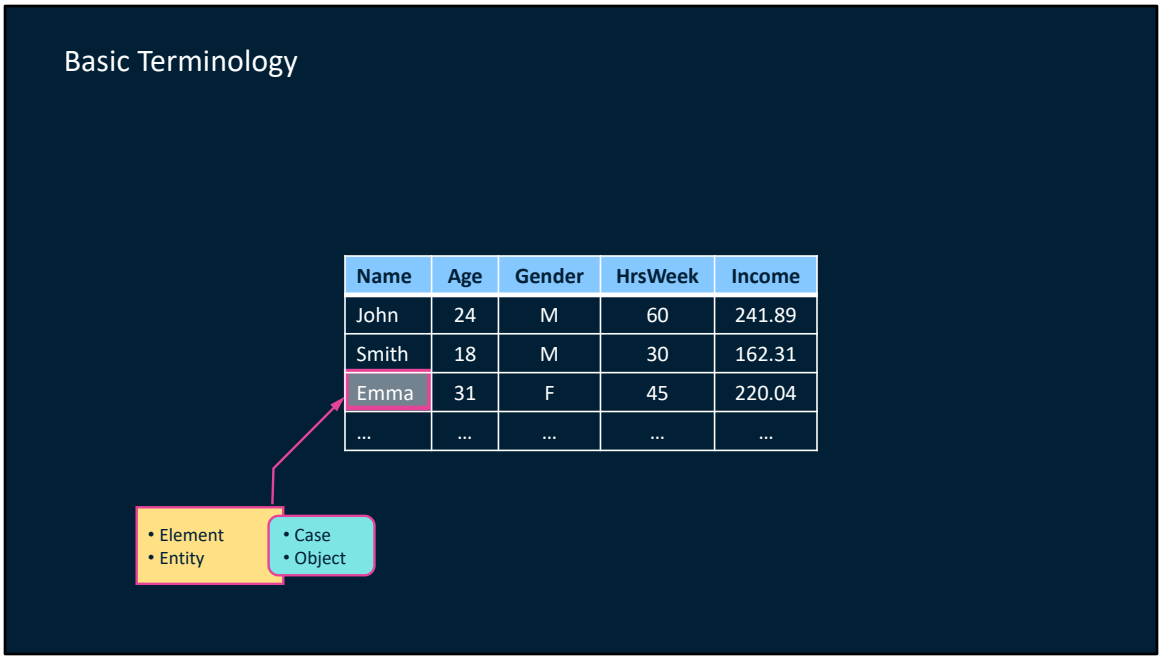
Analytical terms might have different meanings in different contexts, specifically statistics versus machine learning (ML). There are many words for the same thing, and one word might have different meanings to different users. For example, a statistician might say: "*What are all these new fangled and confusing terms?*" Similarly, a data scientist might feel: "*What are all these archaic, outmoded, and confusing terms?*" Always be sure to clarify with your audience that they are using the same definition that you are, and that everyone understands the terms being used.



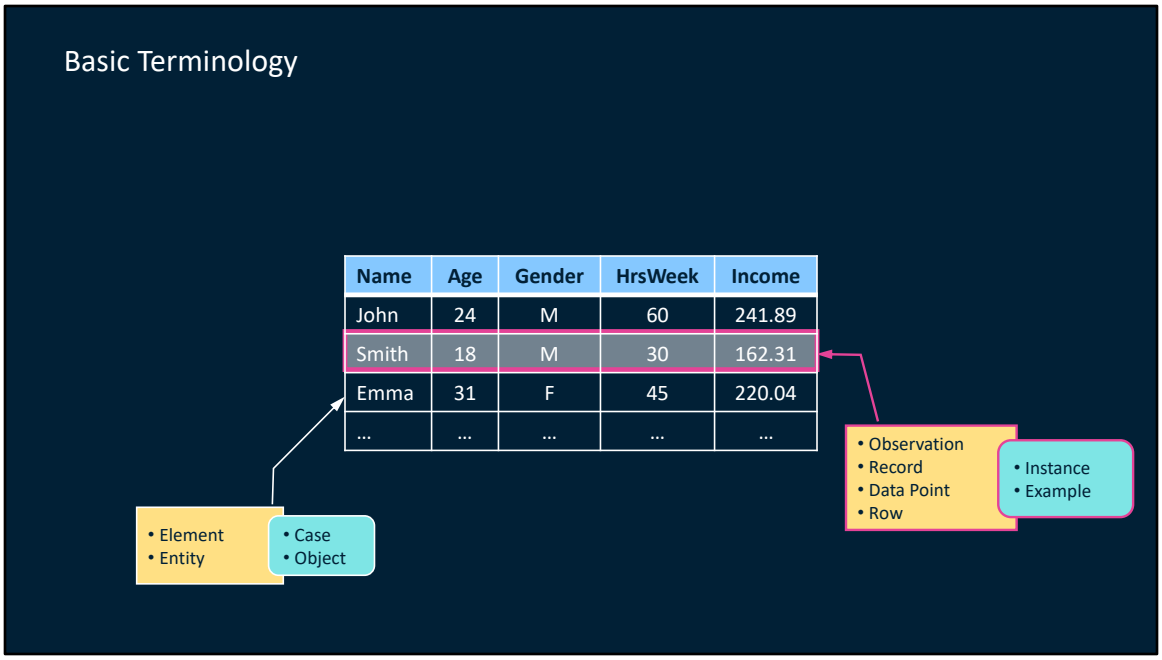
Consider a simple data set containing information on age, gender, working hours per week and income. A *data set* is a table of data from which the machine learns. Let's get acquainted with basic analytical terminology and see some examples of synonyms.

Statistical terms are shown in yellow boxes and their equivalent machine learning terms are shown in the aqua boxes.

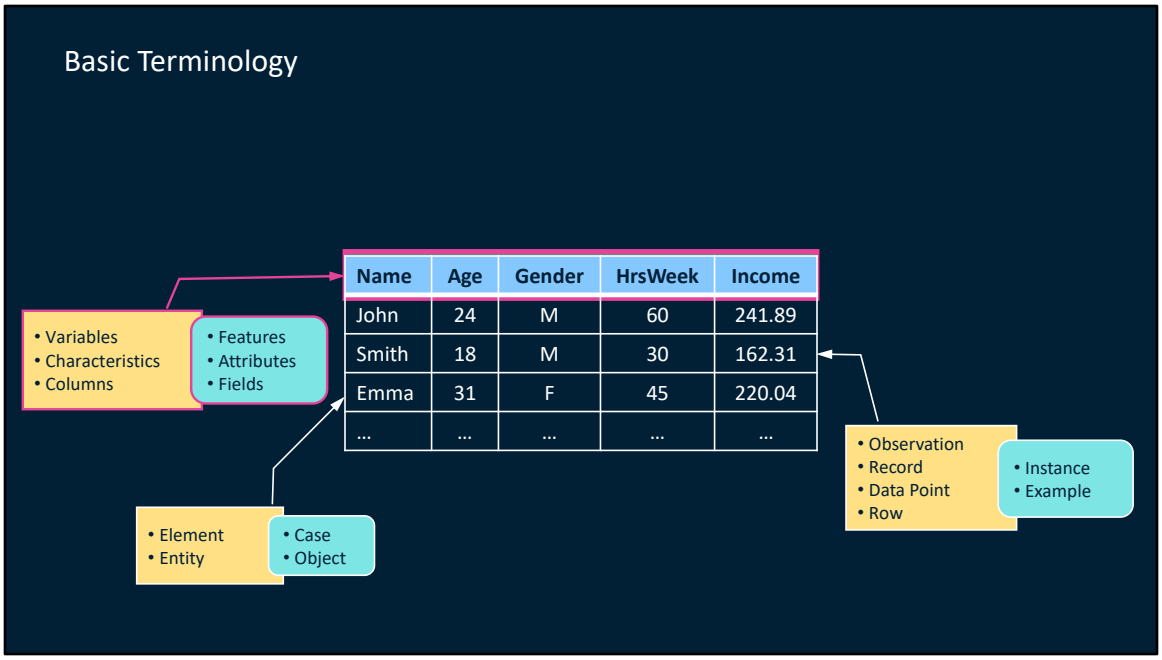
Basic Terminology



The elements are the entities on which data are collected or extracted. In machine learning terminology, they are called cases or objects. John, Smith and Emma are cases here.

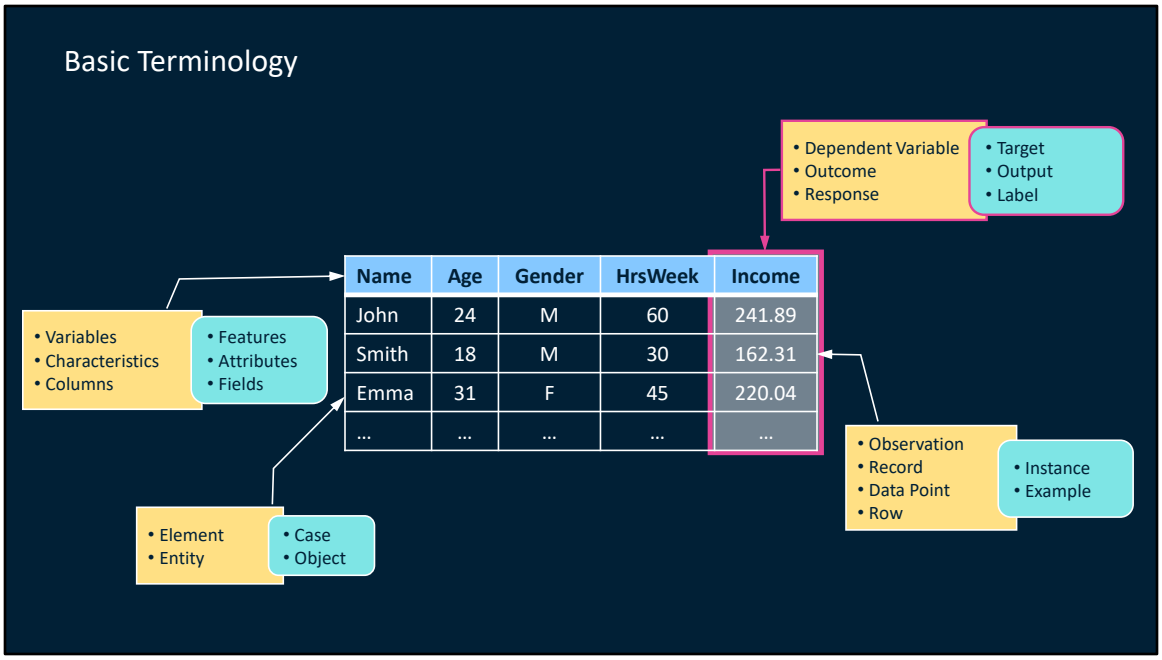


The set of measurements collected or extracted for a particular element is called an *observation* or a *record* or a *data point*. Due to specific format of analysis-based tables, at times it is also called a *row*. In machine learning terminology, it is known as an *instance* or an *example*. The highlighted row is an instance here.

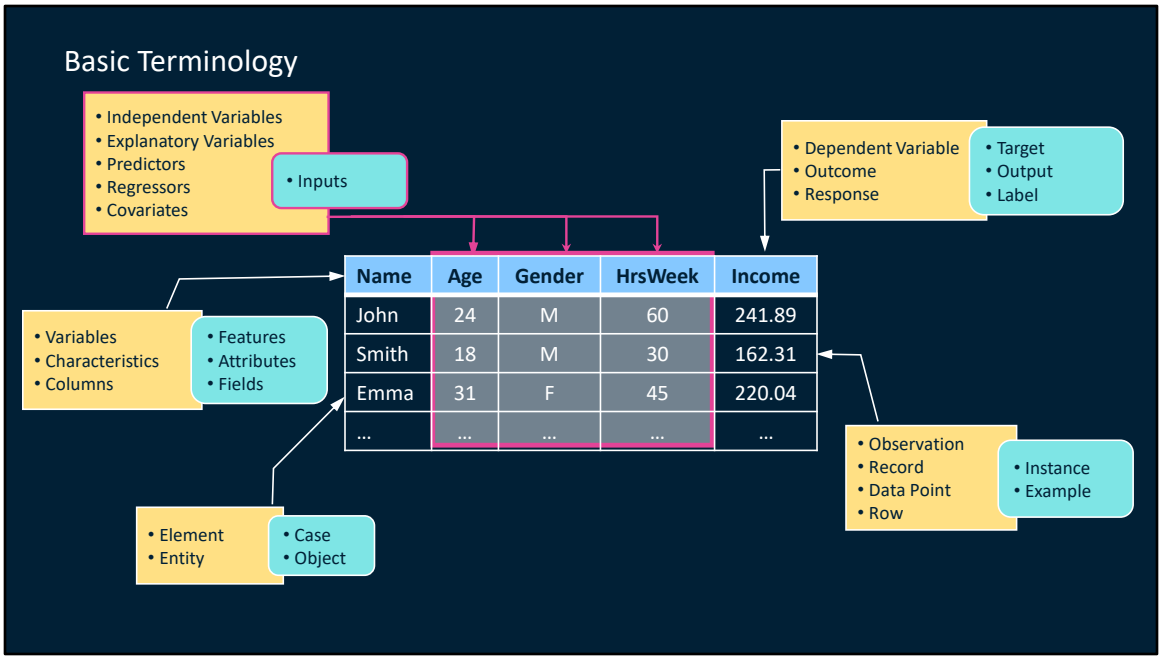


A *variable* is a *characteristic* of interest for the elements. Due to specific format of analysis base table, at times, they are also called *columns*. The total number of data values in a data set is the number of elements multiplied by the number of variables.

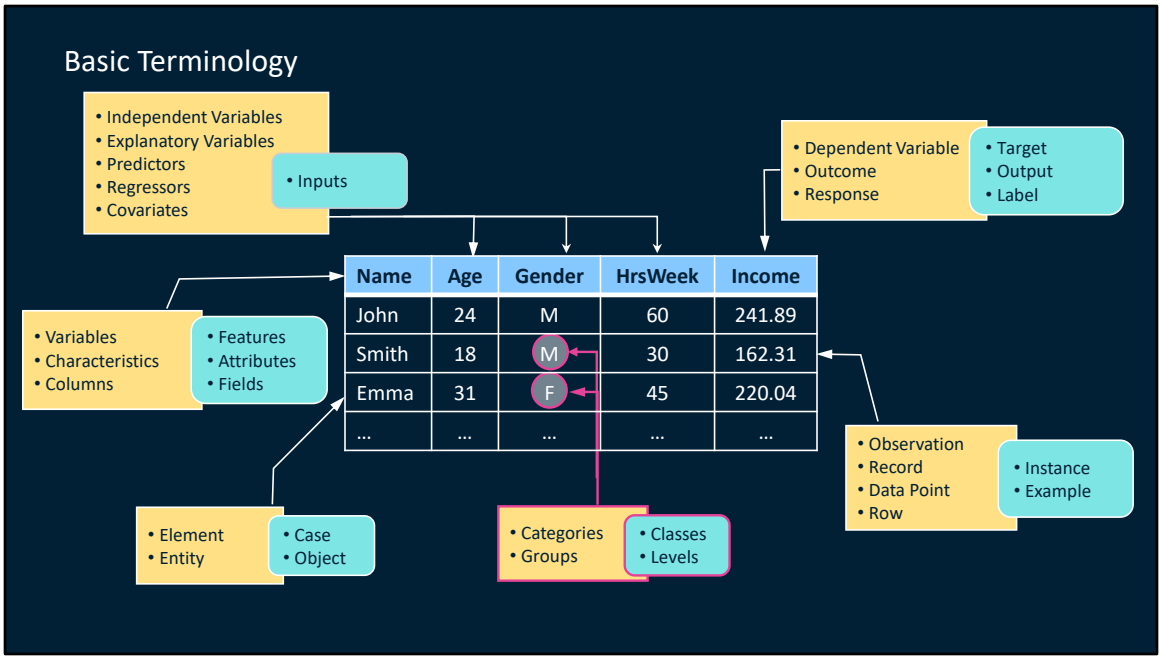
Similarly, an *attribute* or a *field* is an aspect of an instance. For example, age, gender, hours per week, and income are variables. Attributes are often called features in machine learning.



Dependent variable receives this name because, in an experiment, its values are studied under the supposition that they depend on the values of other variables. For example, how much you earn depends on your age and hours per week. In traditional statistical terms, it is also known as *outcome* or *response*. It is sometimes called a variable whose value need to be predicted. In machine learning terminology, it is simply called a *target* or *output.* In case of a categorical target, it can be called a *label*. It is seldom called a training value in data science.



Independent variables, in turn, are not seen as depending on any other variable in the scope of the experiment in question. Independent variables are variables that stand on their own, and you have complete control over which independent variables you choose. During an experiment, you usually choose independent variables that you think will affect dependent variable. For example, age and hours per week might affect the income, but not gender. In traditional statistical terms, they are also known as *explanatory variables* or *predictors* or *regressors* or *covariates*. In machine learning terminology, they are simply called *inputs*.



In this example, the variable gender is somewhat different from variables: age, hours per week, and income. Gender does not have fractional values, rather it observes only three values – male, female, and others. They are called *categories* or *groups* in statistical terminology or *classes* or *levels* in the machine learning world. Discussed next is the difference between continuous and categorical variables.

Question: Basic Terminology (1)

Which of the following pairs of terms define the highlighted part in the data table? (Select all that apply.)

a. Target / Response
b. Instance / Example
c. Predictor / Input
d. Observation / Record

Company	Stock Exchange	Annual Sales ($M)	Earning per Share ($)
MedTech	AMEX	84.20	0.33
EnergyWest	NYSE	372.85	1.89
DataDig	BSE	74.65	2.10
...	...	...	...

Answer: Basic Terminology (1)

Which of the following pairs of terms define the highlighted part in the data table? (Select all that apply.)

a. Target / Response
b. Instance / Example
c. Predictor / Input
d. Observation / Record

Company	Stock Exchange	Annual Sales ($M)	Earning per Share ($)
MedTech	AMEX	84.20	0.33
EnergyWest	NYSE	372.85	1.89
DataDig	BSE	74.65	2.10
...	...	...	...

Correct answer: b and d.

The highlighted row - a set of measurements collected or extracted for EnergyWest company - is an *observation* or a *record.* In machine learning terminology, it is known as an *instance* or an *example*.

Note that the terms in **a** are equivalent for the outcome variable or the dependent variable. The terms in **c** are equivalent for the independent variable also known as explanatory variable.

Question: Basic Terminology (2)

Which of the following pairs of terms define the highlighted part in the data table? (Select all that apply.)

a. Variable / Characteristic
b. Category / Group
c. Class / Level
d. Feature / Attribute

Company	Stock Exchange	Annual Sales ($M)	Earning per Share ($)
MedTech	AMEX	84.20	0.33
EnergyWest	NYSE	372.85	1.89
DataDig	BSE	74.65	2.10
...	...	...	...

Answer: Basic Terminology (2)

Which of the following pairs of terms define the highlighted part in the data table? (Select all that apply.)

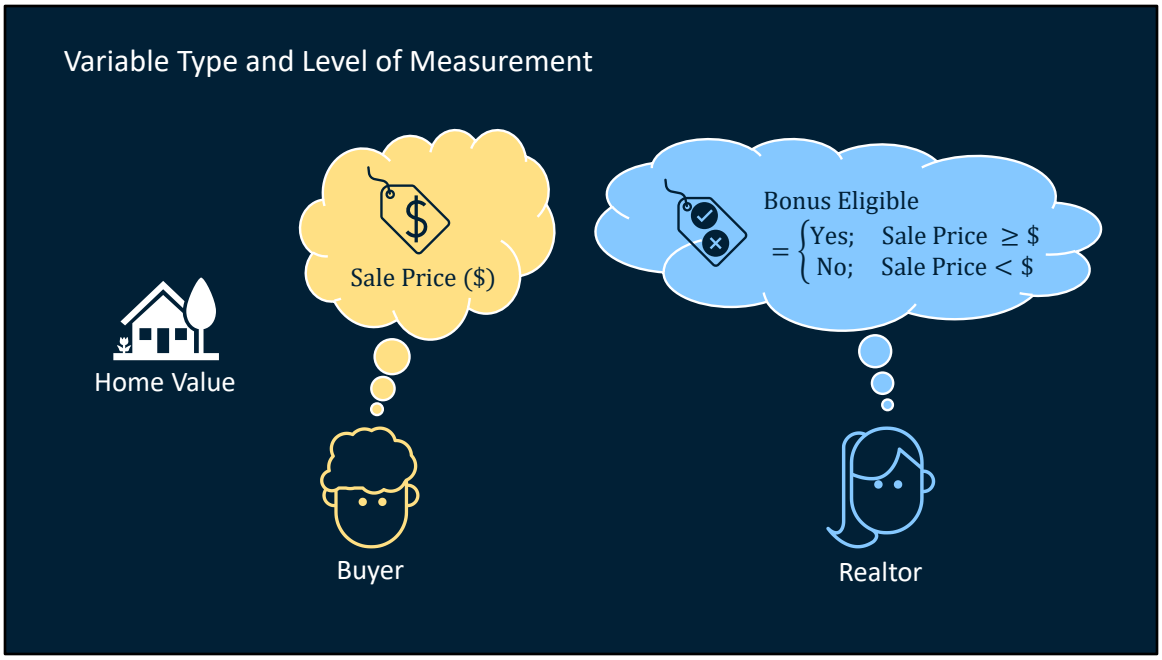
a. Variable / Characteristic
b. Category / Group
c. Class / Level
d. Feature / Attribute

Company	Stock Exchange	Annual Sales ($M)	Earning per Share ($)
MedTech	AMEX	84.20	0.33
EnergyWest	NYSE	372.85	1.89
DataDig	BSE	74.65	2.10
...	...	...	...

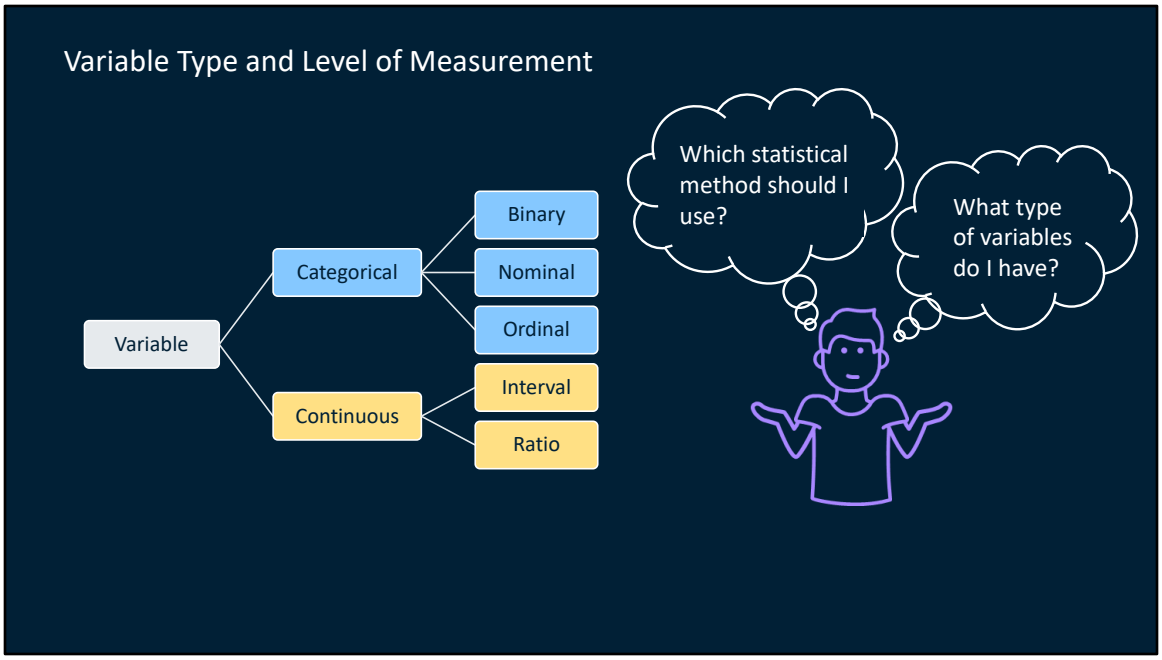
Correct answer: **a** and **d**.

The highlighted column is a *variable* or a *characteristic* of interest. Similarly, an *attribute* is an aspect of an instance. Attributes are often called features in machine learning.

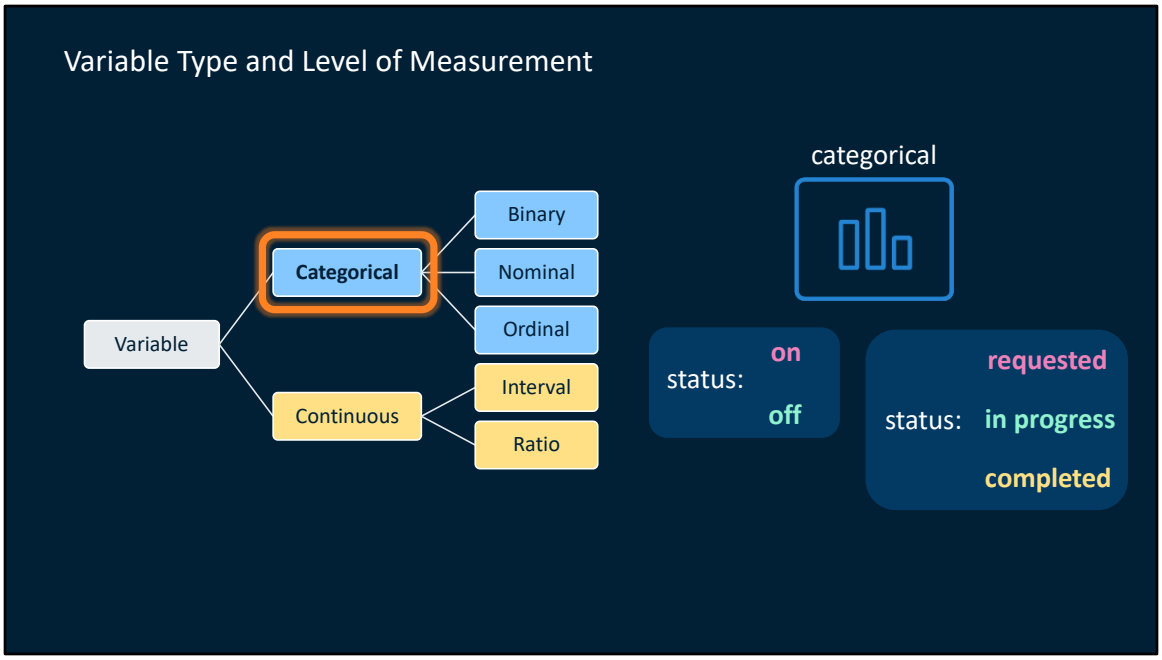
Note that the terms in **b** and **c** are equivalent. They are used to represent the variable values that have observed values rather than fractional values. Examples are the Stock Exchange values of AMEX, NYSE, BSE, and so on. They are called *categories* or *groups* in statistical terminology or *classes* or *levels* in the machine learning world.



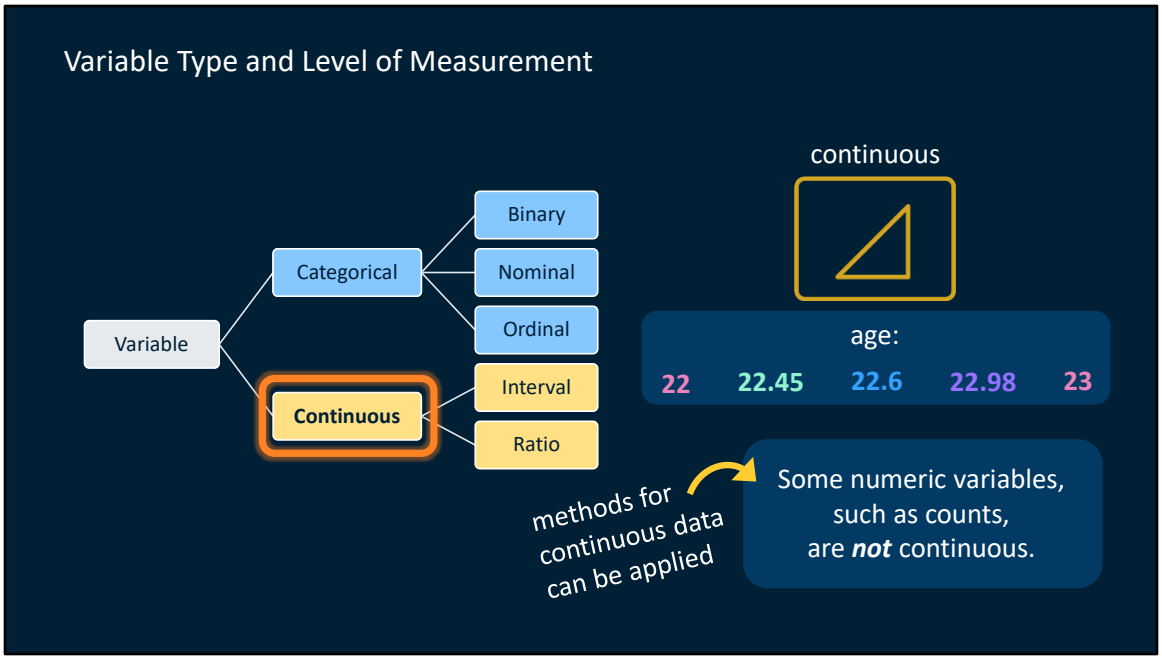
Data comes in several different types. For example, consider the sale of individual residential property. A buyer would be interested in exact sale price in dollars. So, the Sale Price variable will have many different dollar values. On the other hand, a realtor would be interested in considering the variable differently. Realtors receive some standard percentage of commission on homes sales if sale price is more than some dollar amount otherwise not. Thus, the same variable turned dichotomous, observing only two values: bonus eligible or not bonus eligible depending on dollar sale price.



There are a variety of statistical methods for analyzing data. To choose the appropriate method, you must determine the type and level of measurement for your variables. The most basic distinction is that between continuous and categorical data, which has a profound impact on the types of statistics that can be used.



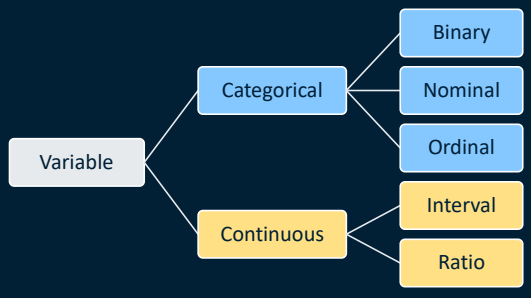
Categorical variables are variables that represent groupings. Categorical variables are also known as discrete or qualitative variables. Categorical variables can be stored as numeric or non-numeric values in SAS. Categorical variables can be further categorized as either binary, nominal, or ordinal.



Continuous variables can take on any of an infinite number of possible values between two numbers, and the relative distances between the values are meaningful. An example of a continuous variable would be age. There is no restriction to the number of values between 22 and 23. You could have an age of 22.45 or 22.98 years. A variable can be continuous even if your measurement system has finite intervals. For instance, we normally report age as 22 years, 23 years, and so on.

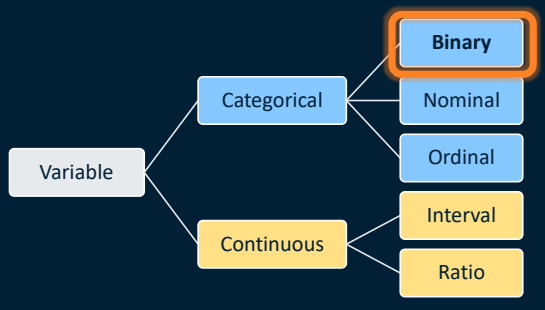
Some numeric variables are not continuous. These include variables such as counts that can only take on specific values (for example, integers). In many statistical applications, methods for continuous data can be applied to these variables as well provided there are many discrete values. Continuous variables can be either "ratio" level or "interval" level.

Variable Type and Level of Measurement



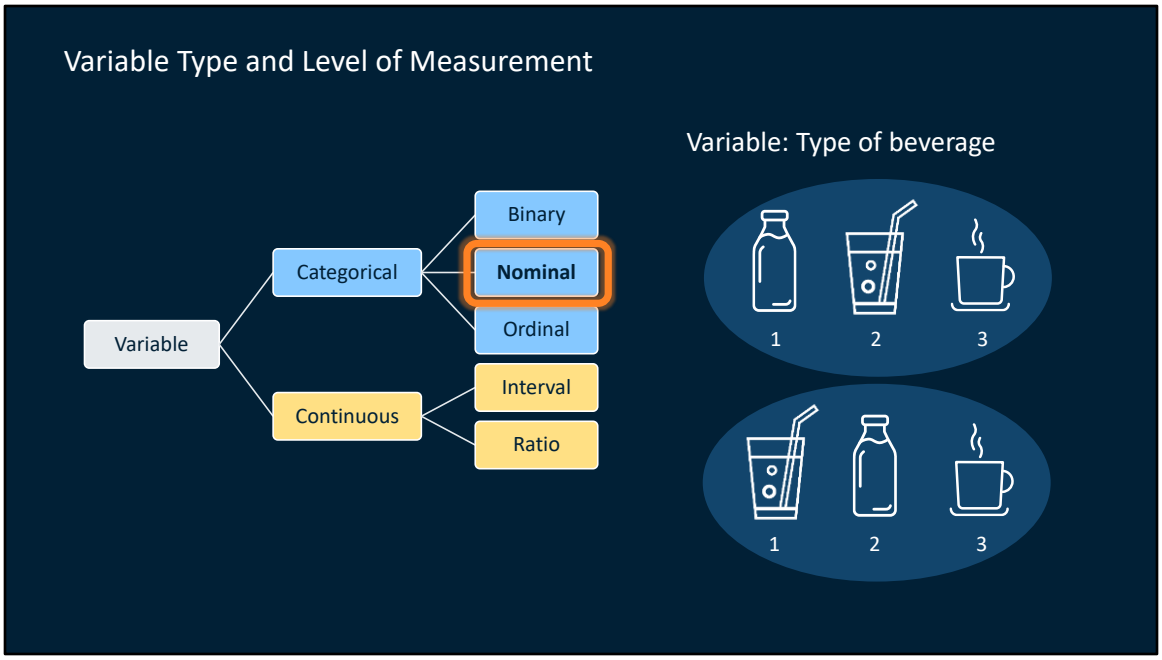
Suppose you have a certain beverage, say, a glass of soda. Keeping this as an object example, we can describe what is binary, nominal, ordinal, interval, and ratio scales.

Variable Type and Level of Measurement

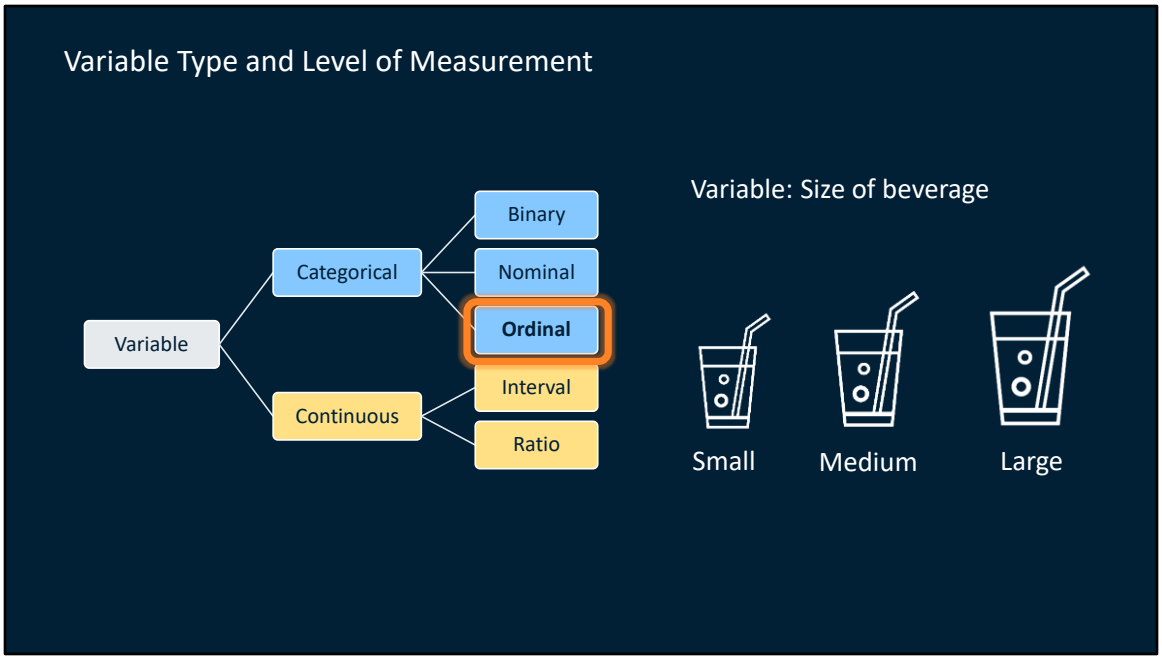


Variable: Do you have a beverage?
1 or 0 , Yes or No

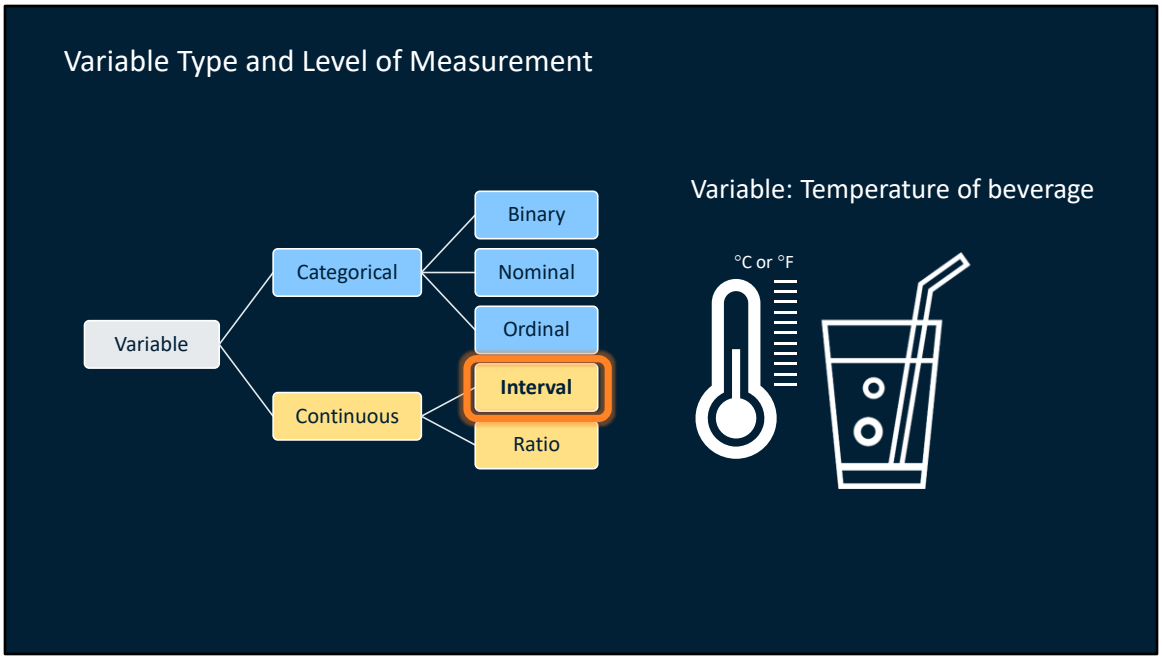
Binary variables are variables that take only two values. For example, do you have a beverage or you do not have beverage. Many variables and questions are naturally binary, but it is often useful to construct binary variables from other types of data.



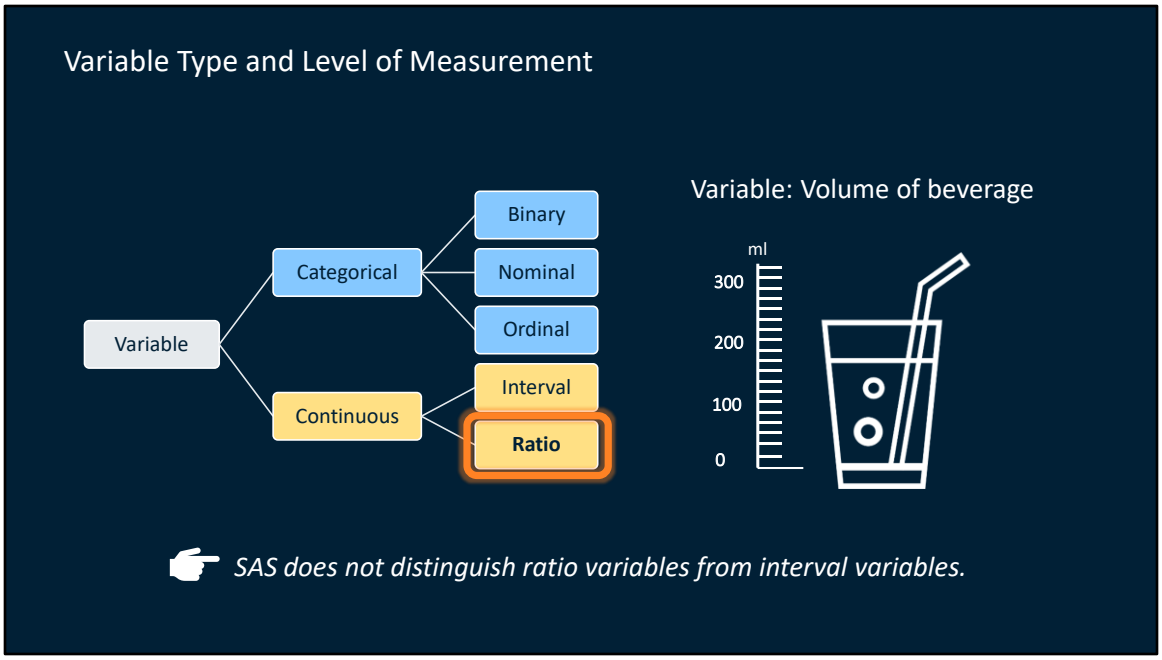
Data on a nominal scale represent a set of categories that differ in some characteristic that does not necessarily have a measurable magnitude. Observations on a nominal scale are classified in one of several non-overlapping categories, also called the *levels* of the variables. *Nominal variables* have values with no logical ordering. In this example, beverage type is a nominal variable, even though numeric values are assigned to the categories.



Ordinal variables have values with a logical order. However, the relative distances between the values are not clear. In this example, drink size is ordinal. The level assignments to the categories - small, medium, and large - convey information of the relative size of the drinks, but not precise information about the quantitative differences. For example, this scale did not suggest that small size was twice as small as the medium size nor that the difference between the small size and the medium size was the same as the difference between medium size and the large size beverage.



Interval measures have no true zero point, for example, temperature of beverage. Zero degrees Celsius does not imply the absence of heat. Interval scales are so called because when two measurements differ by an additive factor d, the larger is said to be d units higher than the lower. For example, if the temperature of your beverage is 40 degrees Fahrenheit and the temperature of another beverage is 30 degrees, then your drink is 10 degrees warmer, but it would be inaccurate to say that your drink is 25% hotter (as for data on a ratio scale). The difference between the volume in ml and the temperature in degrees Fahrenheit is that the Fahrenheit scale has an arbitrary 0 (at approximately −18 degrees Celsius), just as the Celsius scale has an arbitrary 0 at 32 degrees F.



Ratio measures have a true zero point, for example, volume of a beverage. Zero milliliters implies the absence of beverage. Ratio variables differ from that of interval variable in the existence of a "true" zero point. This zero point allows ratios to be calculated. They are called ratio scales because when two measurements differ by a multiplicative factor r, then the one value is said to be r times larger or smaller than the other. For example, if the volume of a drink is 100 ml and subsequently you poured 50 ml more in it, you can say that the volume has increased by 50%. The ratio factor is r=1.5. Likewise, a rise from 100 ml to 200 ml would represent a doubling in volume.

In SAS, ratio variables are not distinguished from interval variables, because the analytical methods used for them are the same.

Question: Measurement Scales

Which of the following represents a nominal measurement scale?

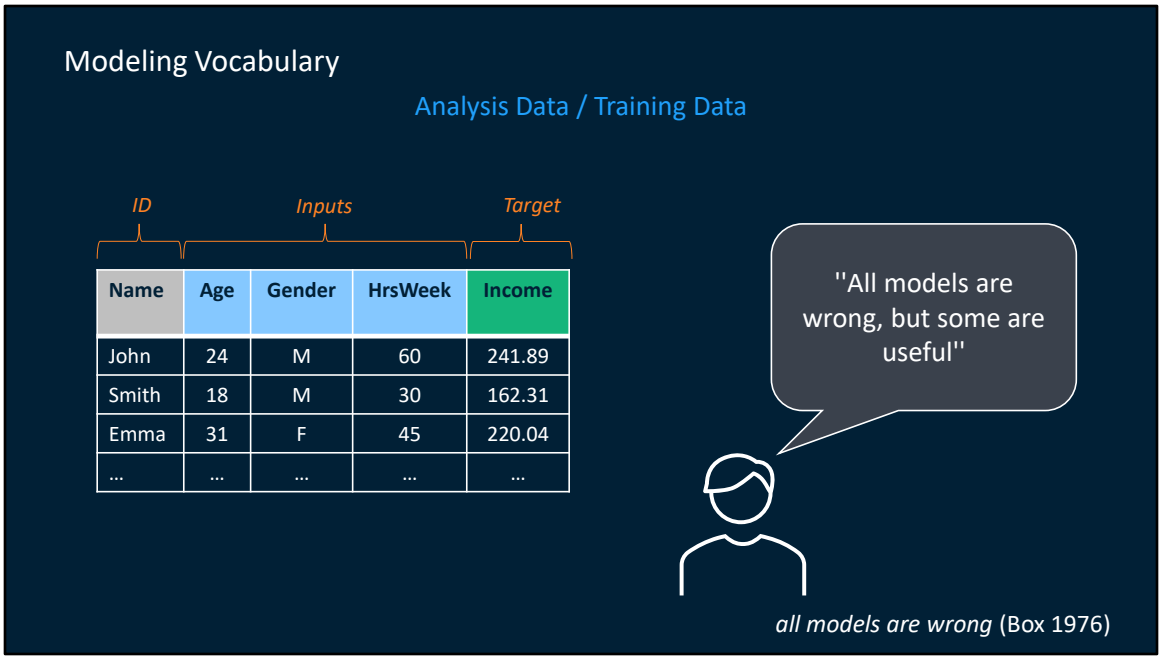
a. Jersey Numbers Assigned to Runners			
b. Rank of Winners	 Third place	 Second place	 First place
c. Performance Rating on a 0 to 10 Scale	8.2	9.1	9.6
d. Time to Finish, in Seconds	15.2	14.1	13.4

Answer: Measurement Scales

Which of the following represents a nominal measurement scale?

a. Jersey Numbers Assigned to Runners			
b. Rank of Winners	 Third place	 Second place	 First place
c. Performance Rating on a 0 to 10 Scale	8.2	9.1	9.6
d. Time to Finish, in Seconds	15.2	14.1	13.4

The correct answer is a. Rank of winners = ordinal. Performance rating on a 0-10 scale = interval. Time to finish in seconds = ratio.



You might be familiar with the phrase ''All models are wrong, but some are useful'' attributed to George Box. Well, in an era of massively abundant data, we don't have to settle for wrong models.

Let's begin our discussion of modeling with an overview of common modeling vocabulary. Modeling starts with a set of *analysis data* called *training data*. The observations in a training data set are known as *training cases* or simply *cases*. The variables are called *inputs* and *targets*. The measurement scale of the inputs and the target can be varied. The inputs and the target can be numeric variables, such as **age**, **hours per week**, and **income**. They can be nominal variables, such as **gender**.

Modeling Vocabulary

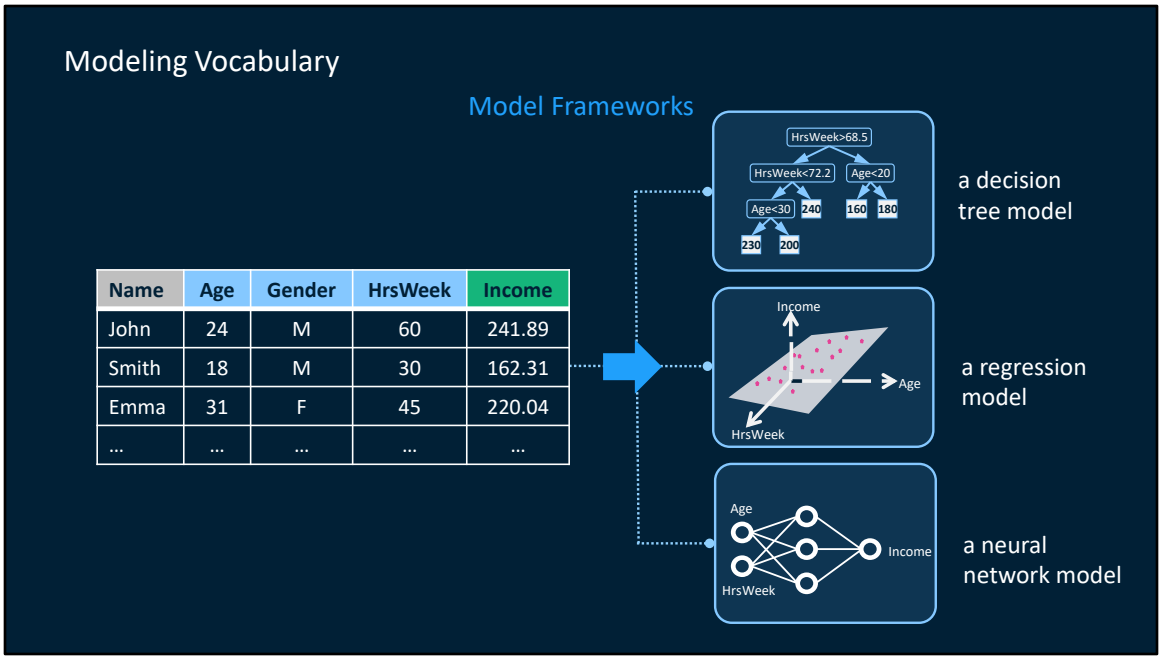
Model

A **model** is a concise representation of the inputs and target association.

Name	Age	Gender	HrsWeek	Income
John	24	M	60	241.89
Smith	18	M	30	162.31
Emma	31	F	45	220.04
…	…	…	…	…

model

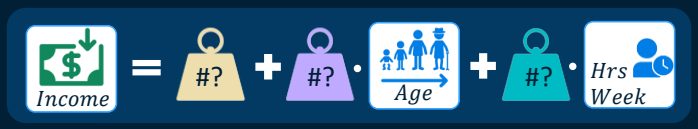
The purpose of the analysis data is to generate a model. A *model* is a concise representation of the relationship between the inputs and the target variables. For example, you might want to create a model to explain variability in income using age, gender, and weekly work hours, or you might want to predict income using these features. As such, a model is a formal exemplification of a theory and is usually specified as a mathematical relationship to approximate the reality.



A model stores a representation of an input–output relationship in many frameworks. It could be a mathematical equation involving parameter estimates in a typical regression model, or a more complex mathematical equation involving weights and biases in a neural network model, or it might have certain prediction rules representing input-output relationships like in a decision tree model.

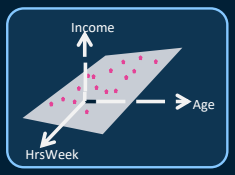
Modeling Vocabulary

Regression Model Example



linear combination

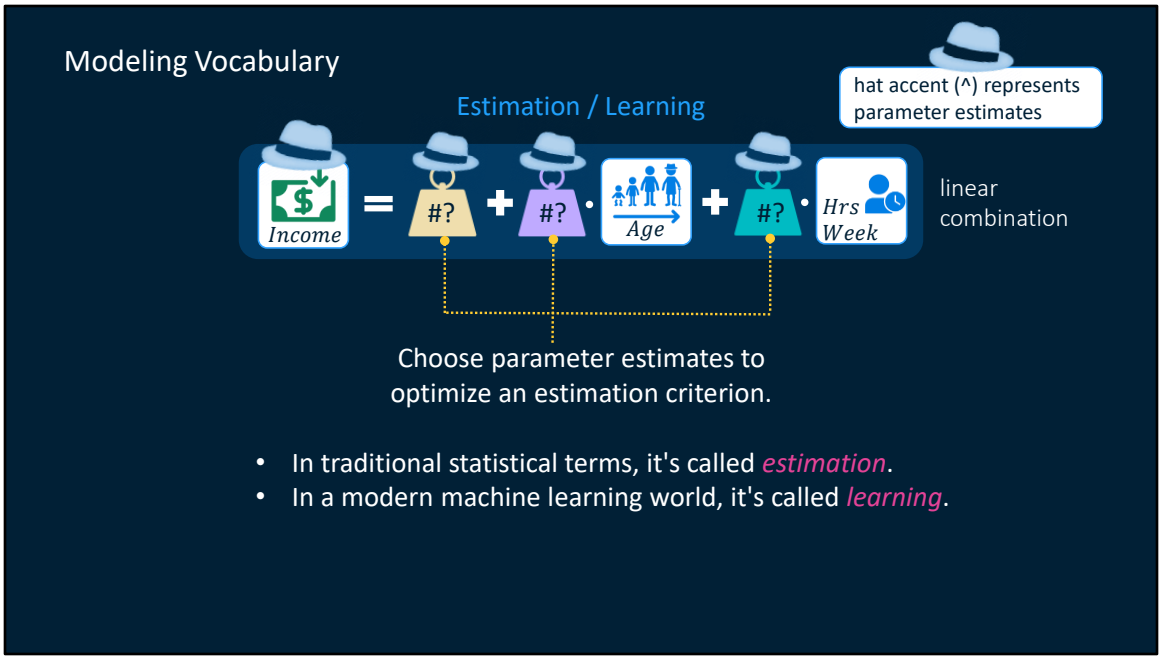
Name	Age	Gender	HrsWeek	Income
John	24	M	60	241.89
Smith	18	M	30	162.31
Emma	31	F	45	220.04
…	…	…	…	…



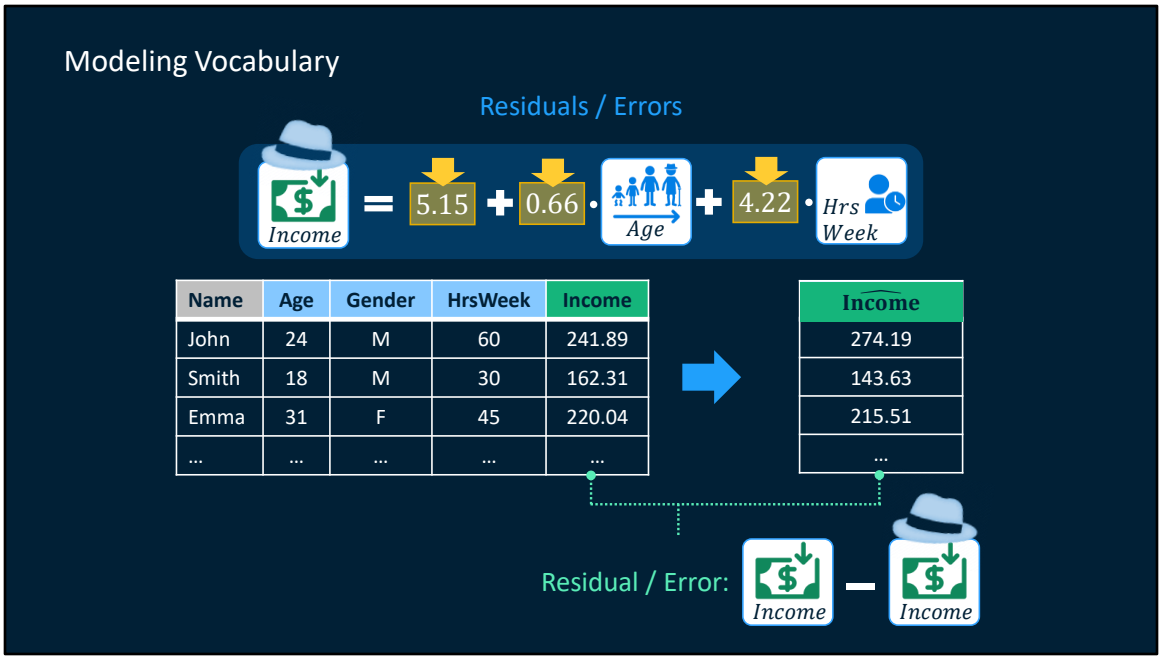
a regression model

Let's understand more modeling vocabulary using a simple example of a regression model. Linear regressions are usually deployed for targets with an interval measurement scale, like income as shown here with two inputs – age and weekly working hours. A prediction estimate for the target variable is formed from a linear combination of the inputs. Each input is weighted differently to determine the trend strength between each input and the target. A bias weight is also incorporated that centers the range of predicted target values. These are called model parameters and they are unknown.

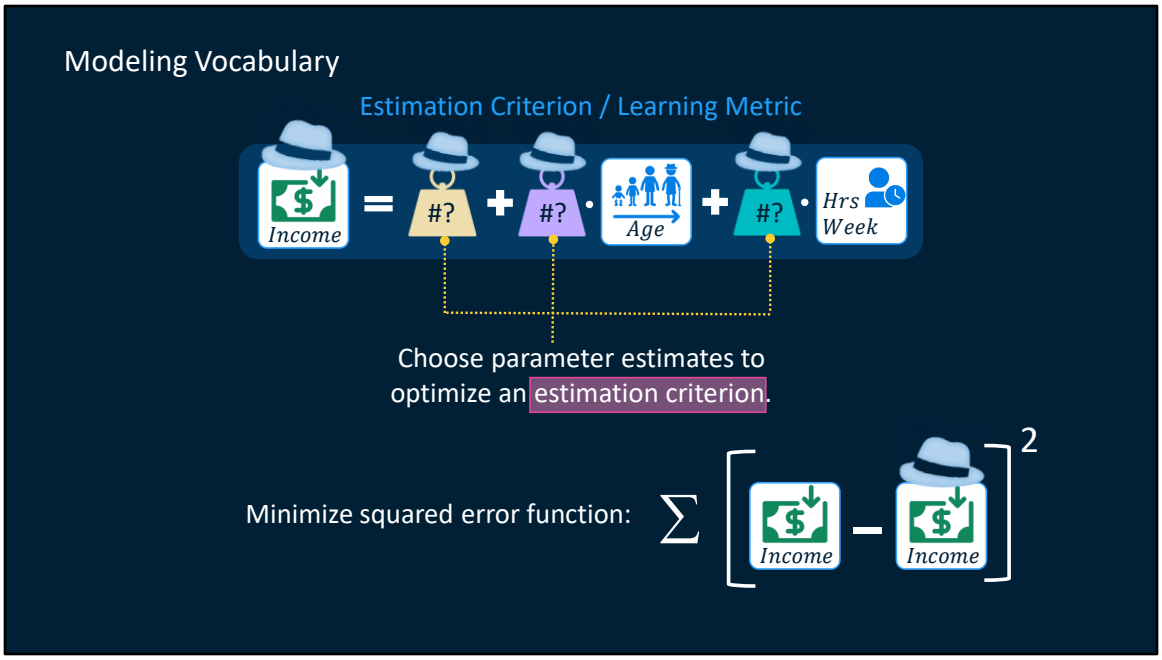
You will learn about regression models in detail in one of the upcoming lessons.



The primary task in a regression model is to obtain parameter estimates. Statisticians generally represent parameter estimates using a hat accent. The estimates are best chosen by optimizing an estimation criterion represented mathematically. In traditional statistical terms, this process is called *estimation,* and in a more modem machine learning world, it's called *learning.* The bottom line is that the parameters are estimated or learned to optimize a mathematical function. The function you want to optimize is called the objective function, or estimation criterion.

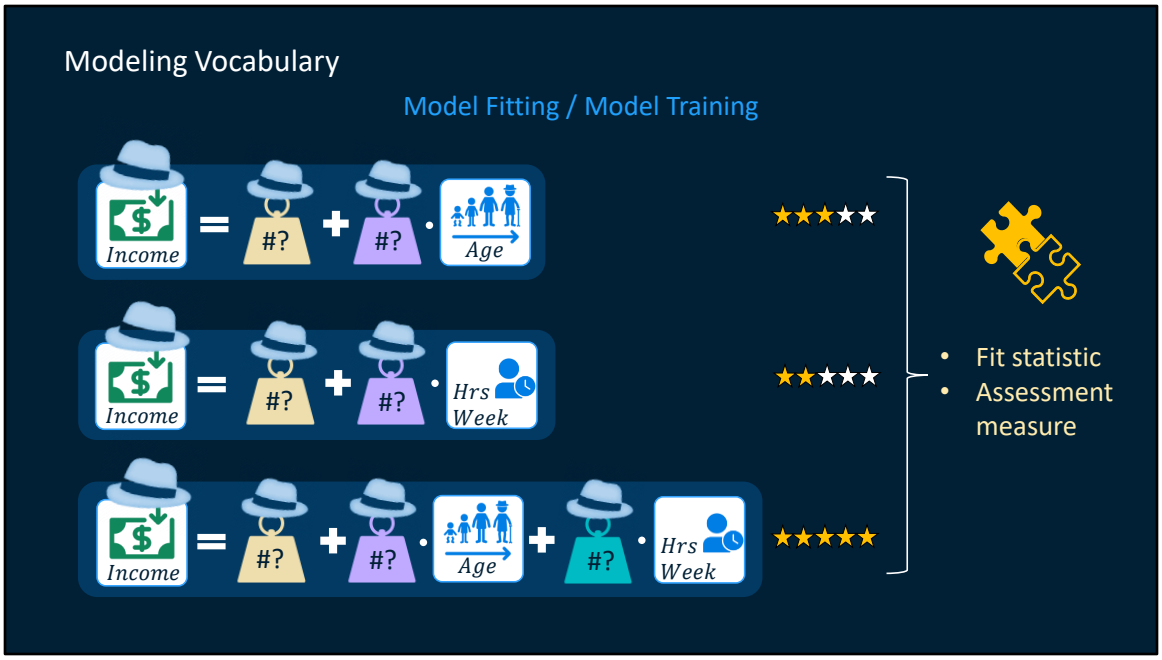


After the parameters are estimated, predicted values on the analysis data can be generated, such as predicted income. The difference between observed and predicted values of data are called *residuals* in a statistical or machine learning model. They are a diagnostic measure used when assessing the quality of a model. They are also known as *errors*.



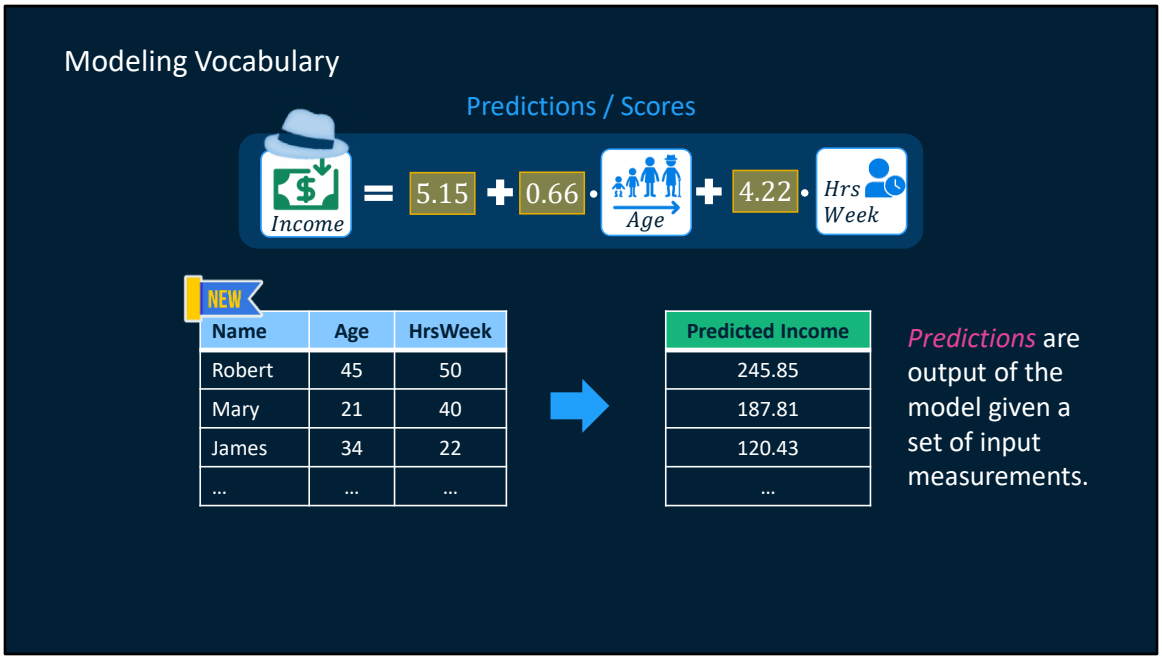
At its core, the estimation criterion or learning metric is incredibly simple: It's a method of evaluating how well your algorithm models your data set.

For example, in a linear regression model, *parameter estimates* are chosen to minimize the squared error between the predicted and observed target values. This method is called *least squares estimation*. The prediction estimates can be viewed as a linear approximation to the expected (average) value of a target conditioned on observed input values.



Fitting a model to data means choosing a statistical model that predicts values as close as possible to the ones observed in your population. In the machine learning world, determining the best fit model is also known as *model training*.

For example, several regression models can be created with multiple combinations of input variables. The purpose of model training is to build the best mathematical representation of the relationship between inputs and a target. While doing such a statistical analysis, it's very important to make sure about the goodness of fit of the model used. To do so, there are different *fit statistics* that you can use, such as R-square, the coefficient of determination. Fit statistics are generally called *assessment measures* in machine learning terminology.



When the *best-fit model* is determined, *predictions* can be generated. The outputs of the predictive model are referred to as *predictions*. Predictions represent your best guess for the target given a set of input measurements. The predictions are based on the associations learned from the training data by the predictive model. Predictions are generally called *scores* in machine learning terminology.

Question: Modeling Vocabulary (Model)

Which of the following defines the analytical term ***model***?

a. Choosing model parameter estimates by optimizing an estimation criterion
b. A measure of evaluating how well your algorithm models your data set
c. A concise representation of the inputs and target association
d. A person with a role either to promote, display, or advertise commercial products

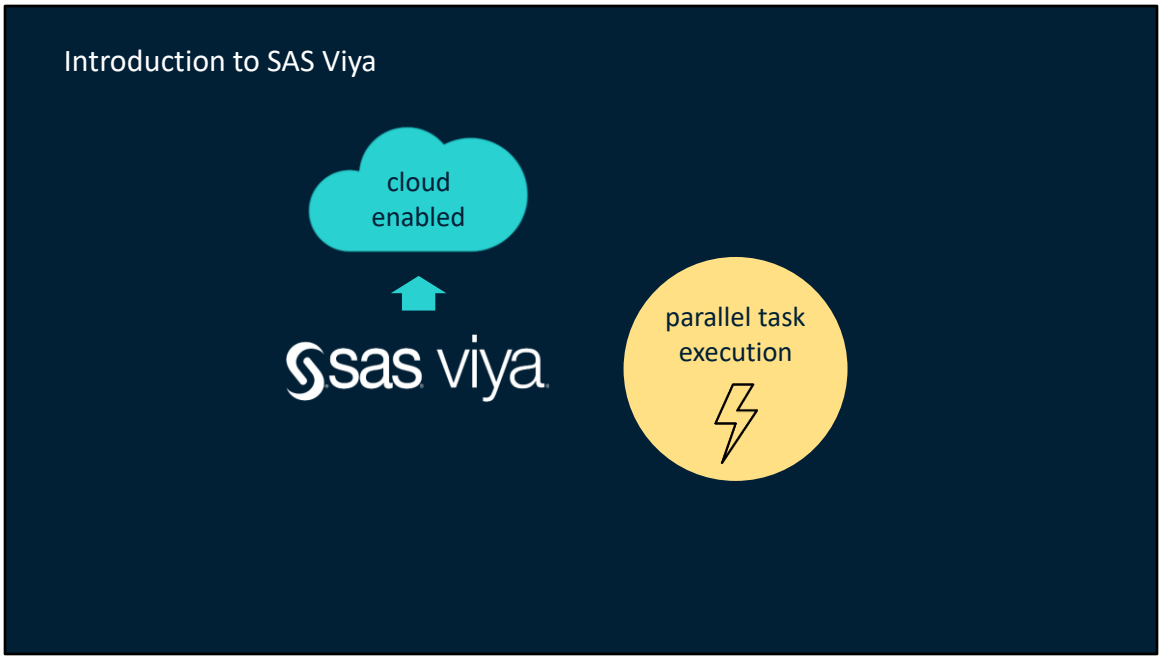
Question: Modeling Vocabulary (Model)

Which of the following defines the analytical term ***model***?

a. Choosing model parameter estimates by optimizing an estimation criterion
b. A measure of evaluating how well your algorithm models your data set
c. A concise representation of the inputs and target association
d. A person with a role either to promote, display, or advertise commercial products

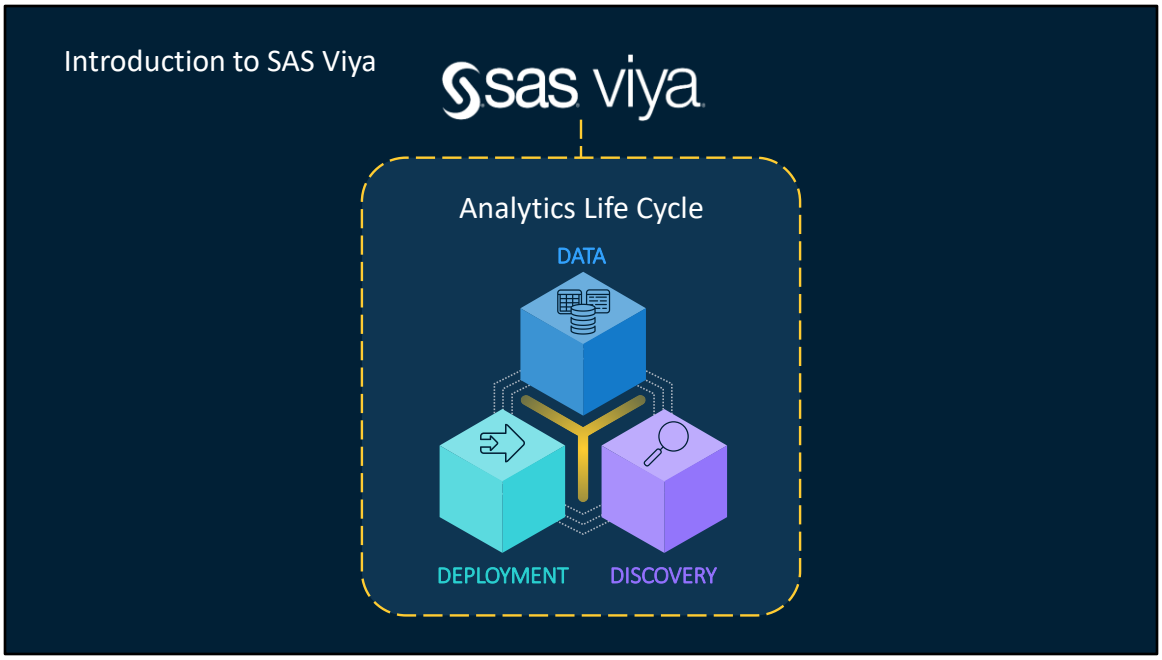
The correct answer is **c**. A model is a concise representation of the relationship between the inputs and the target variables. Option **a** is the definition of estimation/learning. Option **b** is the definition of estimation criterion. Option **d** is irrelevant in analytical vocabulary

1.3 Introduction to SAS Viya and SAS Studio



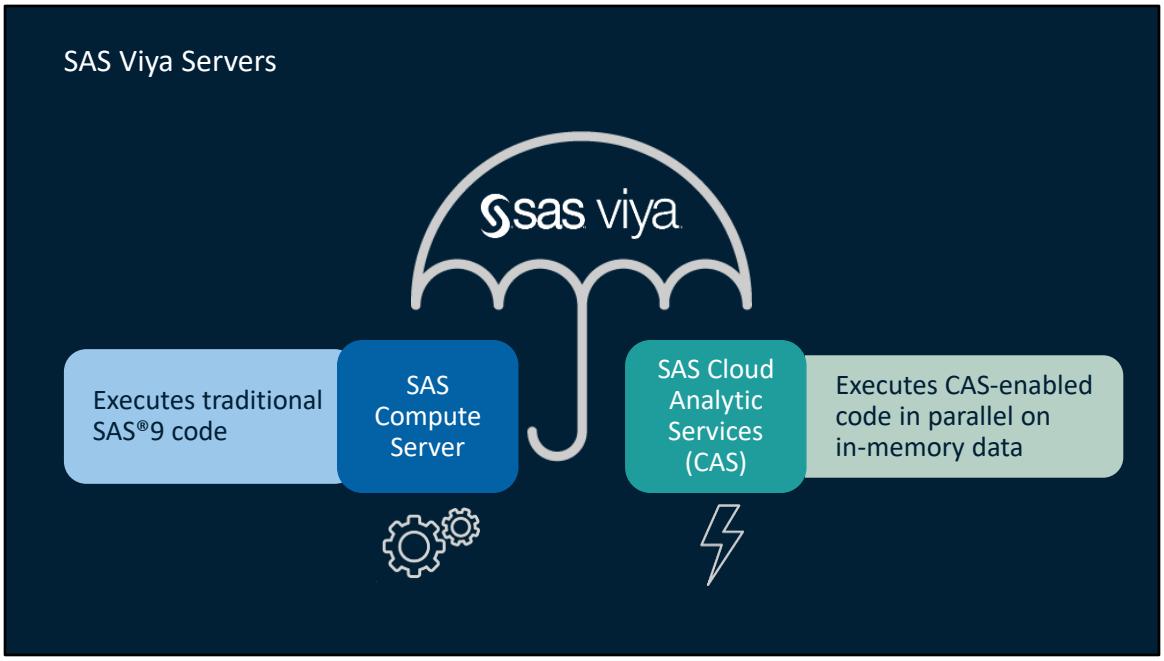
SAS Viya is the latest enhancement of SAS, which is cloud-ready analytics and data management platform. It allows scalable, web-based access for your data processing needs.

SAS Viya offers several products that facilitate parallel task execution. Most computers can execute operations in parallel due to their multicore infrastructure. Performing more than one operation simultaneously has the potential to speed up most tasks and has many practical uses within the field of data science.



Big data demands sophisticated data management and advanced analytics techniques. You can orchestrate your entire analytics journey – from data to tangible results – with a powerful analytics platform.

The "heart" of SAS is the analytics life cycle. The three phases of the analytics life cycle are data, discovery, and deployment. Data are the foundation of everything you do, discovery is the act of finding something that you had not known before, and deployment is where you get the value out of analytics.

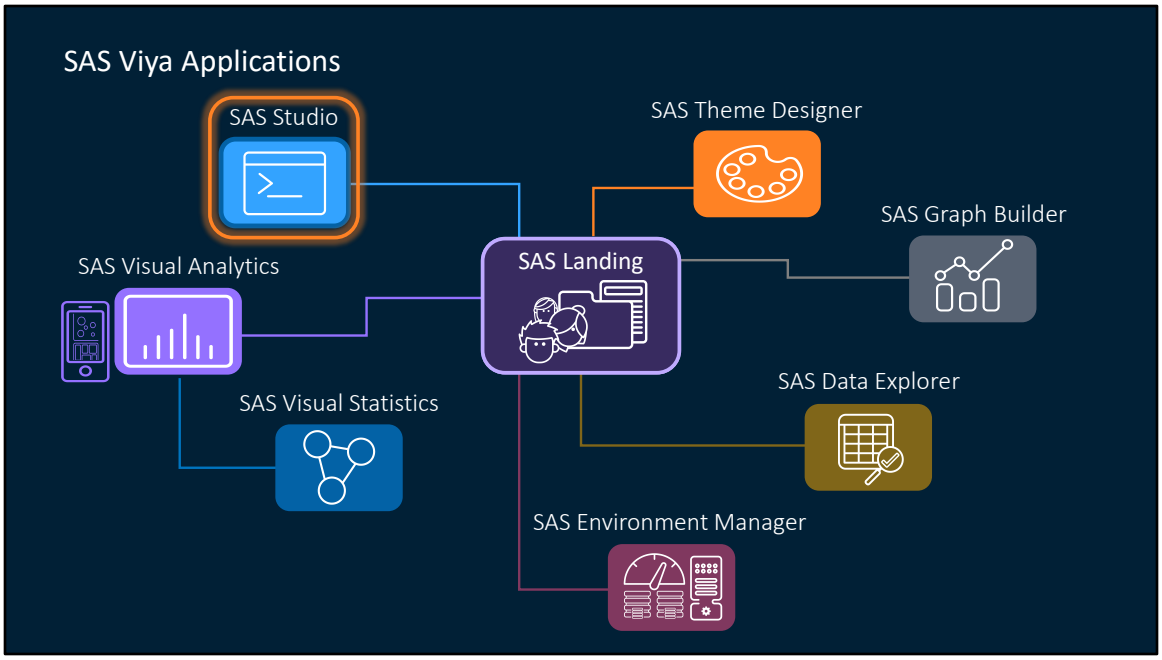


SAS Viya includes multiple servers to execute SAS code. The two primary servers are the SAS Compute Server and SAS Cloud Analytic Services, or CAS. You can think of the SAS Compute Server as equivalent to the SAS®9 workspace server. Your familiar SAS®9 programs can be submitted as is to SAS Viya and, by default, they will run on the Compute Server. There is no need to learn new syntax to use the Compute Server.

SAS Cloud Analytic Services is the high-performance server that performs parallel processing on in-memory data. It is likely that you will use CAS for your big data and complex analytics. Often, only very minor code modifications are required for programs to run in CAS.

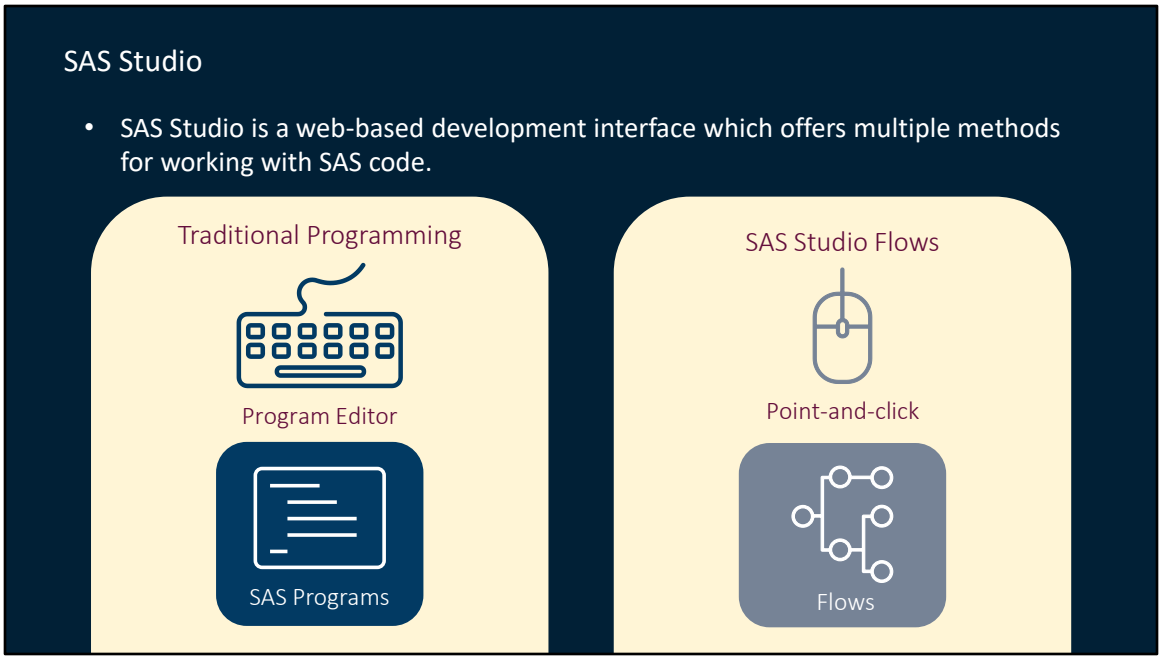
More about SAS Viya

To explore this learning resource in greater detail, please visit the course page on learn.sas.com. There, you'll find comprehensive course materials, tutorials, and updates to support your learning journey.



SAS Viya consists of many different applications that enable you to take advantage of the powerful CAS server. The entry point for SAS Viya is SAS Landing, a collaborative interface for accessing, organizing, and sharing content. This application is the gateway to all other applications in Viya.

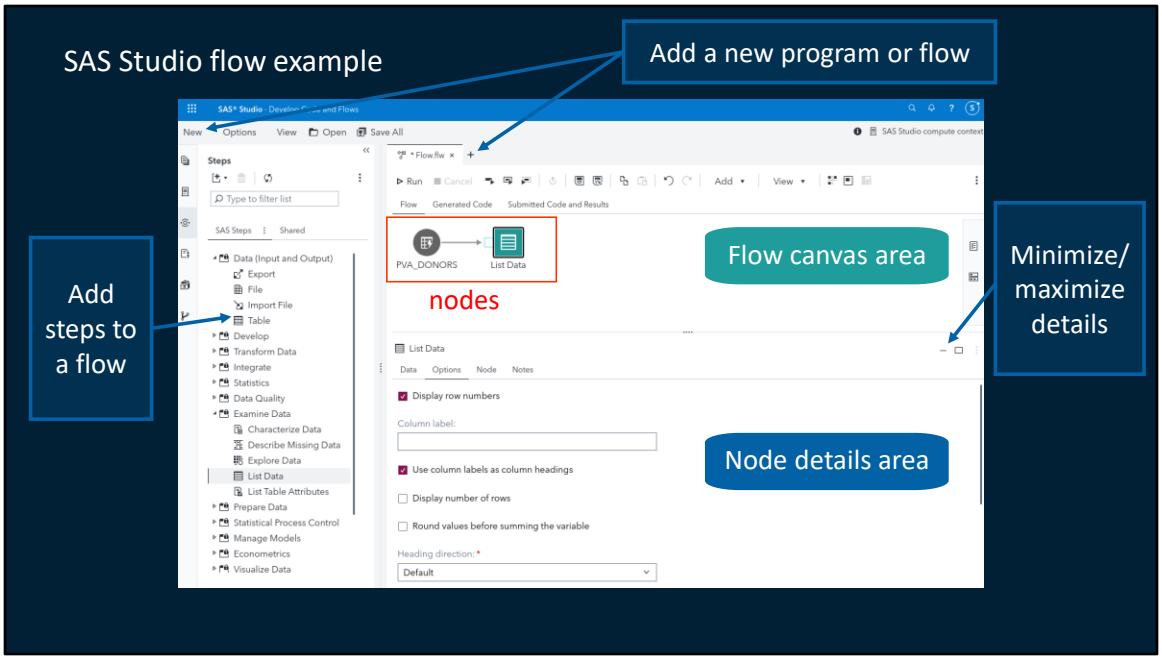
Our focus will be SAS Studio in this course.



SAS Studio is a web-based development interface which offers multiple methods for working with SAS code.

The traditional method is to use the program editor to write SAS code and develop SAS programs.

Another method is to use pre-built point-and-click steps which generate SAS code for tasks such as accessing, analyzing, transforming, or visualizing data. These steps are connected in a dynamic flow chart, allowing you to visualize your data transformation process.



Here is an example flow. A new flow can be created using the plus sign to the right of an open tab or by using the New menu at the top left side of SAS Studio.

A flow is a sequence of operations on data. Data and operations are represented in SAS Studio by steps that can be added to a flow by expanding the Steps menu in the Navigation pane on the left and double clicking on individual steps. The steps will appear as nodes in the Flow canvas area in the center of the open flow tab. The current flow has 2 nodes: a table node labeled PVA_DONORS connected to a List Data node. Options for each node can be set in the Node details area through the point and click interface. The details can be minimized and maximized using the buttons on the upper right side of the details pane.

More information about constructing and executing SAS Studio flows will be shown in the course demonstrations

SAS Studio Flow steps

SAS Step Categories		
Data (Input and Output) ★	Enrichment	Prepare Data ★
Develop (Code)	Econometrics	Examine Data ★
Transform Data	Data Quality	Statistical Process Control
Integrate	Visualize Data	Text Analytics
Statistics ★	Machine Learning ★	Custom Steps
Manage Models	Optimization and Network Analysis	

Flow steps are organized into categories based on the user's desired task.
Specific steps are available to:

- Work with input or output data
- Develop code
- Transform data
- Integrate data
- Generate statistics
- Manage models
- Enrich data
- Work with econometrics
- Enhance data quality
- Visualize data
- Manage machine learning
- Perform optimization and network analysis
- Prepare data
- Examine data
- Execute statistical process control, or
- Complete text analytics

Additionally, SAS Studio offers a separate interface for users to design completely custom steps based on their business needs. Steps within the starred categories will be used to construct flows throughout this course.

Question: SAS Viya

SAS Viya is a cloud-ready analytics and data management platform that facilitates parallel task execution.

a. True
b. False

Answer: SAS Viya

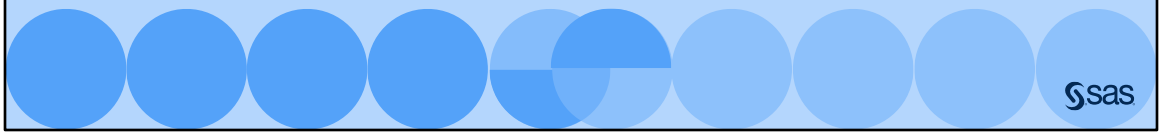
SAS Viya is a cloud-ready analytics and data management platform that facilitates parallel task execution.

a. True
b. False

Answer: True
SAS Viya is the latest enhancement of SAS, which is a cloud-ready analytics and data management platform. SAS Viya offers several products that facilitate parallel task execution.

Lesson Summary

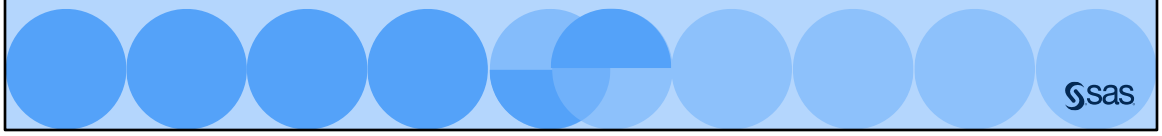
To explore this learning resource in greater detail, please visit the course page on learn.sas.com. There, you'll find comprehensive course materials, tutorials, and updates to support your learning journey.



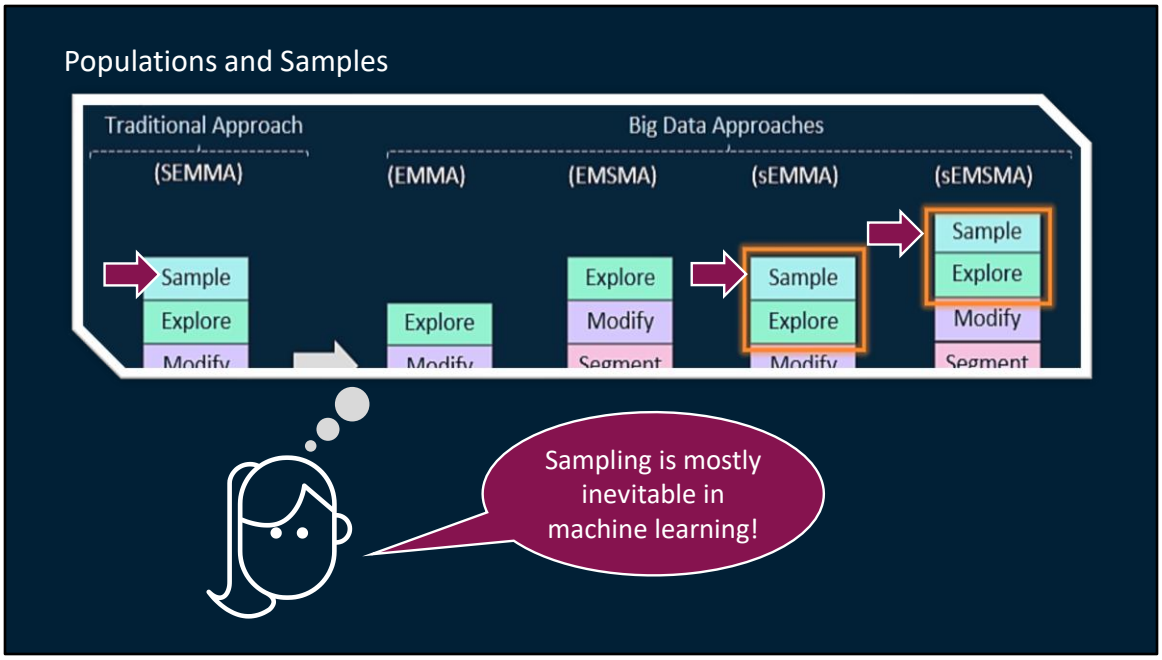
Lesson 2: Fundamental Statistical Concepts

Lesson Overview

To explore this learning resource in greater detail, please visit the course page on learn.sas.com. There, you'll find comprehensive course materials, tutorials, and updates to support your learning journey.

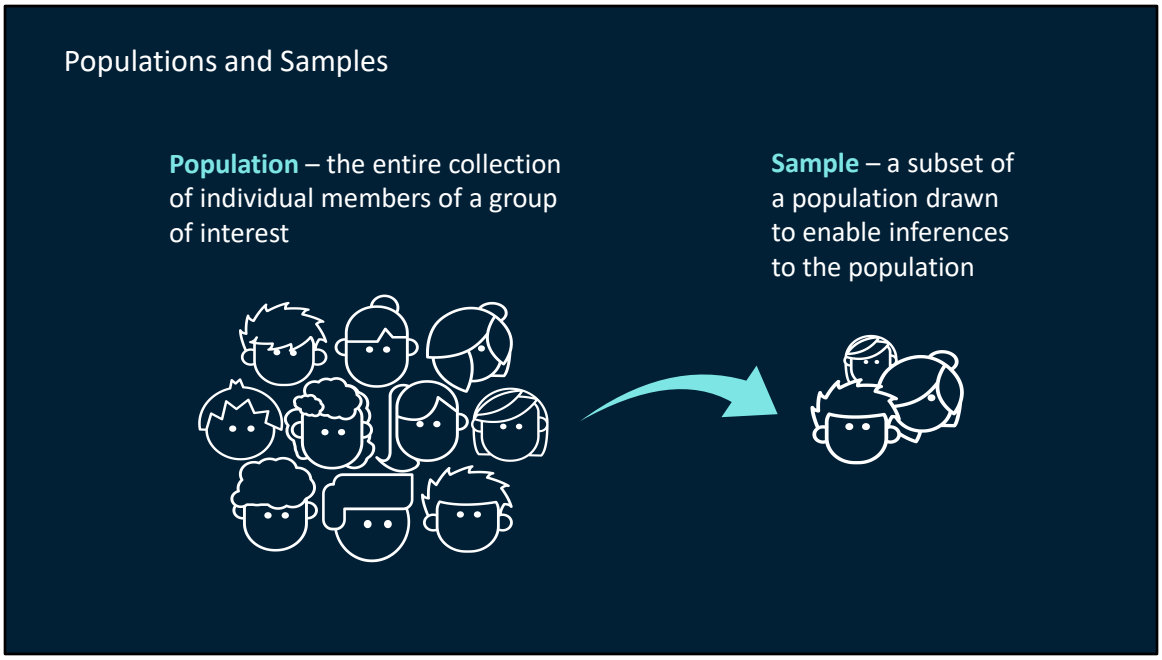


2.1 Introduction to Statistical Analysis



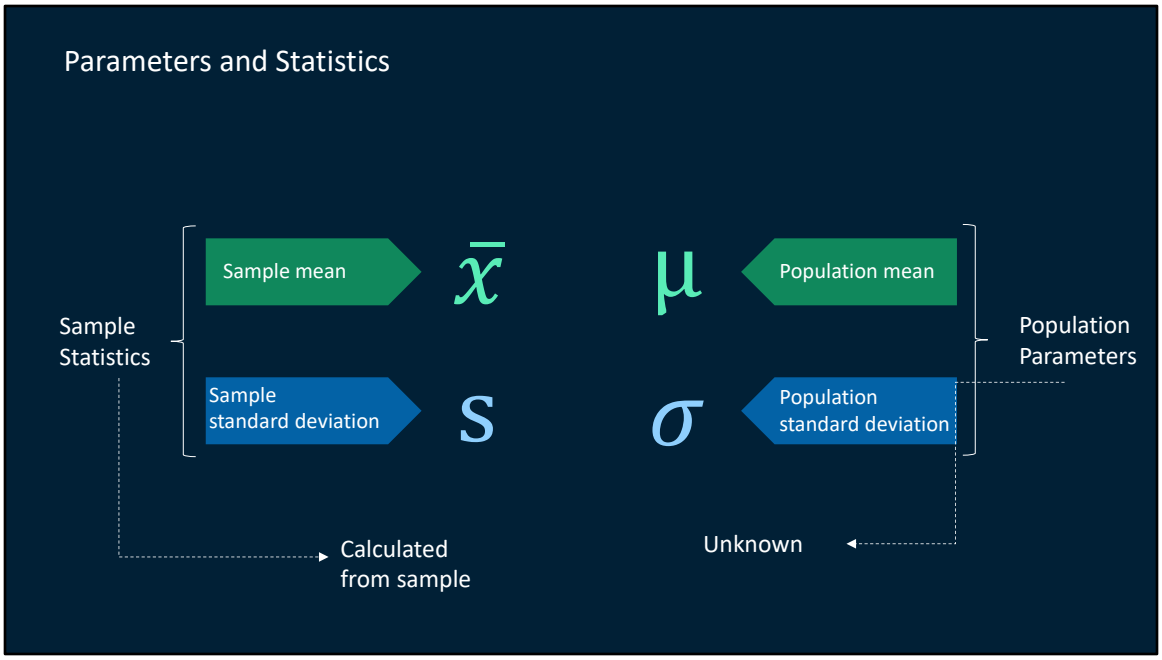
Recall the difference between traditional modeling approaches and big data modeling approaches that we discussed in the previous lesson. Even though SAS Viya can handle big data, it's not entirely correct to say that you don't need to sample when you do machine learning on big data.

Preparing your data for analysis starts with exploring your data. Exploring the data can be time-consuming if processed using the entire data. So, sampling is inevitable.



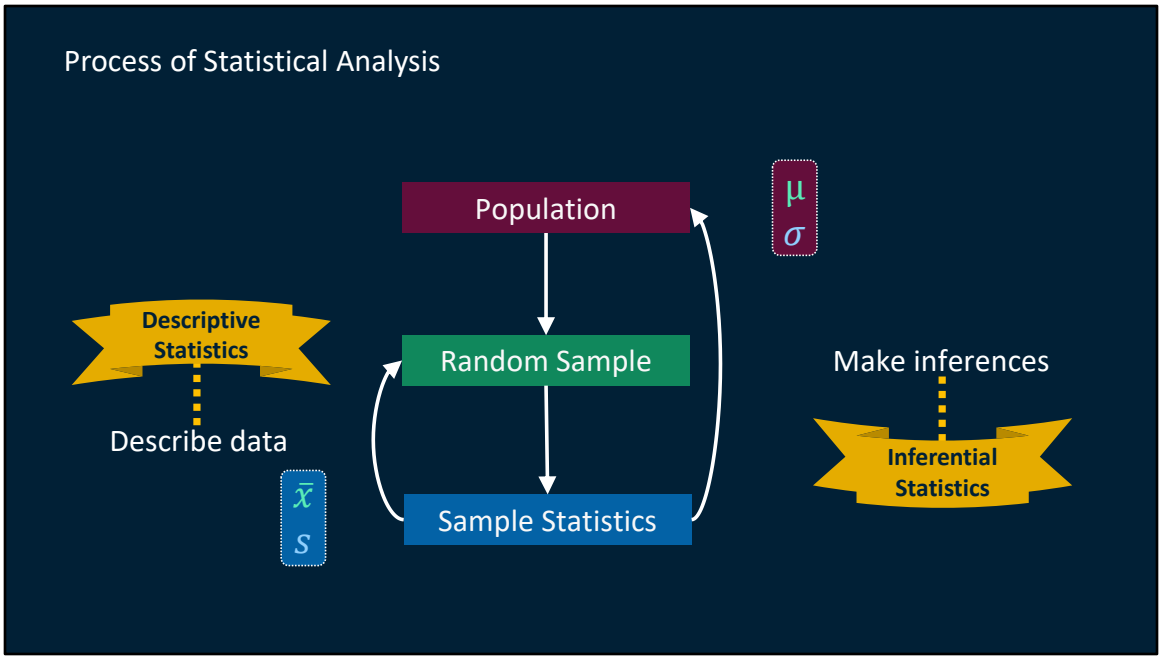
A *population* is a collection of *all* objects about which information is desired. Examples of populations are all people with a certain disease, all drivers with a certain level of insurance, or all customers, both current and potential, at a bank. When population sizes are too large to include all possible observations in a statistical test, samples are used instead.

A *sample* is a subset of the population. The sample should be representative of the population, meaning that the sample's characteristics are like the population's characteristics. A thousand bank customers responding to a survey is an example of sample.



A descriptive measure of the population is called a *parameter*. *Parameters* are essentially the characteristics of population. Because populations usually cannot be measured in their entirety, parameter values are generally unknown. Parameters are usually denoted by Greek letters. Examples of parameters are population mean (μ), and population standard deviation (σ).

A descriptive measure of a sample is called a *statistic*. Statistics are usually denoted by Roman letters. Examples of statistics are sample mean ($\bar{x}$) and sample standard deviation (s). Statistics are quantities calculated from the values in the sample.



Statistical analysis is the process of collecting and analyzing data in order to draw meaningful conclusions from the raw data. You generally start with identifying the population of interest. And you often want to estimate the value of a parameter or conduct tests about the parameter, for example, μ and σ. However, the calculation of parameters is usually either impossible or infeasible because of the amount of time and money required to reckon the full population. In such cases, you take a random sample of the population, compute sample statistics to describe the sample, such as $\bar{x}$ and *s*, and use sample information to make inferences about the population by estimating the value of the parameter. The basis for inferential statistics, then, is the ability to make decisions about parameters without enumerating the entire population.

This leads us to describe the two branches of statistics: descriptive statistics and inferential statistics.

If you are using data gathered on a group to describe or reach conclusions about the same group, the statistics are called *descriptive statistics*. Descriptive statistics are used to organize, summarize, and focus on the main characteristics of your data. Summarizing your data in such a manner also makes it more usable. Here, $\bar{x}$ and *s* are descriptive statistics.

If you gather data from a sample and use the statistics generated to reach conclusions about the population from which the sample was taken, the statistics are *inferential statistics*. Inferential statistics make generalizations or inferences from your data to a larger set of data, based on probability theory. Here, $\bar{x}$ estimates μ, and *s* estimates σ.

Question: Samples

A sample consists of which of the following?

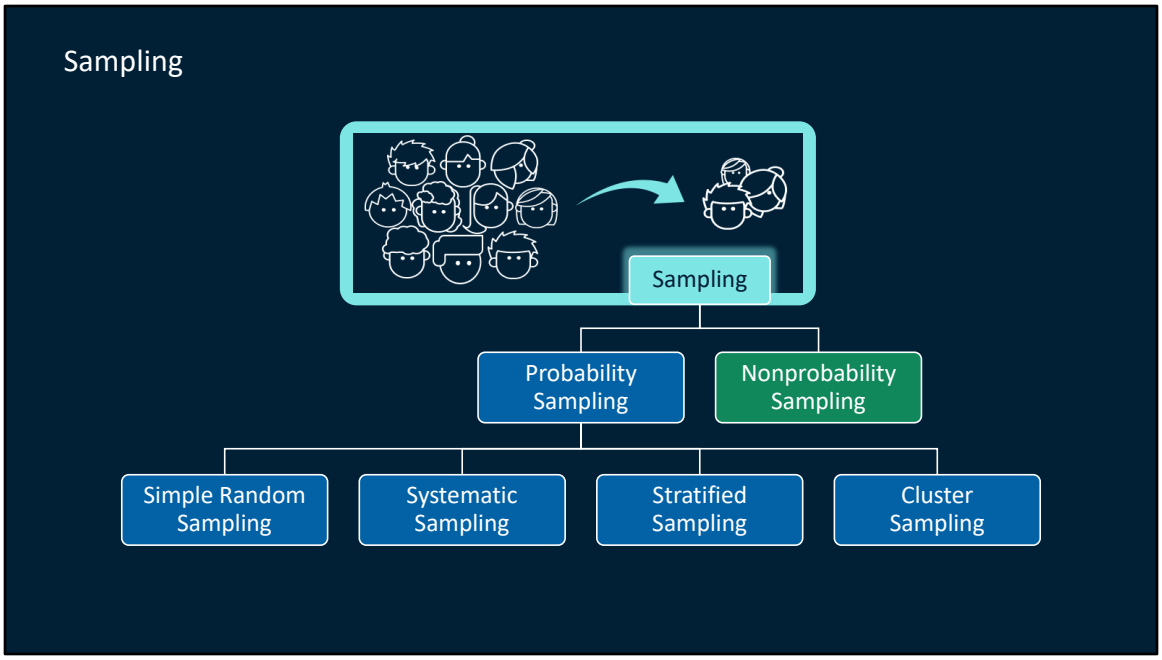
a. all units of the population
b. at least 20% units of the population
c. at the most 50% units of the population
d. any fraction of the population

Answer: Samples

A sample consists of which of the following?

a. all units of the population
b. at least 20% units of the population
c. at the most 50% units of the population
d. any fraction of the population

The correct answer is d.
A sample is a smaller part of the whole, that is, a subset of the population. A sample can be any fraction of the population.



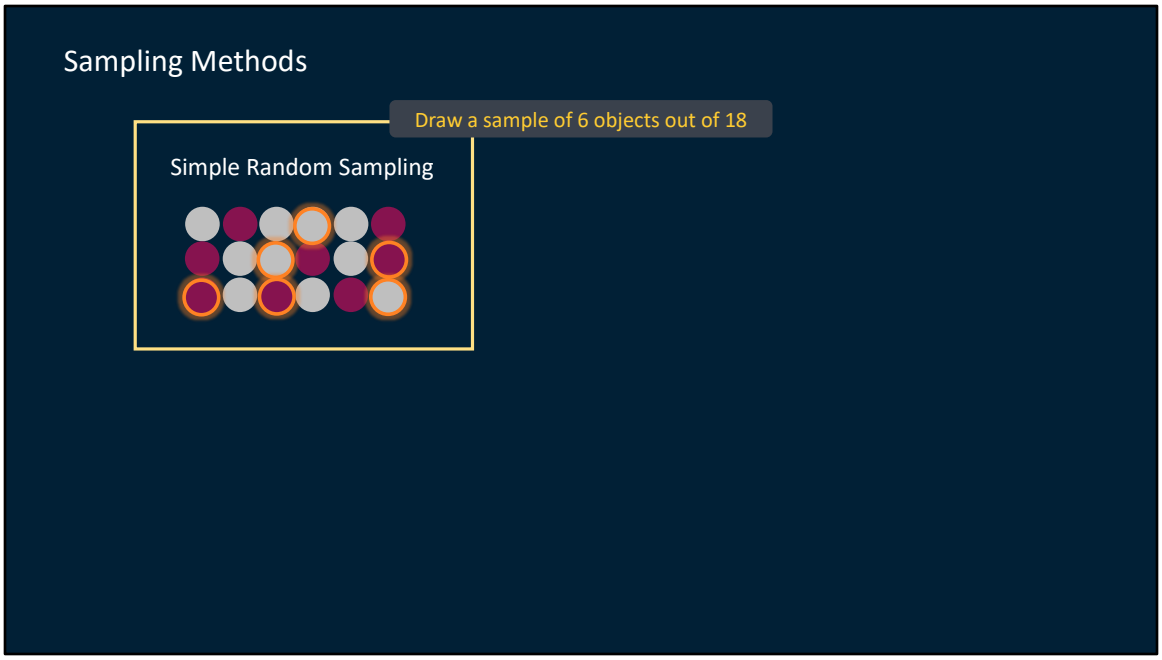
The process used to extract a sample from a population is called *sampling*. Sampling is broadly classified into *probability* and *non-probability* sampling. In probability sampling, each member of the population has a fixed, known opportunity to belong to the sample, which is not the case with non-probability sampling.

In this course, we focus on four basic sampling techniques: simple random sampling, systematic sampling, stratified sampling, and cluster sampling. These sampling techniques are useful in the machine learning world.

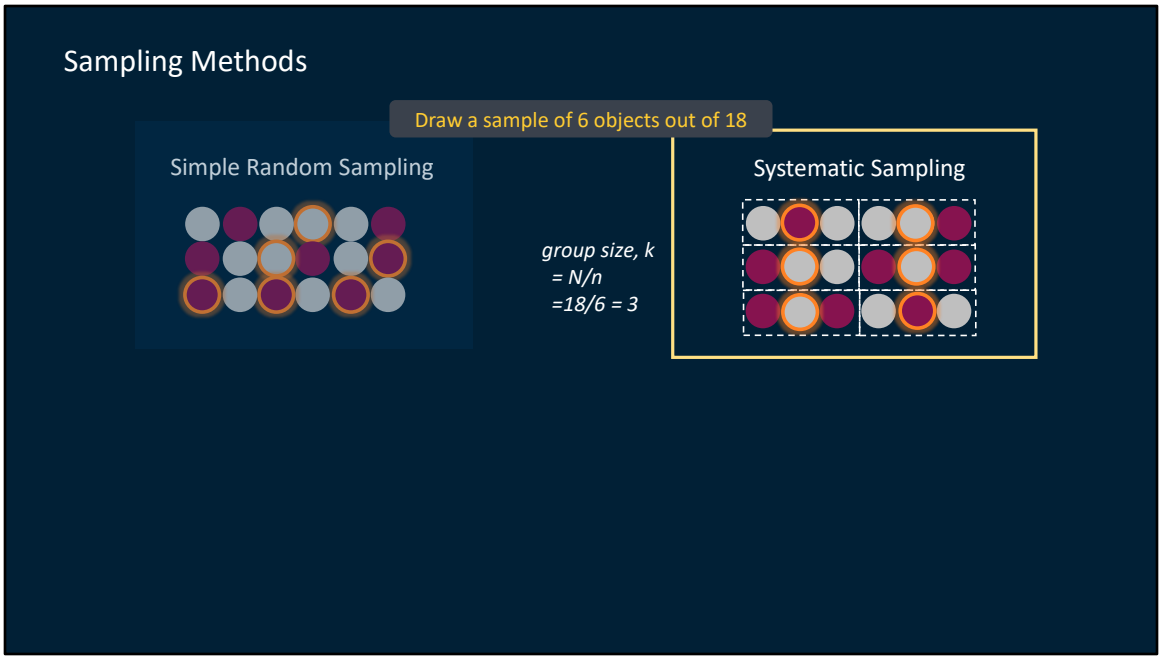
Sampling Methods

Draw a sample of 6 objects out of 18

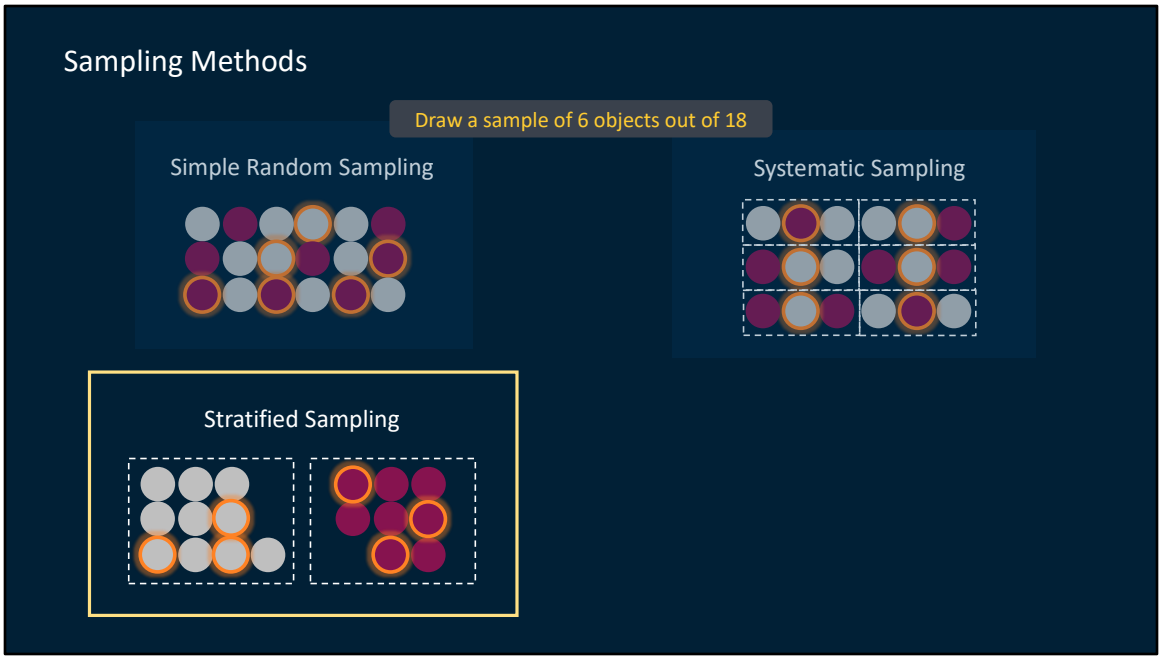
Consider that you have a population of 18 objects of two types, the colors gray and plum. You want to draw a sample of six objects.



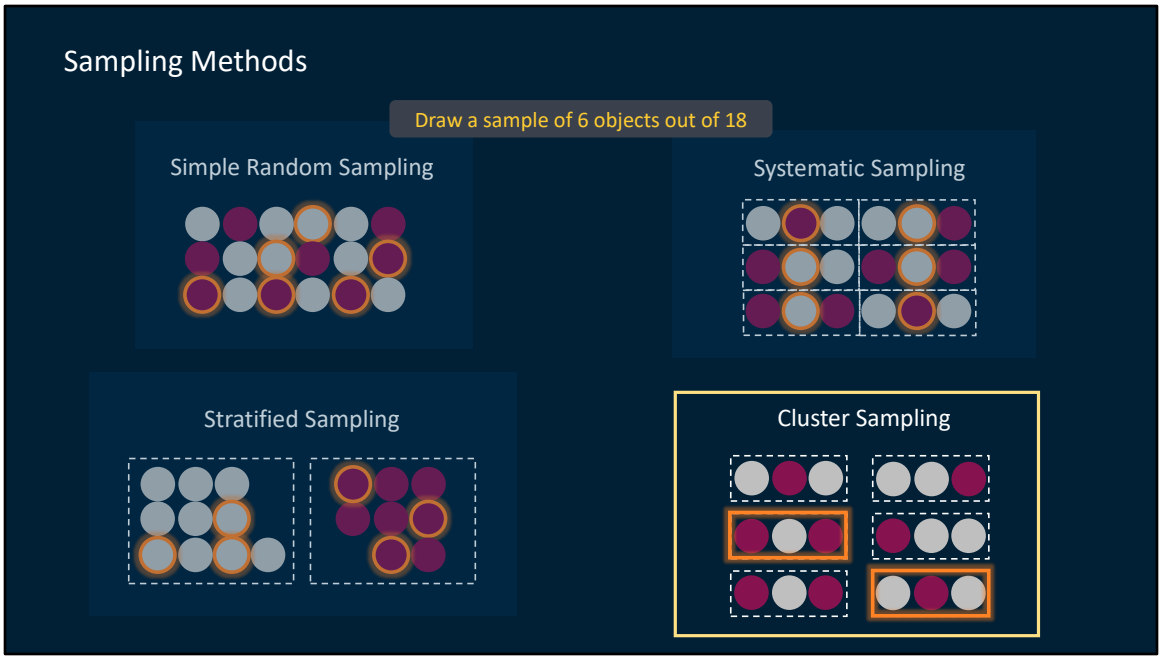
The most elementary of the probability sampling is *simple random sampling.* Simple random sampling is a technique in which each element of the population has an equal probability of being selected. All the elements are selected independent of each other. Six objects are randomly selected without considering the color of the objects.



Systematic sampling involves the selection of objects from an ordered sampling frame. You decide on sample size (*n*) and then divide frame of *N* individuals into groups of *k* individuals where *k*= N/n. It randomly selects a starting point, and then every kth element is selected from the population. For example, here 18 objects are arranged into an ordered sampling frame of six groups, each one of them of size 3. The second object is randomly selected from the first frame and then every third object is selected from the ordered sampling frame.



Stratified sampling is a type of random sampling in which the population is divided into various non-overlapping strata and then elements are randomly selected into the sample from each stratum. Here, the dots are divided into two strata - gray objects and plum objects. Three objects are randomly selected from each stratum.

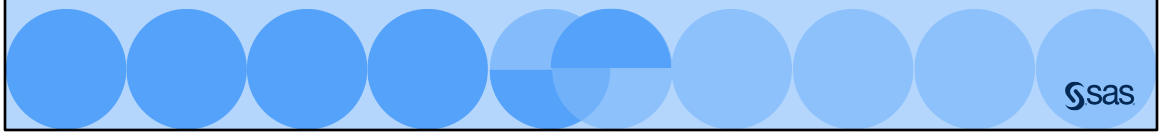


Unlike stratified sampling, *cluster sampling* divides the population into several clusters such that each cluster is representative of the population and then some of the clusters are randomly selected for the sample. Here, the population is divided into six clusters each comprised of gray and plum objects. Then two clusters are randomly selected giving a sample of six objects.

In cluster sampling, the sampling is done on a population of clusters. Therefore, cluster is considered a sampling unit. In stratified sampling, elements within each stratum are sampled. Stratified sampling has homogeneity within strata but heterogeneity between strata whereas cluster sampling has heterogeneity within clusters but homogeneity between clusters.

More about Sampling

To explore this learning resource in greater detail, please visit the course page on learn.sas.com. There, you'll find comprehensive course materials, tutorials, and updates to support your learning journey.



Question: Sampling Methods

In FM radio markets, age of listener is an important determinant of the type of programming used by a station. If a random sample is extracted from each of the three subpopulations of listeners (below 25 years, 25-40 years, and above 40 years), what type of sampling is this?

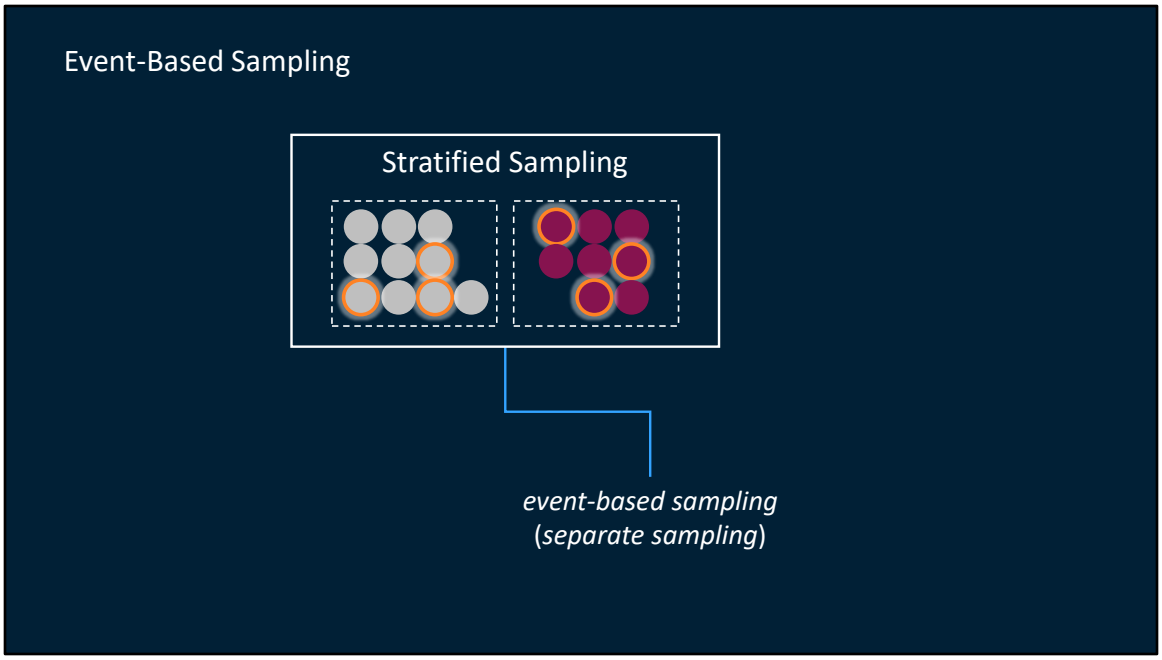
a. simple random sampling
b. stratified sampling
c. systematic sampling
d. cluster sampling

Answer: Sampling Methods

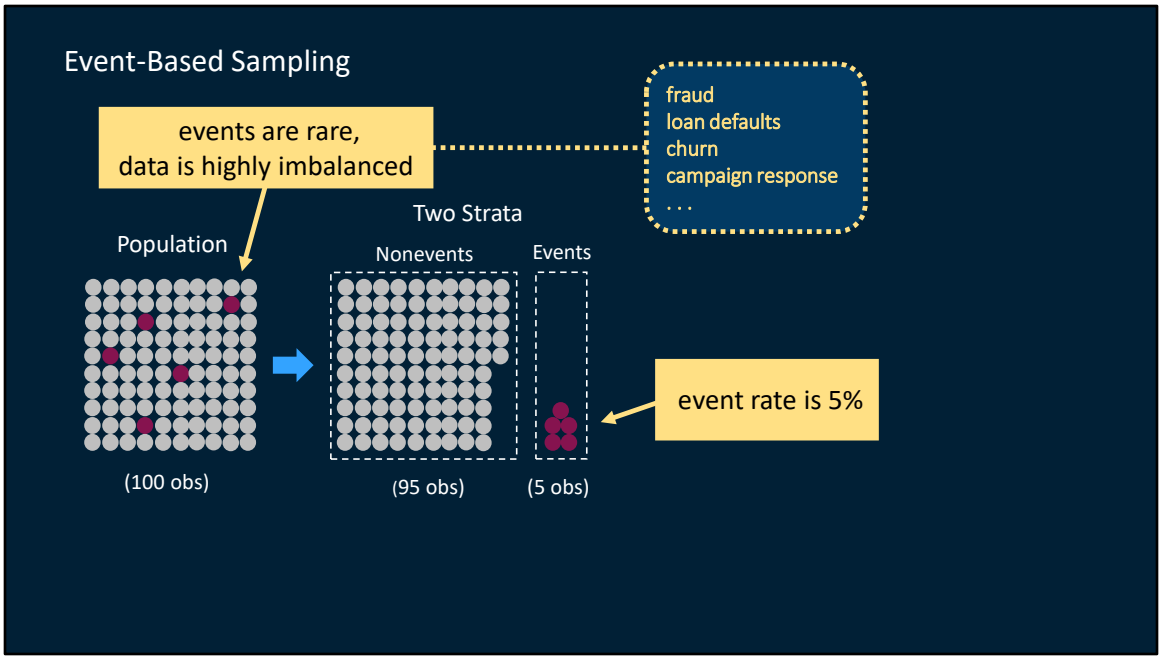
In FM radio markets, age of listener is an important determinant of the type of programming used by a station. If a random sample is extracted from each of the three subpopulations of listeners (below 25 years, 25-40 years, and above 40 years), what type of sampling is this?

a. simple random sampling
b. stratified sampling
c. systematic sampling
d. cluster sampling

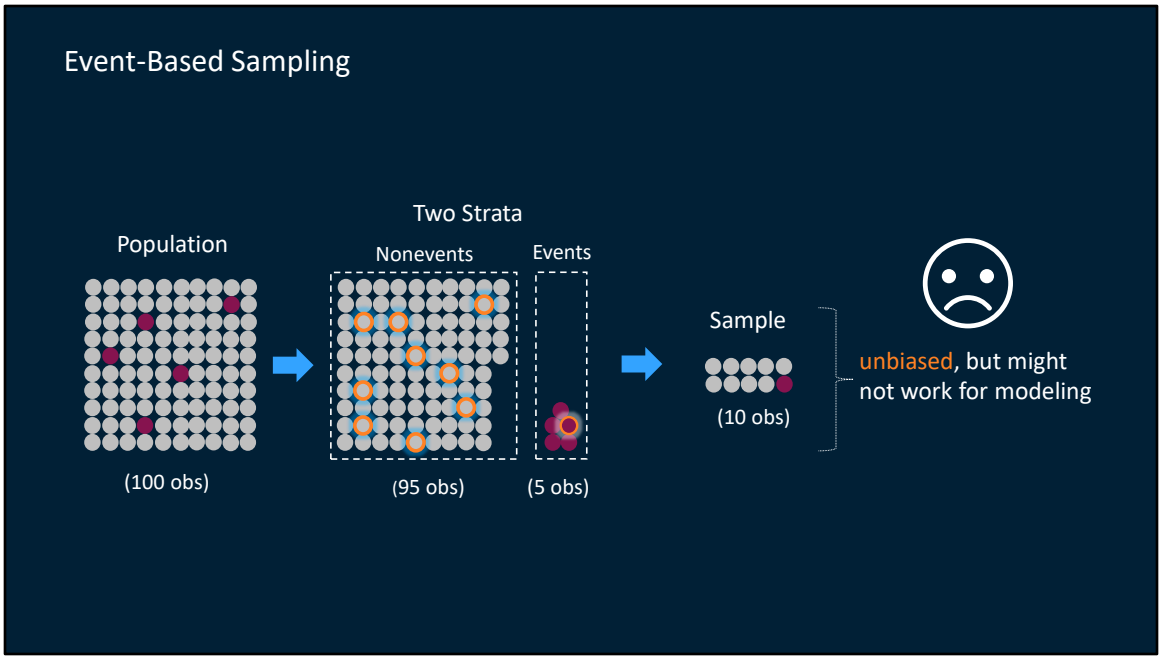
Correct Answer: b
In stratified sampling, the population is divided into nonoverlapping subpopulations called *strata*. The researcher then extracts a random sample from each of the subpopulations.



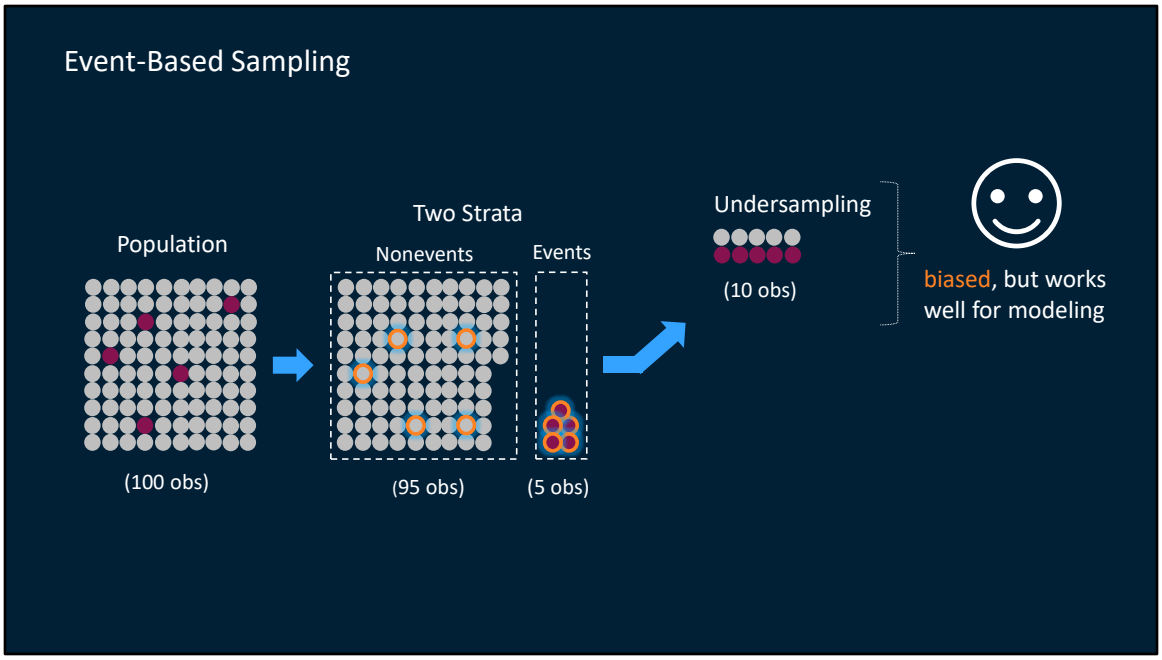
One type of stratified sampling, called *event-based sampling* or *separate sampling,* is extremely important in machine learning predictive models.



Consider a binary target with categories logically labeled as events (the plum dots) and nonevents (the gray dots). We are interested in building a predictive model with a focus to predict the event. However, events are rare and thus the data is highly imbalanced. Most of the machine learning problems have rare events like predicting fraud, loan defaults, customer churn, and campaign response. For example, the actual event rate is 5%. That means that if there are 100 observations in the population, only five of them are events and 95 are nonevents.

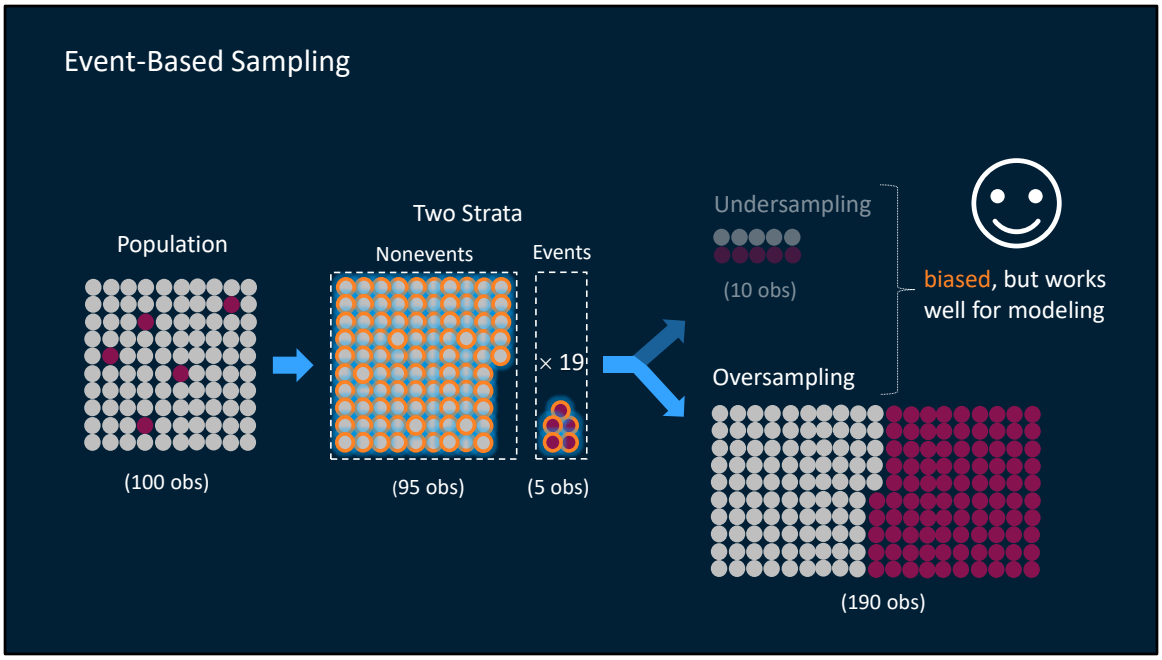


Event-based samples also known as target-based samples are created by considering the events separately from the nonevents. There are two strata and target-based stratified sampling can be used. A natural practice is to build models from a sample with the same event proportion as there is in the population. For example, to draw a sample of 10 observations, one observation is randomly selected from events and nine observations are randomly selected from nonevents. In this sample, the nonevents far outweigh the events. Although this type of sample would be unbiased, if we train a binary classification model without fixing this problem, the model will be completely biased toward nonevents. For example, a trivial model that predicts every case as a nonevent would be as accurate as 95%. But always predicting nonevents does not help because our primary interest is to build a model to predict the events and not the nonevents.

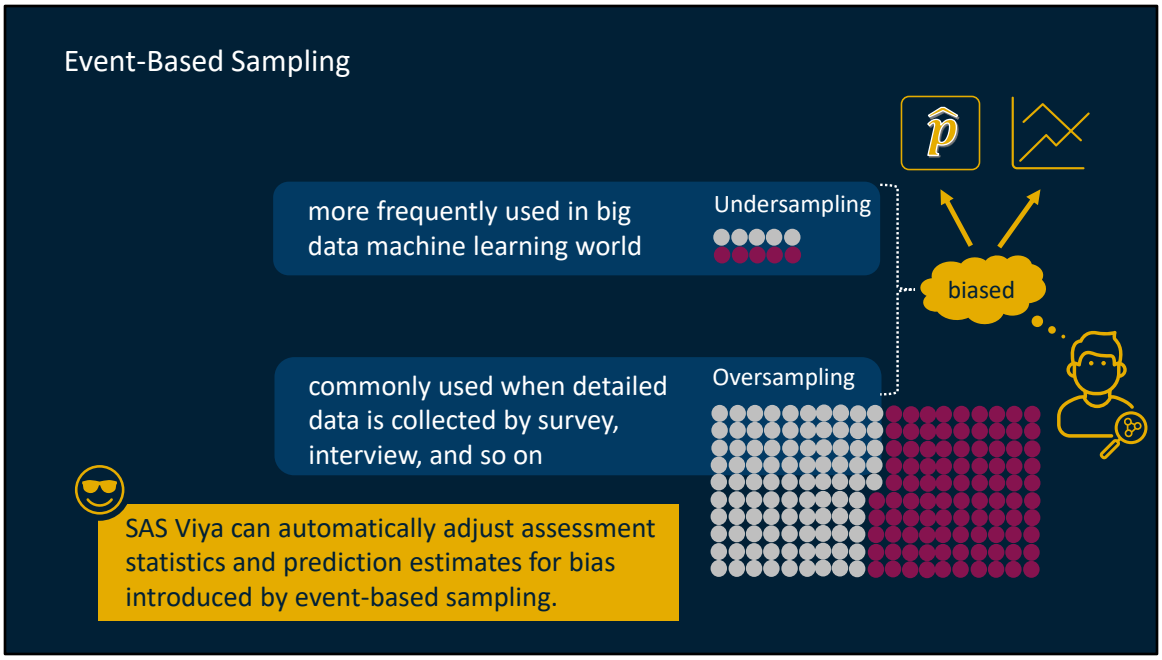


A common predictive modeling practice is to build models from a sample with an event proportion different from the true population proportion. This is typically done when the ratio of events to nonevents is small. Separate sampling derives its name from the technique used to generate the modeling data, that is, samples are drawn separately based on the target outcome. There are two ways to perform event-based sampling – undersampling and oversampling, both of which have their pros and cons.

Undersampling is deleting samples from the majority class. In the case of a rare event, usually all events are selected. Then, each event is matched by one or (optimally) more nonevents. For example, to draw a sample of 10 observations, all five events are selected, and an equal number of nonevents are randomly selected from nonevents. In this sample, the events and nonevents are balanced.



Oversampling, on the other hand, is duplicating the observations of the minority class to obtain a balanced data set. You can select random examples from the minority class with replacement, and supplement the training data with multiple copies of this observation. Hence, it is possible that a single observation can be selected multiple times. Here, a sample of 190 observations is drawn by selecting all the nonevents and including 19 copies of events to make it a balanced sample.



Oversampling duplicates or creates new synthetic examples in the minority class, whereas undersampling deletes or merges examples in the majority class. Both types of resampling can be effective when used in isolation. Note that some sources describe undersampling as "oversampling" since the event of interest is over-represented in the sample.

Undersampling is generally employed more frequently than oversampling in the big data machine learning world. Overabundance of already collected data is usually encountered in the "big data" scenario, and the reasons to use undersampling are mainly practical and related to resource costs. Specifically, while you need a suitably large sample size to draw valid statistical conclusions, the data must be cleaned before it can be used. Cleansing typically involves significant efforts and time and are therefore costly. So, smaller samples resulting in undersampling are going to be more useful.

Conversely, oversampling is commonly used when the detailed data has yet to be collected by survey, interview, or otherwise. Because of the tremendous cost involved in collecting such information, undersampling is employed much less frequently.

Despite its advantages, event-based sampling (whether oversampling or undersampling) is fundamentally biased and thus introduces some analysis complications.

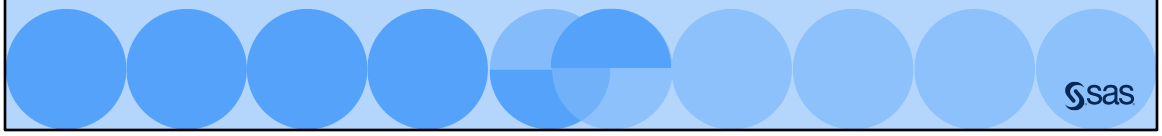
First, most model fit statistics, especially those related to prediction decisions (like misclassification rate), and some of the assessment plots (like cumulative lift) are closely tied to the outcome proportions in the training samples. If the outcome proportions in the training and validation samples do not match the outcome proportions in the scoring population, model performance can be greatly misestimated.

Second, if the outcome proportions in the training sample and scoring populations do not match, model prediction estimates are biased.

Fortunately, SAS Viya can adjust assessments and prediction estimates to match the scoring population if you specify *prior probabilities*, the scoring population outcome proportions.

Oversampling and Undersampling: Advantages and Disadvantages

To explore this learning resource in greater detail, please visit the course page on learn.sas.com. There, you'll find comprehensive course materials, tutorials, and updates to support your learning journey.



Question: Event-Based Sampling

Both oversampling and undersampling involve introducing bias by having more cases from one class than from another, to compensate for an imbalance representation in the data.

a. True
b. False

Answer: Event-Based Sampling

Both oversampling and undersampling involve introducing bias by having more cases from one class than from another, to compensate for an imbalance representation in the data.

a. True
b. False

Correct answer: True
Undersampling and oversampling are techniques used to combat the issue of unbalanced classes in a data set. Undersampling induces bias by deleting samples from the majority class, and oversampling induces bias by duplicating samples from the minority class.

Analysis Data

A veterans' organization seeks continued contributions from lapsing donors.

Population

- 3.5 million individuals
- actual response rate of approximately 5%

Sample

- 9,686 individuals selected from previous year's campaign
- sample balances response and non-response rate

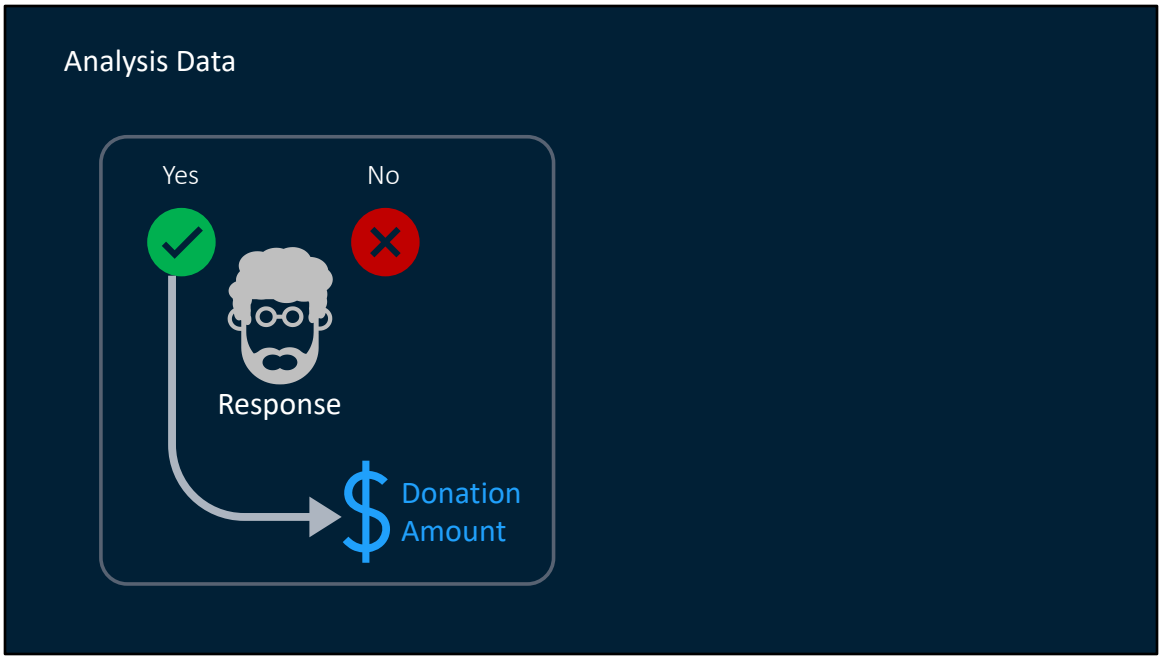
Source: ACM 1998 KDD-Cup competition. Archive at http://kdd.ics.uci.edu/.

A national veterans' organization wants to better target its solicitations for donation. By soliciting only the most likely donors, the organization can spend less money on solicitation efforts and more money on charitable concerns.

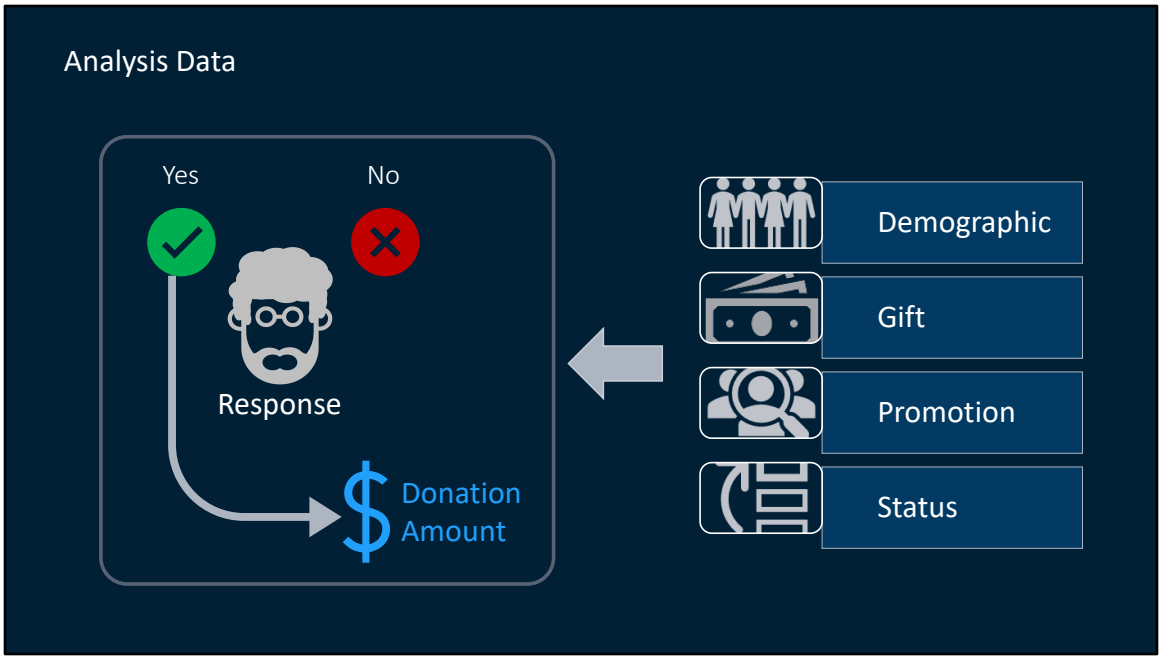
The organization has a mailing database of 3.5 million individuals. These individuals are classified by their response behaviors to previous solicitation efforts. In the original campaign, the response rate was approximately 5%.

The PVA_DONORS table is extracted from a previous year's campaign and is a sample containing 9,686 observations. The data were sampled using event-based sampling so that there are an equal number of responders and non-responders.

For details about the data, see the Data Dictionary in the course overview.



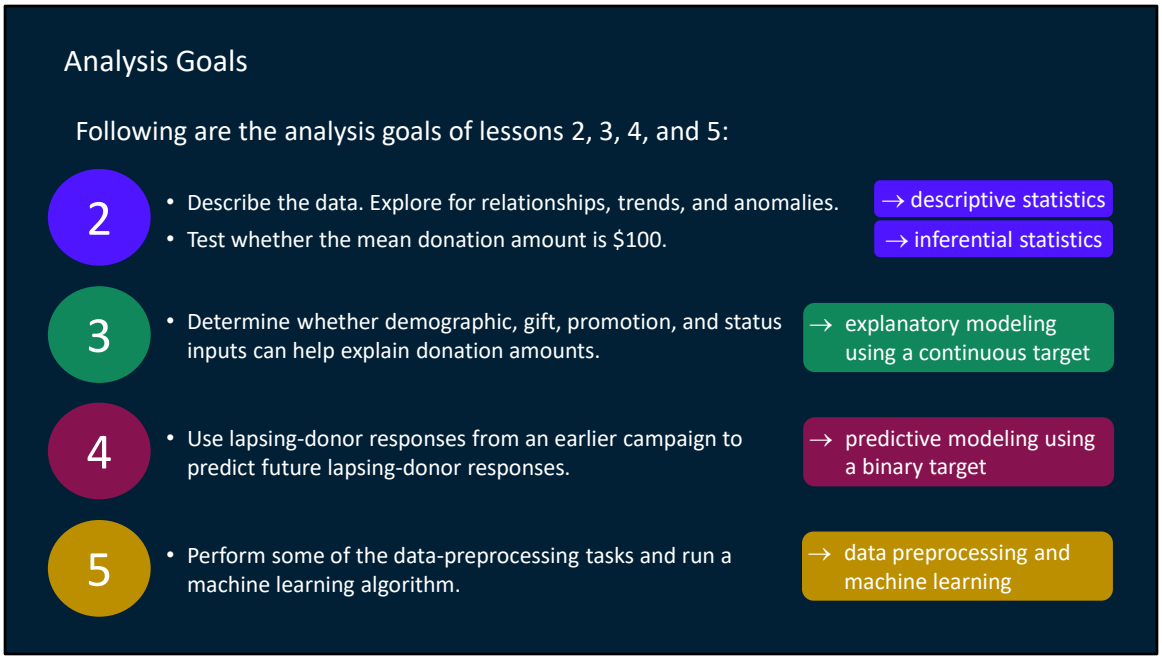
The PVA_DONORS data contains two target variables. The binary variable **Response** is a class target that has two levels, *Yes* and *No*. The interval target variable **Donation_Amt** is the value that represents the dollar amount that an individual donates and is associated with those who have responded to the campaign.



Along with target measurements, several other categorical and continuous measurements are available.

The inputs can be divided into four groups:

- The demographic inputs provide demographic characteristics about the donors and their neighborhoods.
- The gift inputs summarize the previous donation history, providing information about the frequency, size, and recency of past donations.
- The promotion inputs describe the frequency of previous solicitations.
- The status variables categorize individuals into groups, based on recency and frequency measures.



Let's take a moment to look at the analysis goals for the course.

In Lesson 2, we describe the data and explore for anticipated relationships, unanticipated trends, and anomalies in order to gain understanding and ideas. We then test whether the mean donation amount is $100.

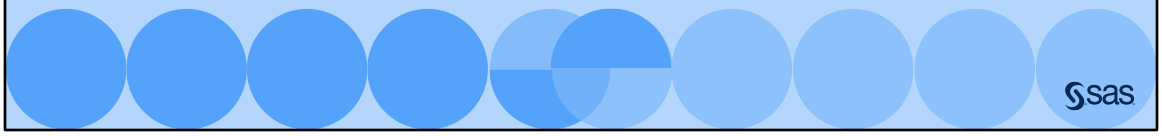
In Lesson 3, we determine whether demographic, gift, promotion, and status inputs can help explain donation amounts from an earlier campaign.

In Lesson 4, we use lapsing-donor responses from an earlier campaign to predict future lapsing-donor responses.

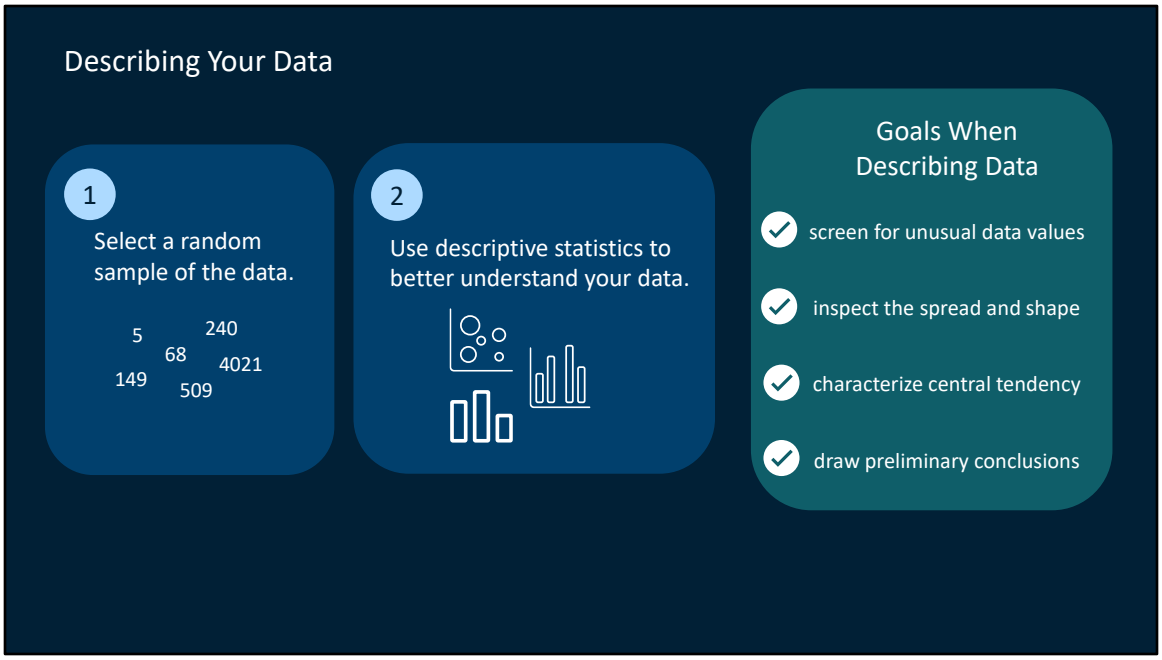
Finally, in Lesson 5, we run a machine learning algorithm with some of the data-preprocessing tasks.

Demo: Listing Observations

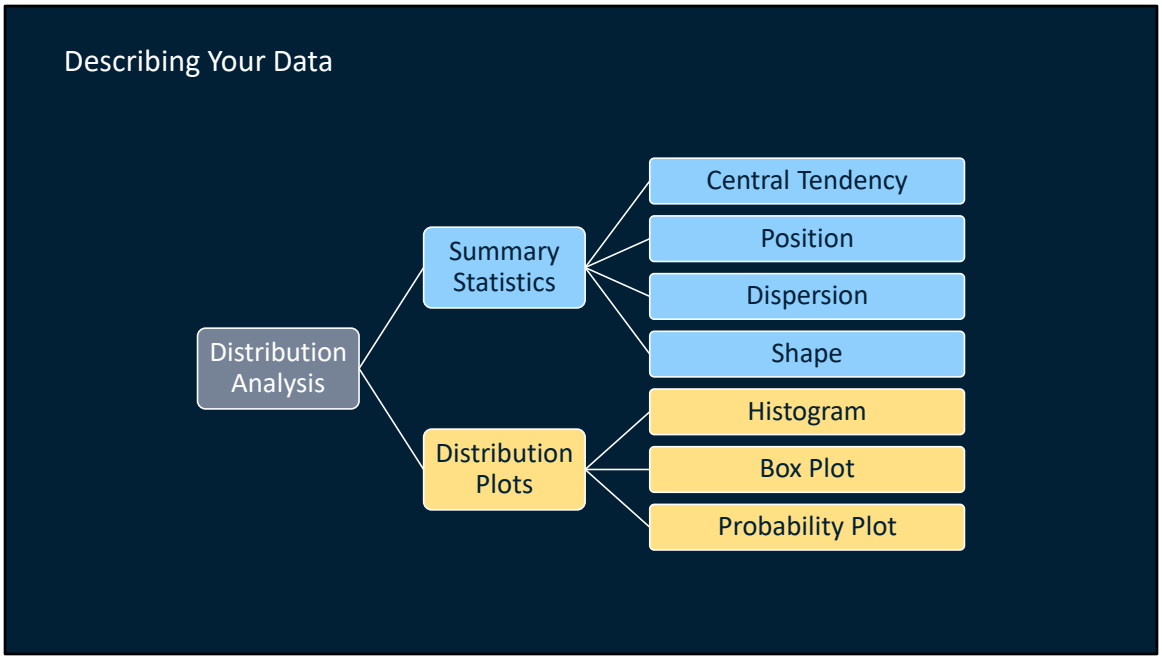
To explore this learning resource in greater detail, please visit the course page on learn.sas.com. There, you'll find comprehensive course materials, tutorials, and updates to support your learning journey.



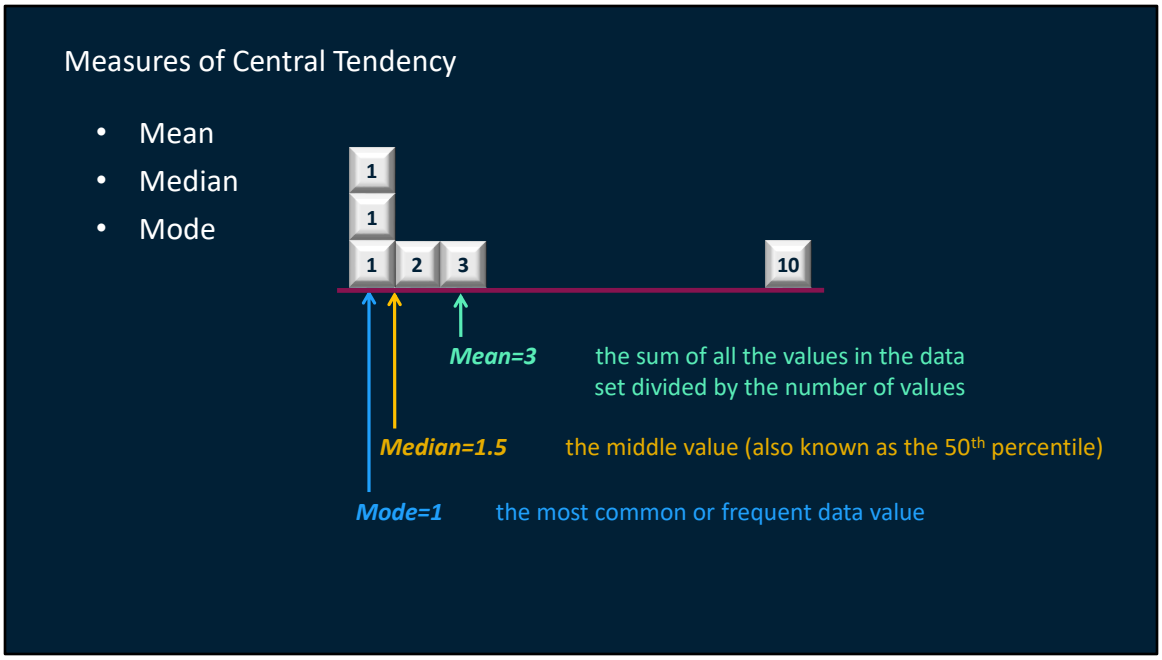
2.2 Descriptive Statistics



After you select a random sample of the data, you can start exploring the data. Although you want to draw conclusions about your population, you first want to describe your data before you use inferential statistics. Descriptive statistics help you to better understand your data by describing and summarizing its basic features. The goals when you're describing data are to screen for unusual data values, inspect the spread and shape of your data, characterize the central tendency, and draw preliminary conclusions about your data.



Distribution analysis typically includes examining the distribution of values of a continuous variable, the range of possible data values, whether the data values accumulate in the middle of the distribution or at one end, and the shape of the distribution. You use summary statistics and distribution plots for distribution analysis. Summary statistics are numerical statistical measures to describe the data, and distribution plots are graphical displays of data's distribution. Graphical techniques, such as histograms, box plots, and probability plots, enable you to make some general observations about the shape and spread of the data. A more complete understanding of the data can be attained by summarizing the data using statistics. This section presents such statistical measures, including measures of central tendency, measures of position, measures of dispersion, measures of shape, and picturing your distribution.

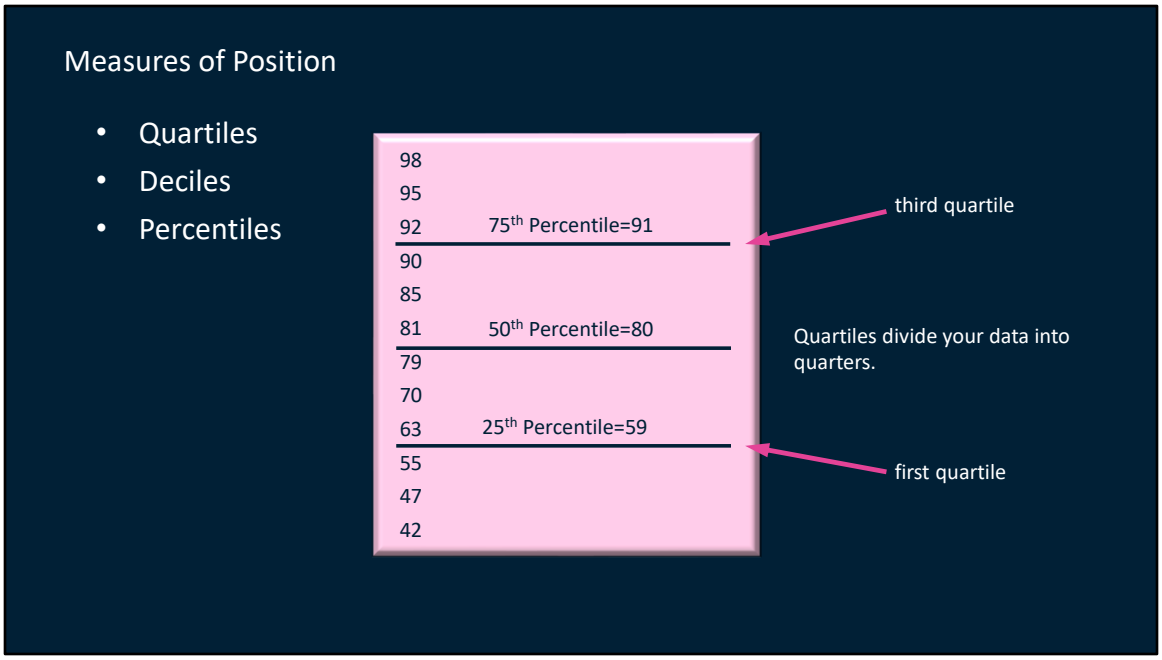


Descriptive statistics that locate the center of your data are called *measures of central tendency*. The most common measure of central tendency is the sample mean. Other measures include median and mode. Consider a simple data set with six observations

The *mean* is the arithmetic balancing point of your data. It is most appropriate for variables measured on an interval or ratio scale with an approximately symmetrical distribution. Mean is affected by each value in the data set, including the extreme values or outliers. A more robust measure of central tendency that is not affected by extreme values is median.

The *median* is the data point in the middle of a sorted sequence. It is appropriate for either rank scores (variables measured on an ordinal scale) or variables measured on an interval or ratio scale with a skewed distribution. Another measure of central tendency that is not affected by extreme values is mode.

The *mode* is the data point that occurs most frequently. It is most appropriate for variables measured on a nominal scale. There might be several modes in a distribution or there might not be any mode in a distribution.



Measures of position provide you a way to see where a certain data point or value falls in a distribution. A measure can tell us whether a value is about the average or whether it's unusually high or low. Like the median divides the data in two equal halves, quartiles divide the data in four equal parts, deciles divide the data in ten equal parts, and percentiles divide the data in hundred equal parts. *Percentiles* locate a position in your data larger than a given proportion of data values. Commonly reported percentile values are the 25th percentile (also called the first quartile), the 50th percentile (also called the median), and the 75th percentile (also called the third quartile).

Question: Central Tendency

You have the data arranged in ascending order. Which one of the following measures can be affected if the smallest and largest observations are removed?

a. mean
b. median
c. 50th percentile
d. none of the above

Answer: Central Tendency

You have the data arranged in ascending order. Which one of the following measures can be affected if the smallest and largest observations are removed?

a. mean
b. median
c. 50th percentile
d. none of the above

Mean is affected by every value in the data including largest and smallest observations.

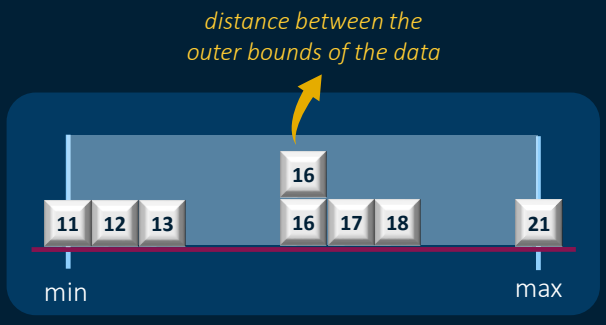
On the other hand, *media*n also called as 50th percentile is less affected by extreme values. Because it is the value that divides the data into two equal halves when arranged in order of magnitude.

Simply locating the center of the data might not be sufficient. Measures of central tendency yield information only about the center, or middle part, of a group of numbers. Looking at the three data examples, all of them have the same mean, which is 15.5, but it is evident that they have different variability. Using measures of variability in conjunction with measures of central tendency makes possible a more complete numerical description of the data.

Measures of Dispersion

- Range

$$Range = x_{max} - x_{min}$$



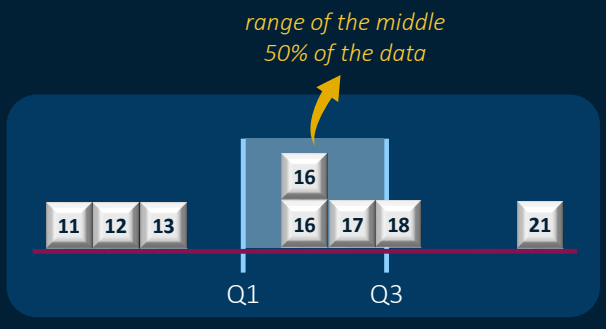
Measures of dispersion enable you to characterize the variability, or spread, of the data.

Range is the simplest measure of dispersion characterized as the difference between the maximum and minimum data values. It is a crude measure of variability, describing the distance between the outer bounds of the data set. A disadvantage of the range is that, because it is computed with the values that are on the extremes of the data, it is affected by extreme values, and its application as a measure of variability is limited.

Measures of Dispersion

- Inter-Quartile Range

$$IQR = Q_3 - Q_1$$



Another measure of variability is the *interquartile range*. The interquartile range is the range of values between the first and third quartile. Essentially, it is the range of the middle 50% of the data and is determined by computing the value of *Q*3 - *Q*1. The interquartile range is especially useful in situations where you are more interested in values toward the middle and less interested in extremes.

Measures of Dispersion

- Mean Absolute Deviation

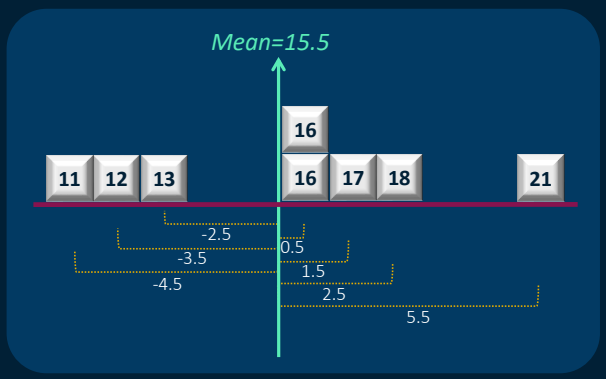
$$MAD = \frac{\sum|x_i - \bar{x}|}{n-1}$$

- Variance

$$s^2 = \frac{\sum(x_i - \bar{x})^2}{n-1}$$

- Standard Deviation

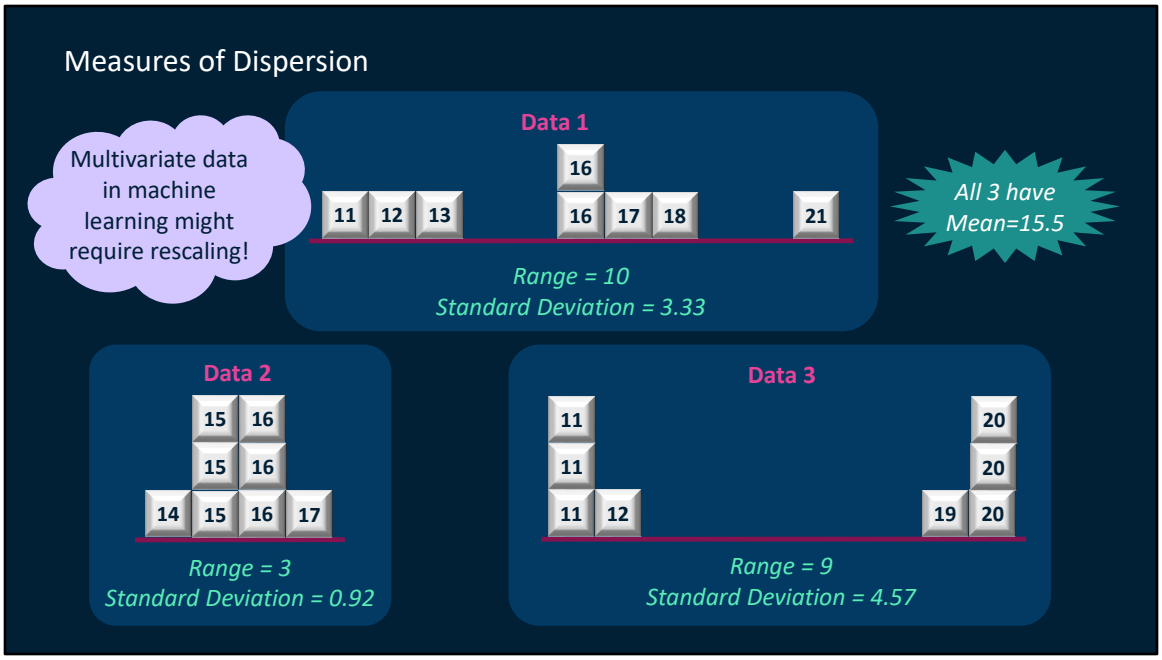
$$s = \sqrt{\frac{\sum(x_i - \bar{x})^2}{n-1}}$$



Three other measures of variability are the variance, the standard deviation, and the mean absolute deviation. They are obtained by subtracting the mean from each value of data and are, therefore, presented together. These measures of dispersion essentially yield the deviation from the mean. Geometrically the negative deviations represent values that are below (to the left of) the mean and positive deviations represent values that are above (to the right of) the mean.

The *mean absolute deviation (MAD)* is the average of the absolute values of the deviations around the mean for a set of numbers. Because absolute values are not conducive to easy manipulation, an alternative mechanism for overcoming the zero-sum property of deviations from the mean is devised. This approach utilizes the square of the deviations from the mean. The result is the variance, an important measure of variability. The *variance* is the average of the squared deviations about the arithmetic mean for a set of numbers. The sample variance is denoted by s-square.

Statistics measured in squared units are problematic to interpret. The standard deviation is a more popular measure of variability. The *standard deviation* is the square root of the variance. The population standard deviation is denoted by *s*. One feature of the standard deviation that distinguishes it from a variance is that the standard deviation is expressed in the same units as the mean and the raw data, whereas the variance is expressed in those units squared.



In the previous example, the three data sets have the same mean but different measures of variability. Their respective range and standard deviation values are shown.

You encounter multivariate data in machine learning applications in which you have multiple features spanning varying degrees of magnitude, range, and units. A significant issue is that the range of the variables can differ a lot. Using the original scale might put more weight on the variables with a large range. Thus, rescaling might be required to have the equal range and/or variance. Feature scaling is discussed in Lesson 5.

Question: Dispersion

Standard deviation and variance essentially yield the deviation from which of the following?

a. maximum value
b. median value
c. mode value
d. mean value

Answer: Dispersion

Standard deviation and variance essentially yield the deviation from which of the following?

a. maximum value
b. median value
c. mode value
d. mean value

Correct answer: d
The variance and the standard deviation essentially yield the deviation from the mean. The *variance* is the average of the squared deviations about the arithmetic mean for a set of numbers. The *standard deviation* is the square root of the variance.

Question: Summary Statistics

Poor Jose has his head in an oven and his feet in a freezer!

You are a statistician examining the temperature dispersion in different parts of Jose's body. Which of the following statements is reasonable?

a. On an average, he must feel just fine.
b. Locating the center of the (temperature) data is sufficient.
c. Characterizing the variability or spread of the (temperature) data is necessary.
d. He is a 'hot head' who has 'cold feet'!

Answer: Summary Statistics

Poor Jose has his head in an oven and his feet in a freezer!

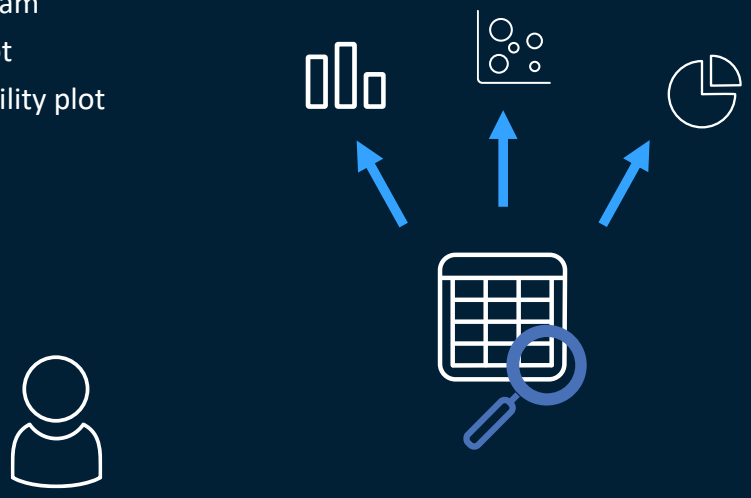
You are a statistician examining the temperature dispersion in different parts of Jose's body. Which of the following statements is reasonable?

a. On an average, he must feel just fine.
b. Locating the center of the (temperature) data is sufficient.
c. Characterizing the variability or spread of the (temperature) data is necessary.
d. He is a 'hot head' who has 'cold feet'!

Correct Answer: c
This is a popular joke about the statistical "average" involving the statistician who put his head in an oven and his feet in a freezer, saying, "On average, I feel fine." Measures of central tendency yield information only about the center, or middle part, of a group of numbers and thus simply locating the center of the data might not be sufficient. Using measures of dispersion in conjunction with measures of central tendency makes possible a more complete numerical description of the data.

Visualizing Distributions

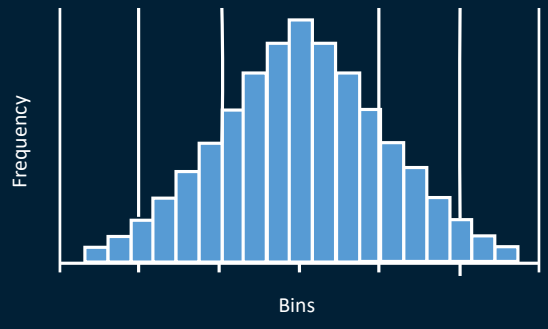
- Histogram
- Box plot
- Probability plot



Another way to understand your data is to display it using graphs and plots. There are several ways to visualize the distribution of your data.

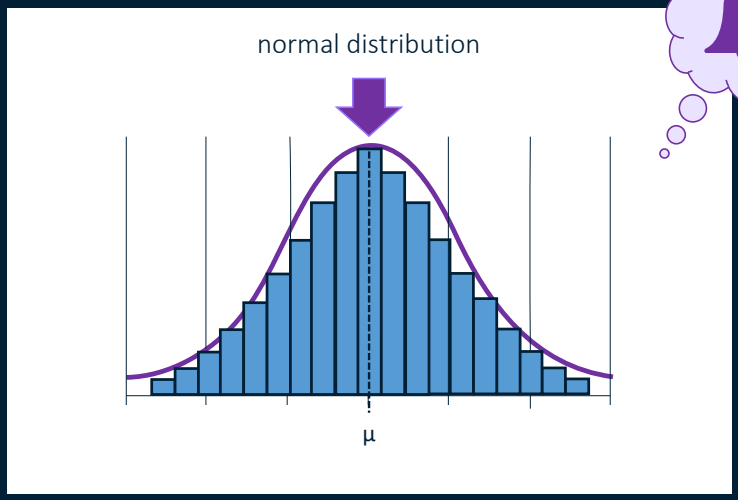
Histograms

visual representation of the frequency distribution of data



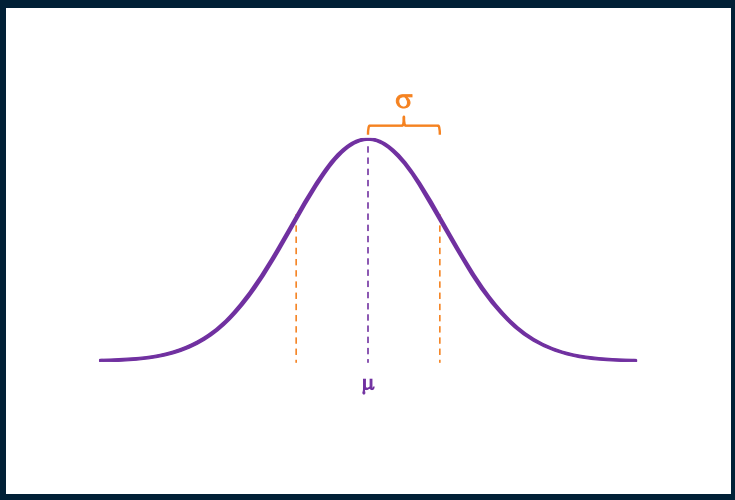
One of the most effective techniques is to plot a histogram. A histogram is a visual representation of the frequency distribution of your data. The frequencies are represented by vertical bars. Each bar in the histogram represents a group of values, sometimes referred to as a bin. The heights of the bars represent the frequency or count of values in the bins.

Histograms

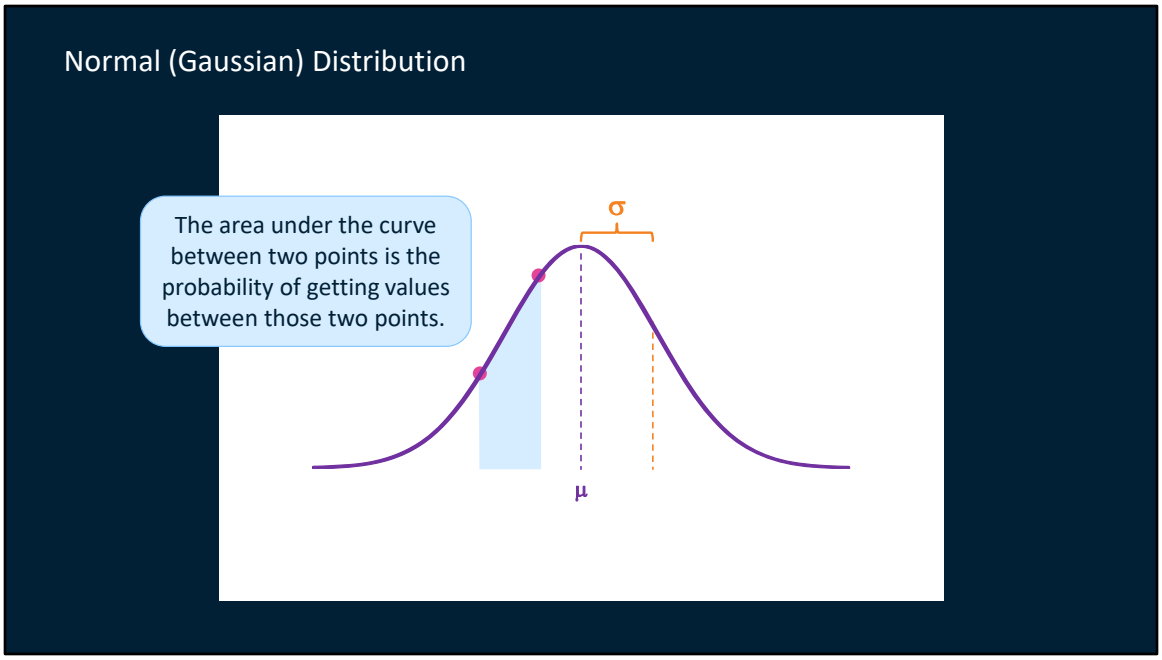


You can easily establish the shape of your data using a histogram. For example, here it shaped like a bell, with values concentrated in the middle. This type of distribution is called a normal distribution. The normal distribution is a common theoretical distribution in statistics.

Normal (Gaussian) Distribution

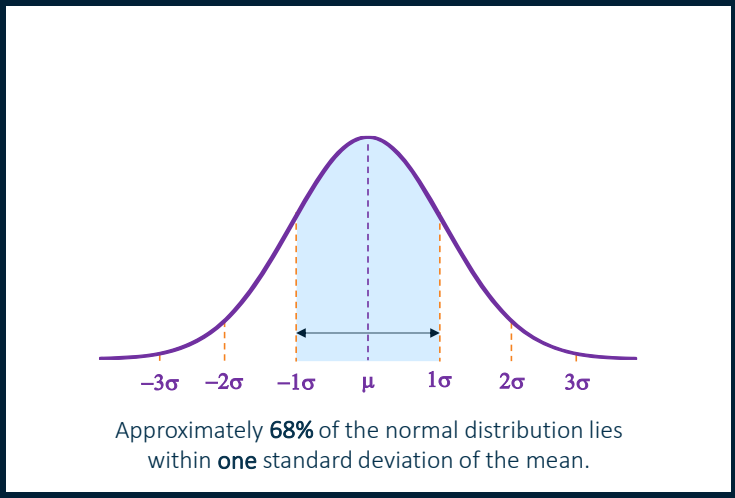


Another name for the normal distribution is the Gaussian distribution. The Gaussian distribution is bell shaped, symmetric, and defined by two parameters, μ (the mean) and σ (the standard deviation). The mean locates the midpoint of the distribution, or the peak of the bell, and the standard deviation describes the variability, or spread of the distribution.



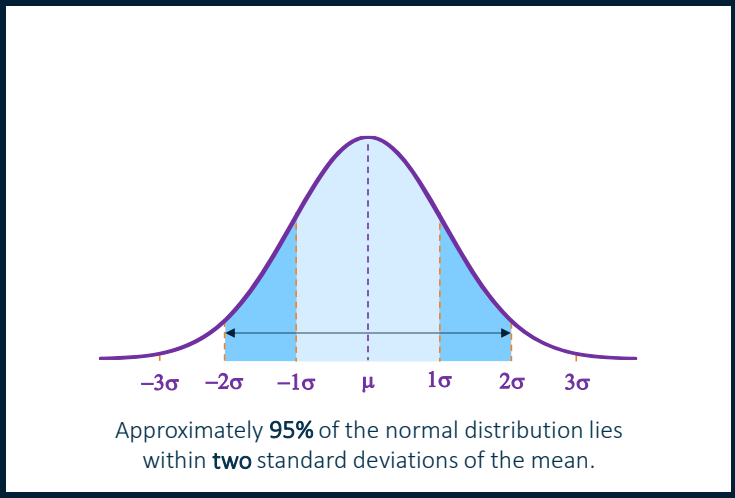
The area under the curve between two points is the probability of getting values between those two points.

Normal (Gaussian) Distribution



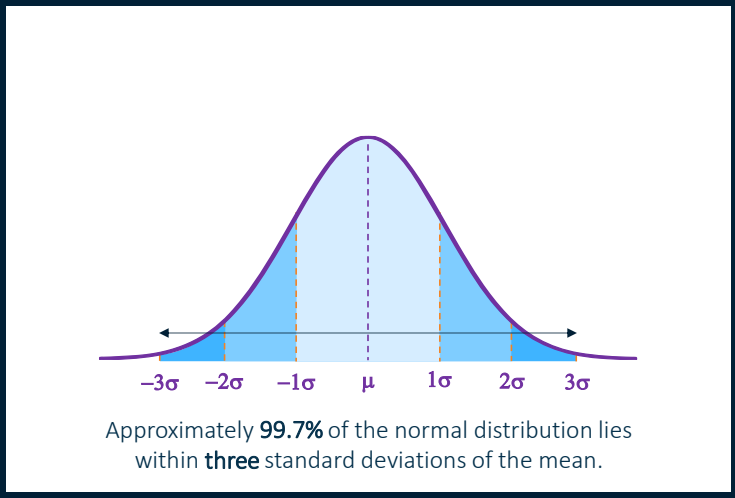
Some well-known probabilities are associated with the mean and standard deviation of the distribution. For example, approximately 68% of the total area under normal distribution lies within ±1 standard deviation of the mean.

Normal (Gaussian) Distribution

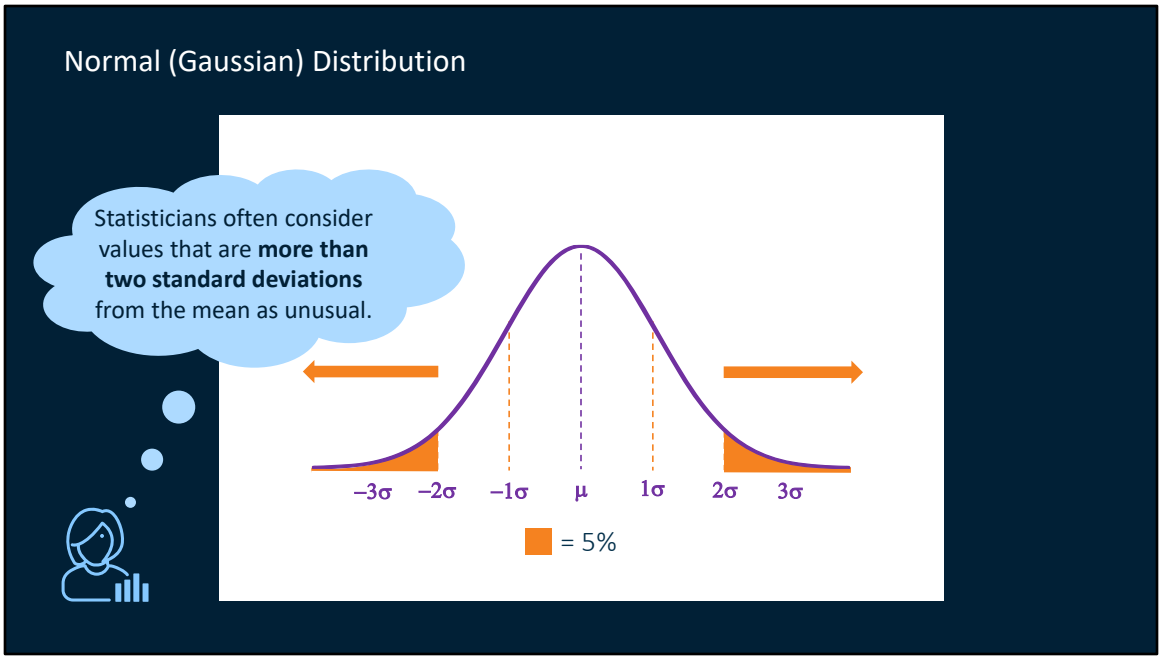


Approximately 95% lies within ±2 standard deviations of the mean.

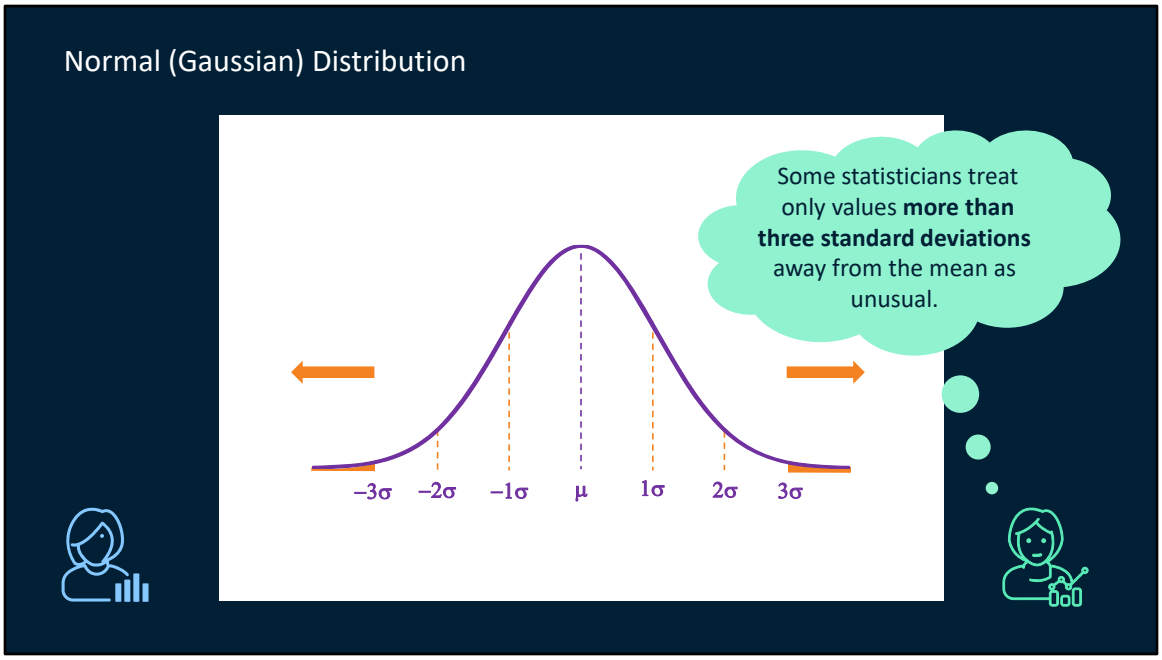
Normal (Gaussian) Distribution



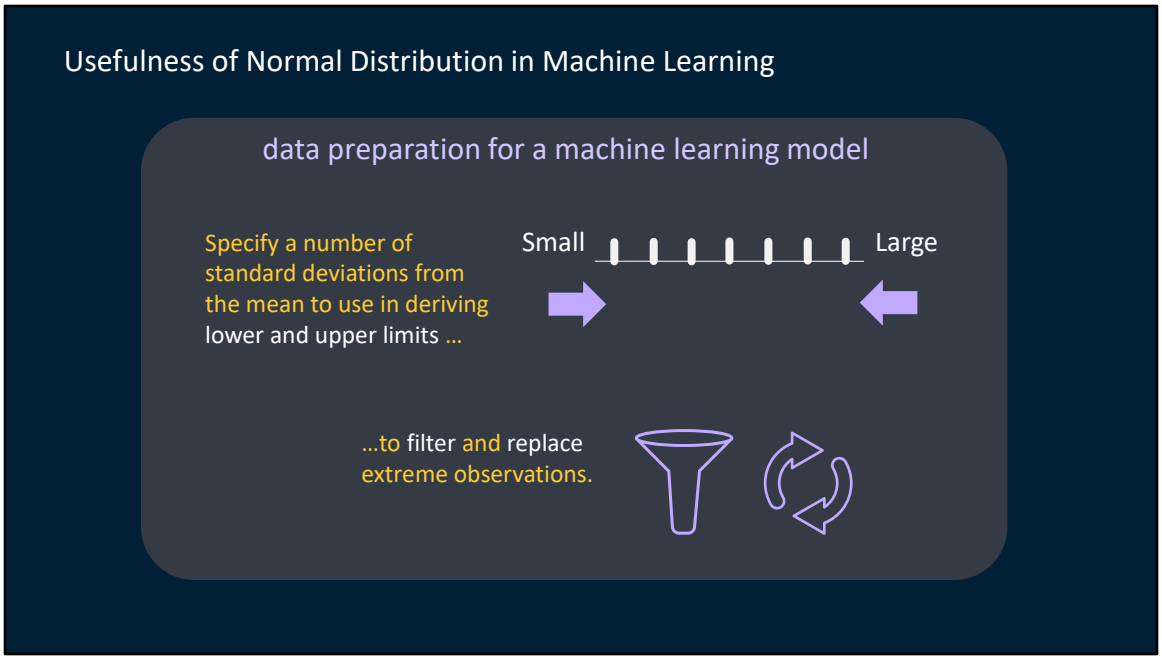
Approximately 99.7% lies within ±3 standard deviations.



Statisticians often consider values that are more than ±2 standard deviations from the mean as unusual. And now you can see why. Only about 5% of all values are that far away from the mean.

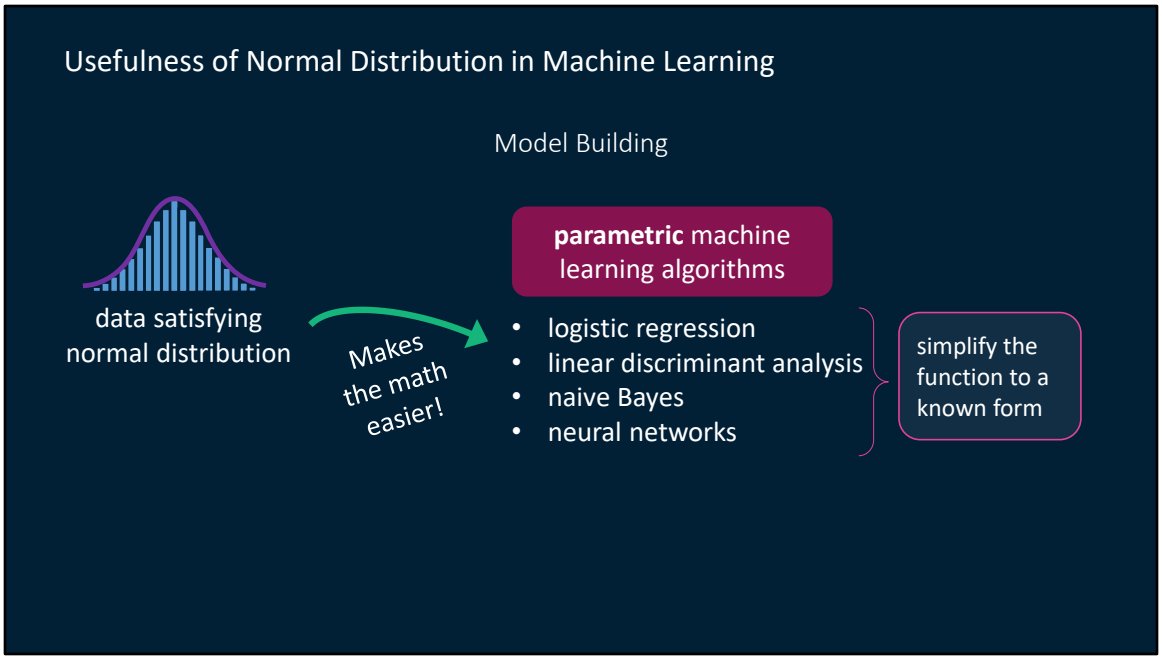


Depending on the context, some statisticians treat only values more than ±3 standard deviations away from the mean as unusual.

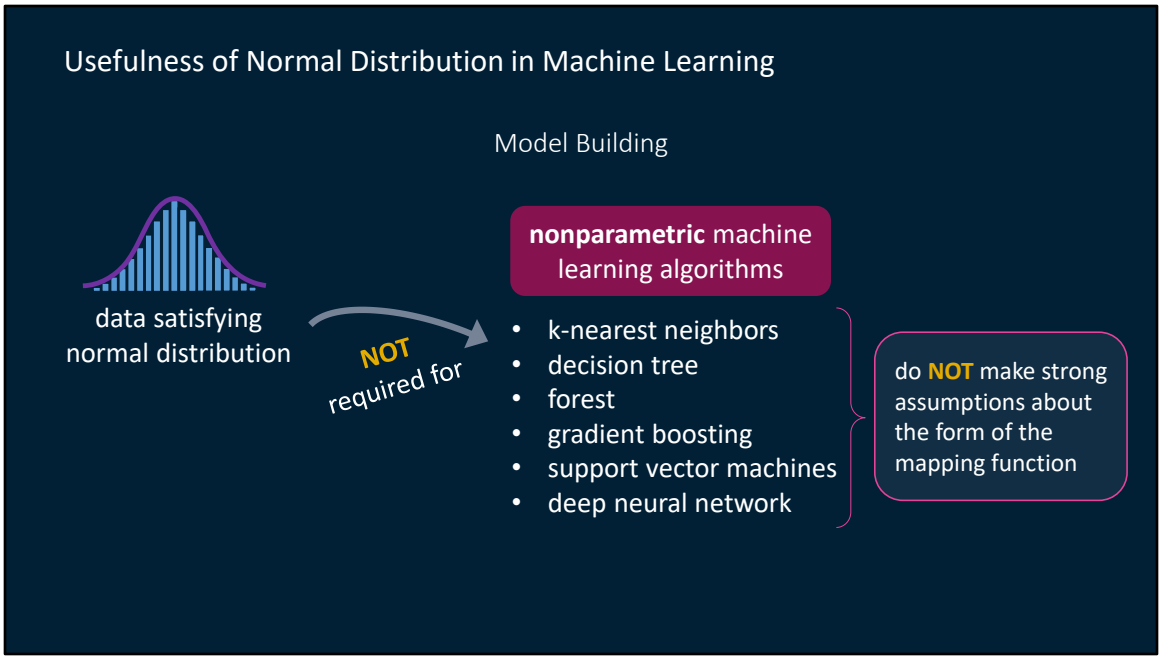


Because the normal distribution has many useful mathematical properties, statistical procedures for data based on a random sample often assume the normal distribution.

Normal distribution assumption is useful in many machine learning tasks including data preparation. You specify the number of standard deviations from the mean to be used in deriving the lower and upper limits for filtering and replacing extreme observations.



Data satisfying normal distribution is beneficial for model building. It makes math easier for models like logistic regression, linear discriminant analysis, naive Bayes, neural networks, and so on. Such algorithms simplify the function to a known form and are called parametric machine learning algorithms. By applying transformation techniques, we can convert the data into a normal distribution.



Data satisfying normal distribution is not required with nonparametric models like k-nearest neighbors, decision trees, forest, gradient boosting, support vector machines, deep neural networks etc. Algorithms that do not make strong assumptions about the form of the mapping function are called nonparametric machine learning algorithms.

Question: Normal (Gaussian) Distribution

Which of the following are typical characteristics of a normal distribution? (Select all that apply)

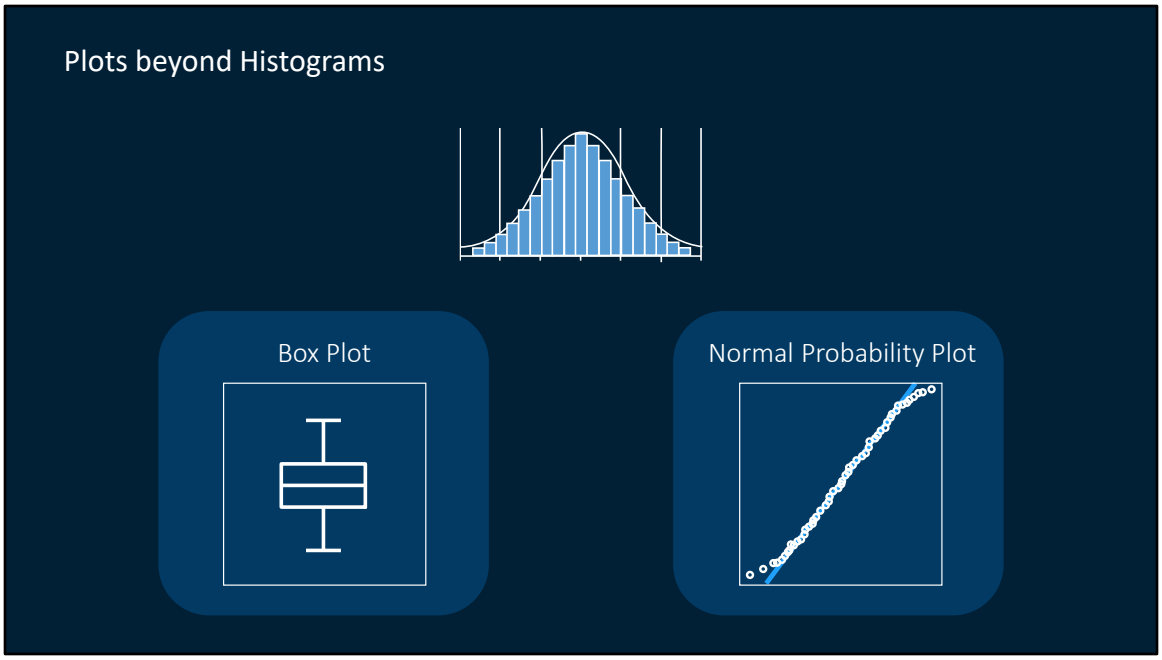
a. multimodal
b. symmetrical, bell shaped
c. mean, median, and mode all coincide
d. approximately 99.7% of values lie within ±3 standard deviation from the mean

Answer: Normal (Gaussian) Distribution

Which of the following are typical characteristics of a normal distribution? (Select all that apply)

a. multimodal
b. symmetrical, bell shaped
c. mean, median, and mode all coincide
d. approximately 99.7% of values lie within ±3 standard deviation from the mean

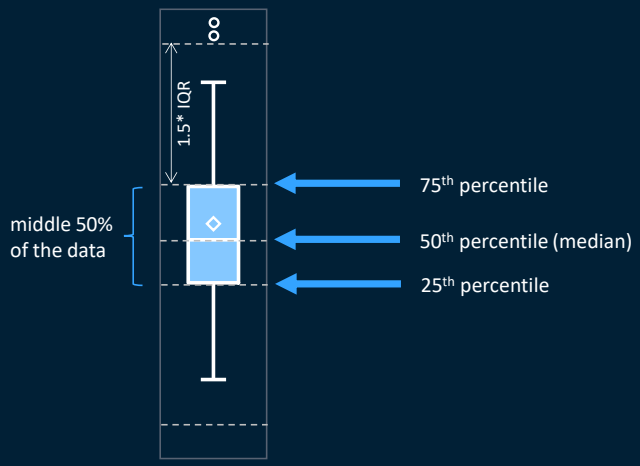
Correct answer: b, c and d.
A normal distribution is unimodal, symmetrical, and bell shaped in which mean, median, and mode are equal. Its empirical rule suggests that approximately 99.7% of values lie within ±3 standard deviation from the mean.



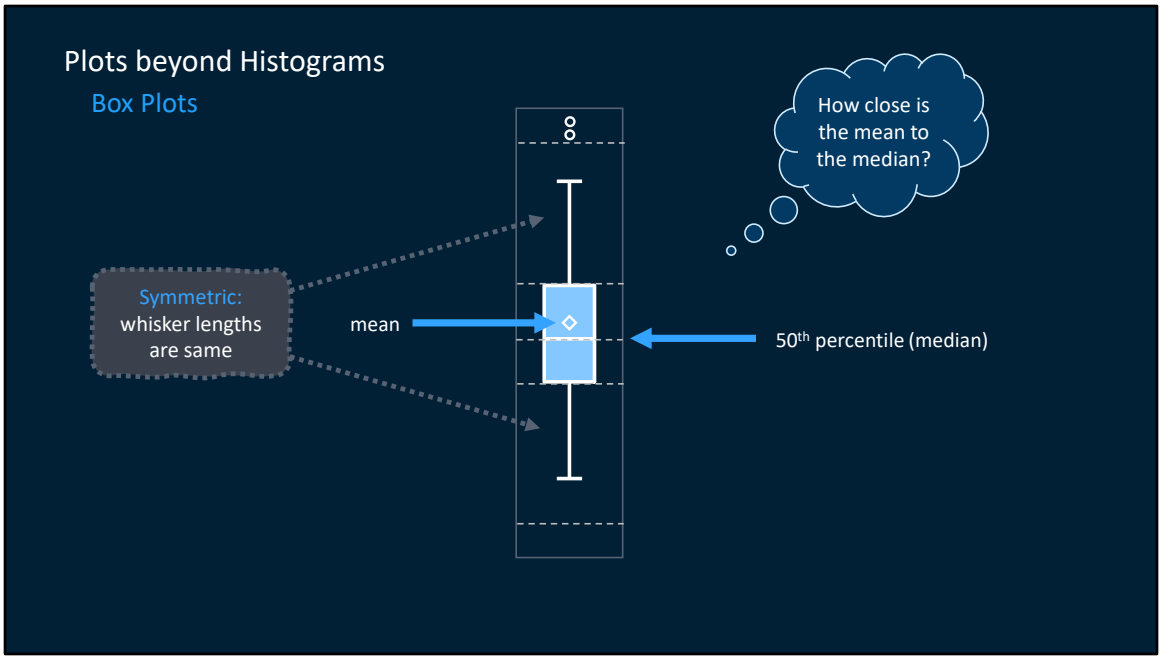
In addition to plotting your data using a histogram, there are other ways to visualize and assess the distribution of your data. You can use box plots and normal probability plots to compare your data with the normal distribution.

Plots beyond Histograms

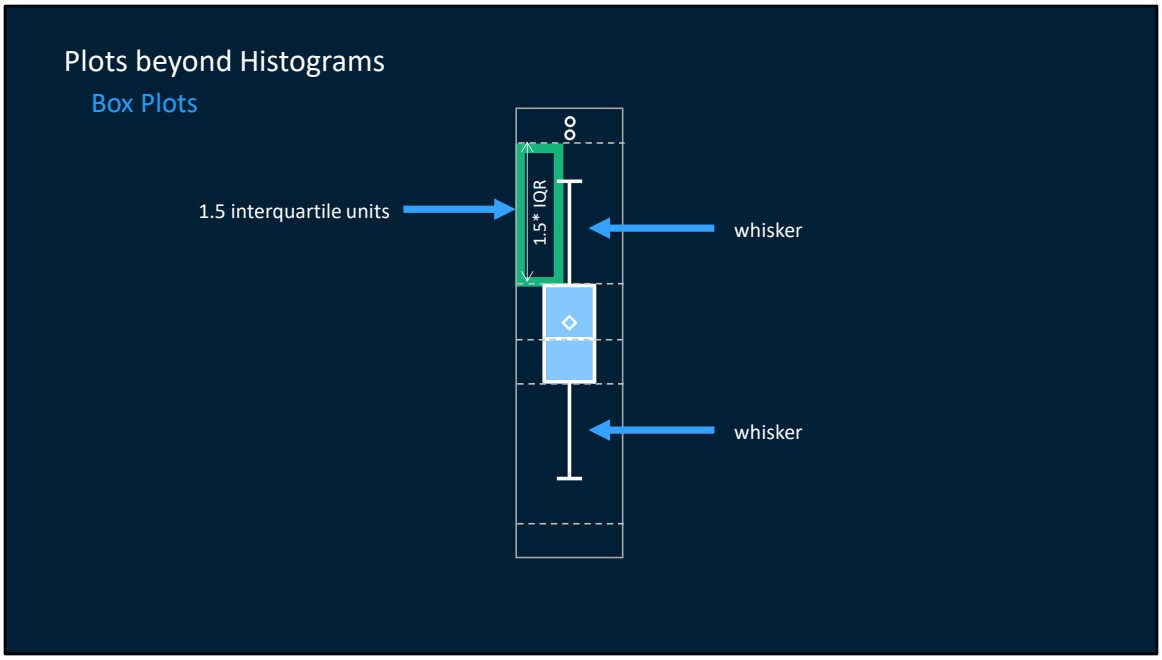
Box Plots



Box plot makes it easy to see how spread out your data is and if there are any outliers. The box represents the middle 50% of your data, which is the interquartile range. The bottom of the box indicates the 25th percentile. The middle line represents the 50th percentile, or the median, and the top line indicates the 75th percentile.



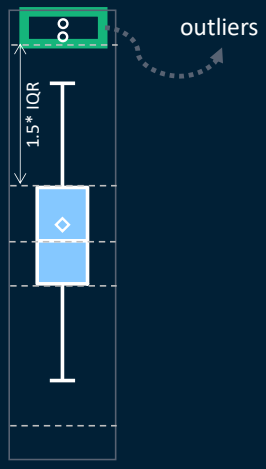
The diamond denotes the mean. Thus, you are able get a rough impression of the symmetry of your distribution by comparing the mean and median, as well as by assessing the symmetry of the box and whiskers around the median line. The median line is close to the center of the box and the whisker lengths are the same. This box plot shows a data distribution that is approximately symmetric.



The whiskers extend from the box as far as the data extends, to a maximum length of 1.5 times the interquartile range (IQR) above the 75th percentile and similarly below the 25th percentile on the other side. This is referred to as 1.5 interquartile units.

Plots beyond Histograms

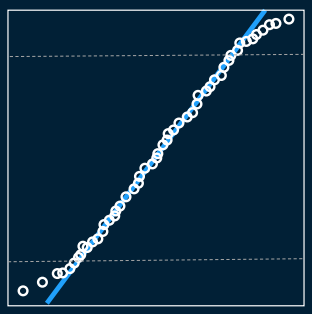
Box Plots



Any data points farther than 1.5 interquartile units away from the box are considered possible outliers and are represented in this plot as circles.

Plots beyond Histograms

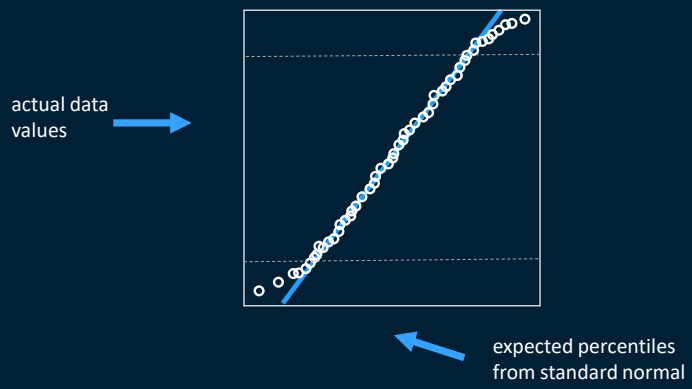
Normal Probability Plots



A normal probability plot is a visual method for determining whether your data comes from a distribution that is approximately normal.

Plots beyond Histograms

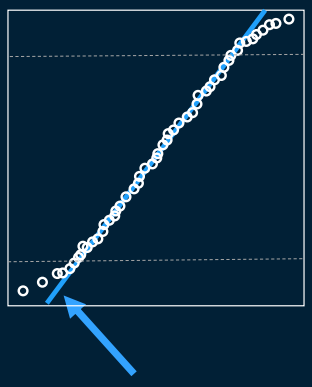
Normal Probability Plots



The vertical axis represents the actual data values, and the horizontal axis displays the expected percentiles from a standard normal distribution.

Plots beyond Histograms

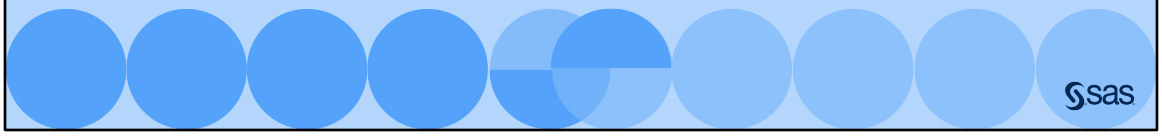
Normal Probability Plots



The normal reference line along the diagonal is where your data would fall if it were perfectly normal. The points hovering around the reference line are the actual data. Here the distribution of data points follows the normal reference line quite closely. Thus, this data distribution looks fairly normal.

Why “1.5” in IQR Method of Outlier Detection?

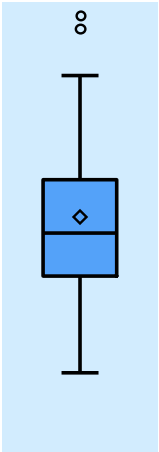
To explore this learning resource in greater detail, please visit the course page on learn.sas.com. There, you’ll find comprehensive course materials, tutorials, and updates to support your learning journey.



Question: Visualizing Distributions

In SAS, how is the median indicated in a box plot?

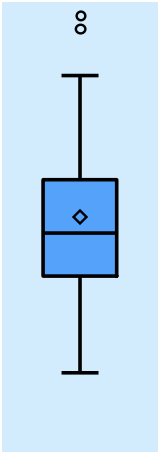
a. a box
b. a whisker
c. a line in the box
d. a point outside 1.5 IQR



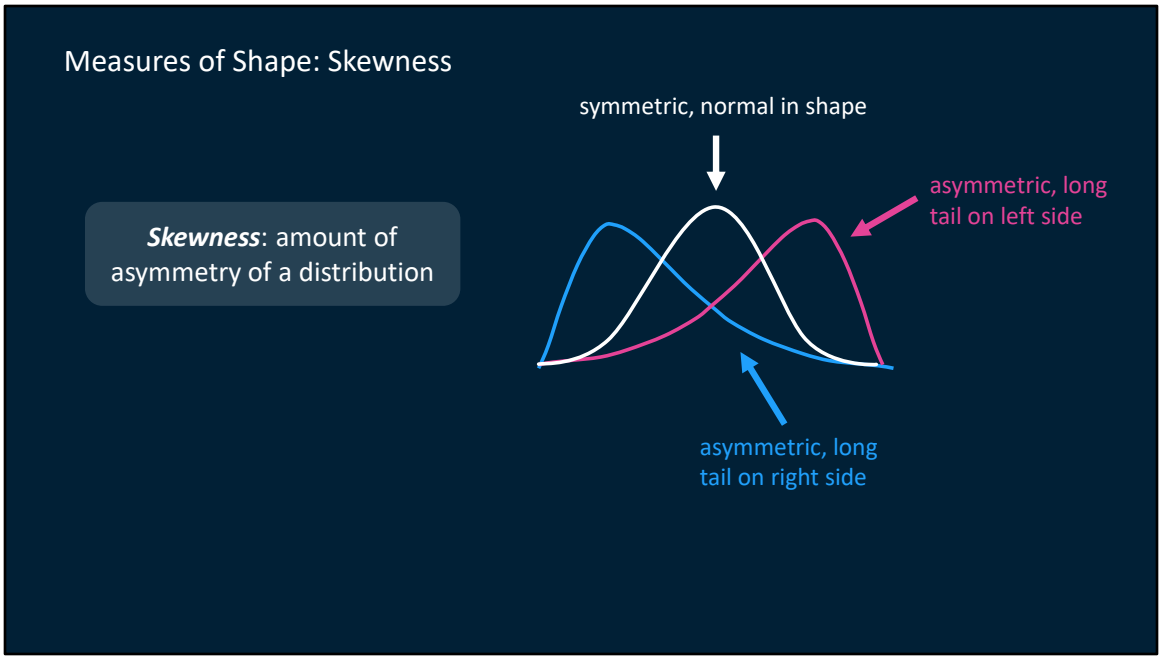
Answer: Visualizing Distributions

In SAS, how is the median indicated in a box plot?

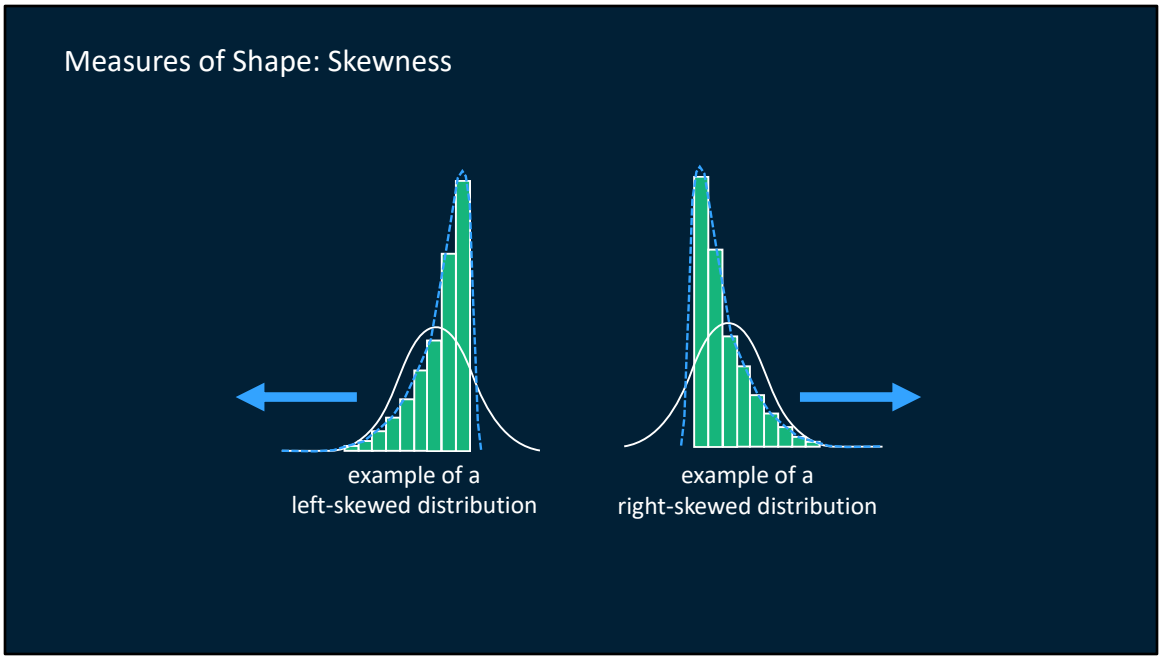
a. a box
b. a whisker
c. a line in the box
d. a point outside 1.5 IQR



Correct answer: c
The middle line in the box represents the median of the data, whereas the box represents the middle 50% of your data, which is the interquartile range.



Simply ascertaining the dispersion of the data might not be sufficient. Measures of dispersion yield information only about the variability, or spread, of a group of numbers. Looking at the three data distributions, all of them have the equal variability, but it is evident that they have different shapes. Not every distribution is a normal distribution. It is important to determine whether the data is symmetric or asymmetric. When data is asymmetric, the distribution tail on one side is longer than on the other side. Using measures of shapes like skewness help understand the tendency of your data to be more spread out on one side of the mean than on the other. Skewness is a measure of the asymmetry of the distribution. Skewness is when a distribution is asymmetrical or lacks symmetry. A distribution of data in which the right half is a mirror image of the left half is said to be symmetrical. One example of a symmetrical distribution is the normal distribution.

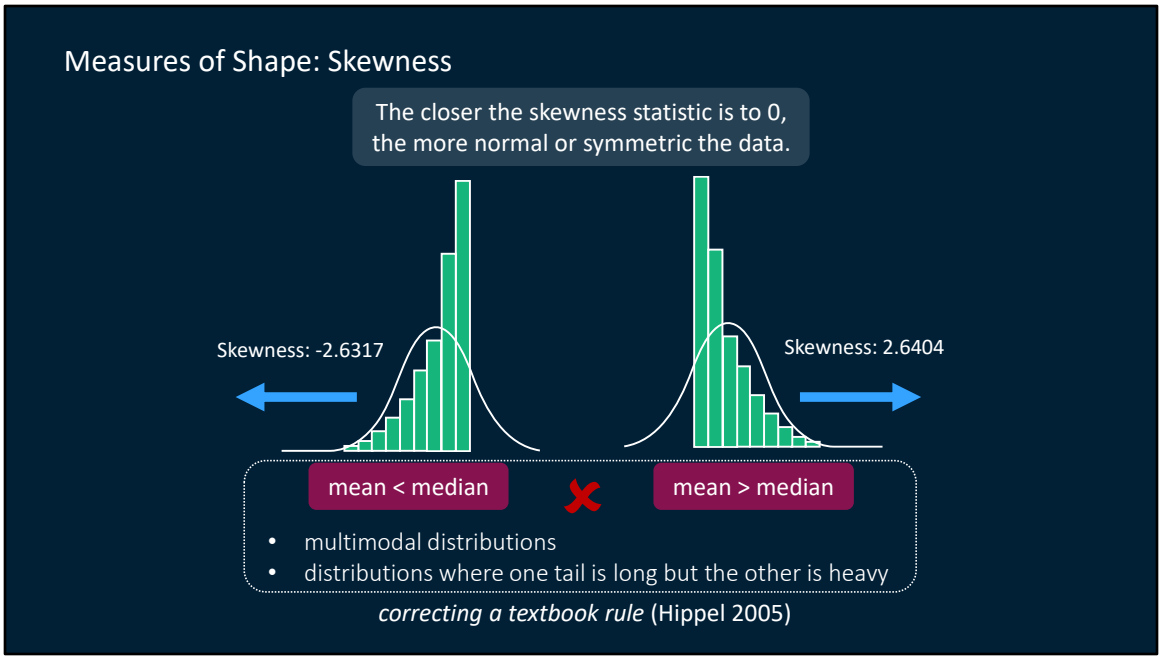


Here are examples of two skewed distributions.

The histogram on the left is an example of a left-skewed distribution, shown by a blue dotted curve. Most values are clustered around the right tail of the distribution, and the left tail of the distribution is longer.

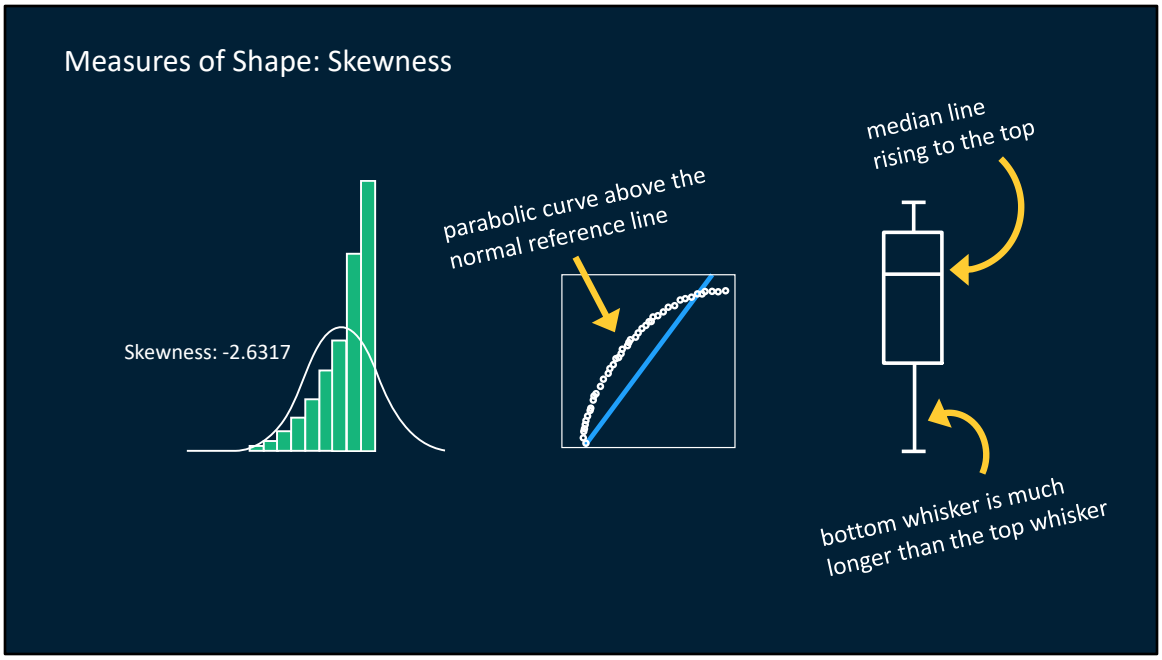
The histogram on the right is an example of a right-skewed distribution, again shown by a blue dotted curve. Most values are clustered around the left tail of the distribution, while the right tail of the distribution is longer.

The white curve on each of these graphs represents the shape of the normal distribution with the mean and variance estimated from the sample data. You can think of the direction of skewness as the direction the data is trailing off to.

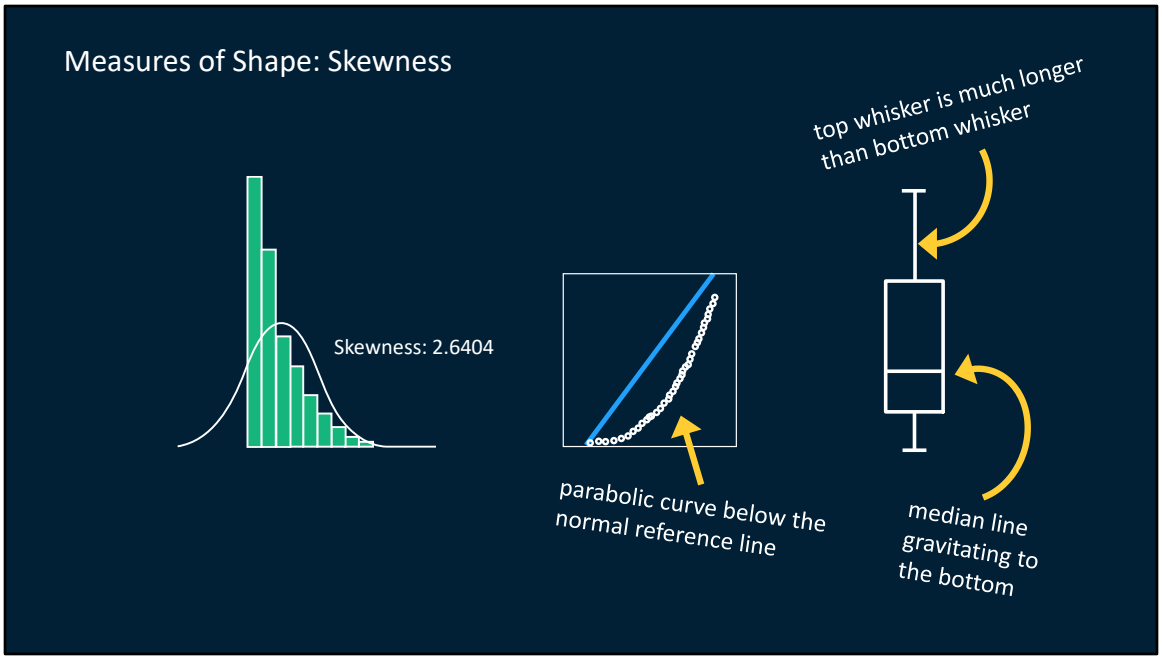


The closer the skewness statistic is to 0, the more normal or symmetric the data. When the statistic is negative, the data is left-skewed. When the statistic is positive, the data is right-skewed.

A left-skewed distribution tells us that the mean is less than the median. A right-skewed distribution tells us that the mean is greater than the median. This rule-of-thumb holds for many familiar continuous distributions, such as the exponential, lognormal, and chi-square distributions. However, this relationship does not hold in general with multimodal distributions or in distributions where one tail is long but the other is heavy.



You can also gauge the shape of the distribution by normal probability plots and box plots. The typical normal probability plot of a left-skewed distribution shows a parabolic curve above the normal reference line. In the box plot of a left-skewed distribution, the bottom whisker is much longer than the top whisker and the median line is rising to the top of the box.



The typical normal probability plot of a right-skewed distribution shows a parabolic curve below the normal reference line. In the box plot of a right-skewed distribution, the top whisker is much longer than the bottom whisker and the median line is gravitating toward the bottom of the box.

Question: Skewness

Which of the following are true in case of a positively skewed distribution? (Select all that apply)

a. it's an asymmetric distribution
b. mean = median = mode
c. most values are clustered around the left tail of the distribution while the right tail of the distribution is longer
d. it is also called left-skewed distribution

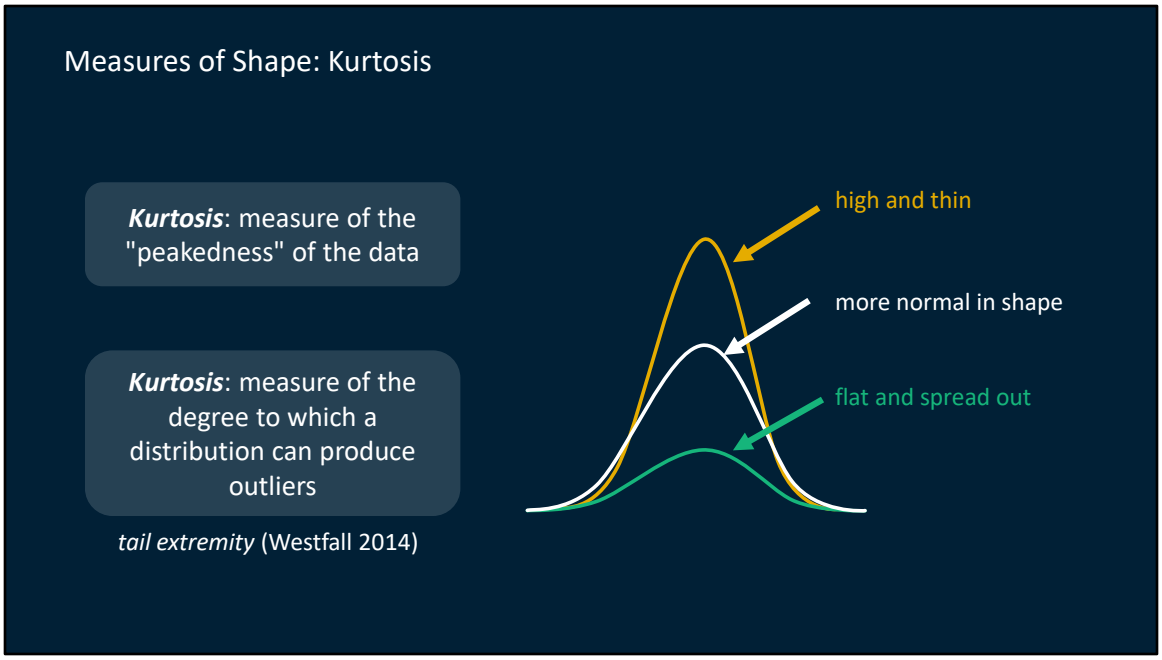
Answer: Skewness

Which of the following are true in case of a positively skewed distribution? (Select all that apply)

a. it's an asymmetric distribution
b. mean = median = mode
c. most values are clustered around the left tail of the distribution while the right tail of the distribution is longer
d. it is also called left-skewed distribution

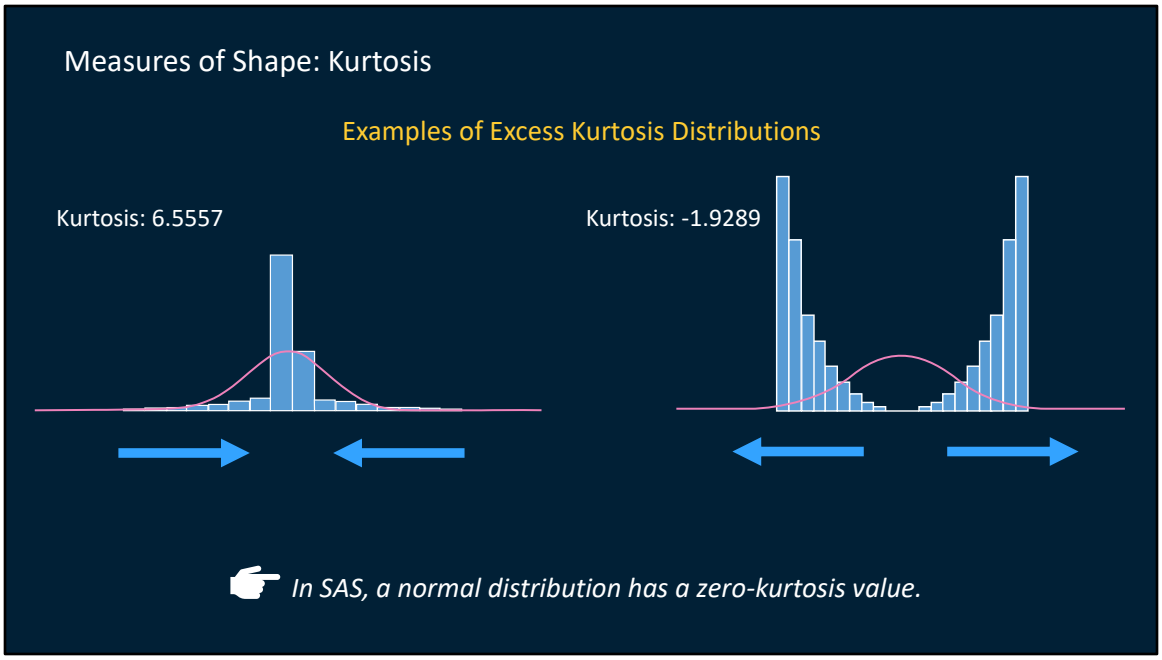
Correct answer: a and c.

A positively skewed (or right-skewed) distribution is a type of distribution in which most values are clustered around the left tail of the distribution and the right tail of the distribution is longer and the mean is greater than the median thus asymmetric.



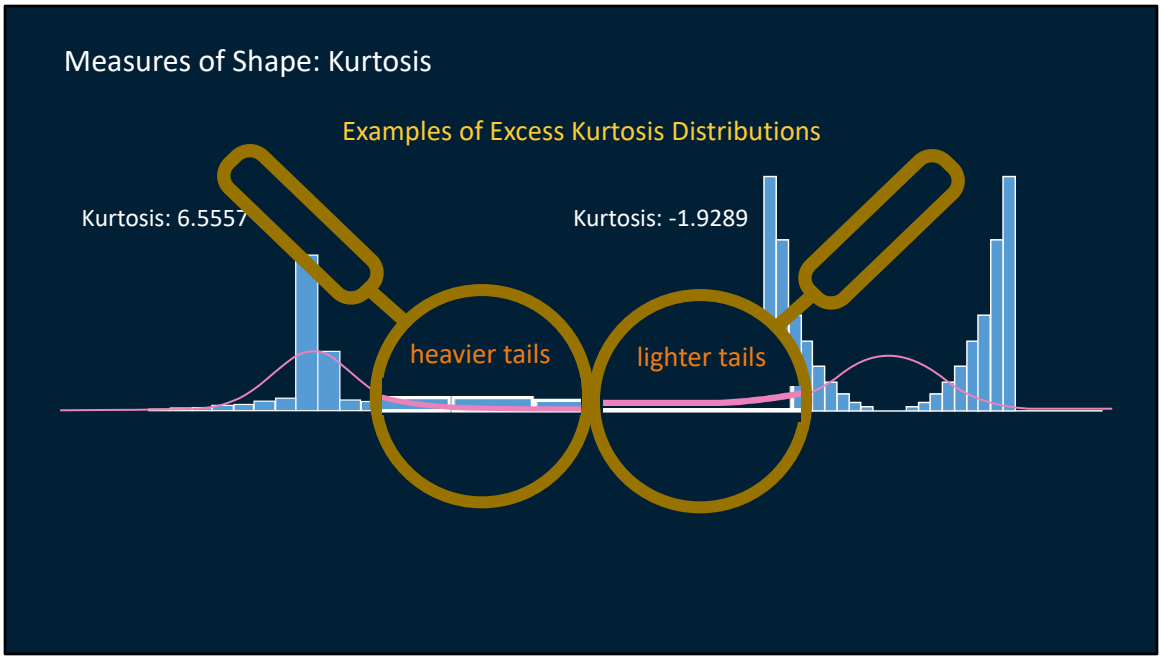
Again, merely ensuring that the data distribution is symmetrical might not be sufficient. Looking at the three data distributions, they all have the same central tendencies as well as equal variability. Also, all three are symmetrical distributions, but it is evident that they have different *peakedness*. Kurtosis is sometimes said to measure the peakedness of a distribution as opposed to its flatness. A distribution of data that is neither high-and-thin nor flat-and-spread-out is said to be more normal in shape.

However, Westfall (2014) and other authors have pointed out that this interpretation is incorrect. Westfall makes the following distinction:
Kurtosis tells you virtually nothing about the shape of the peak—its only unambiguous interpretation is in terms of tail extremity, that is, either existing outliers (for the sample kurtosis) or propensity to produce outliers (for the kurtosis of a probability distribution).

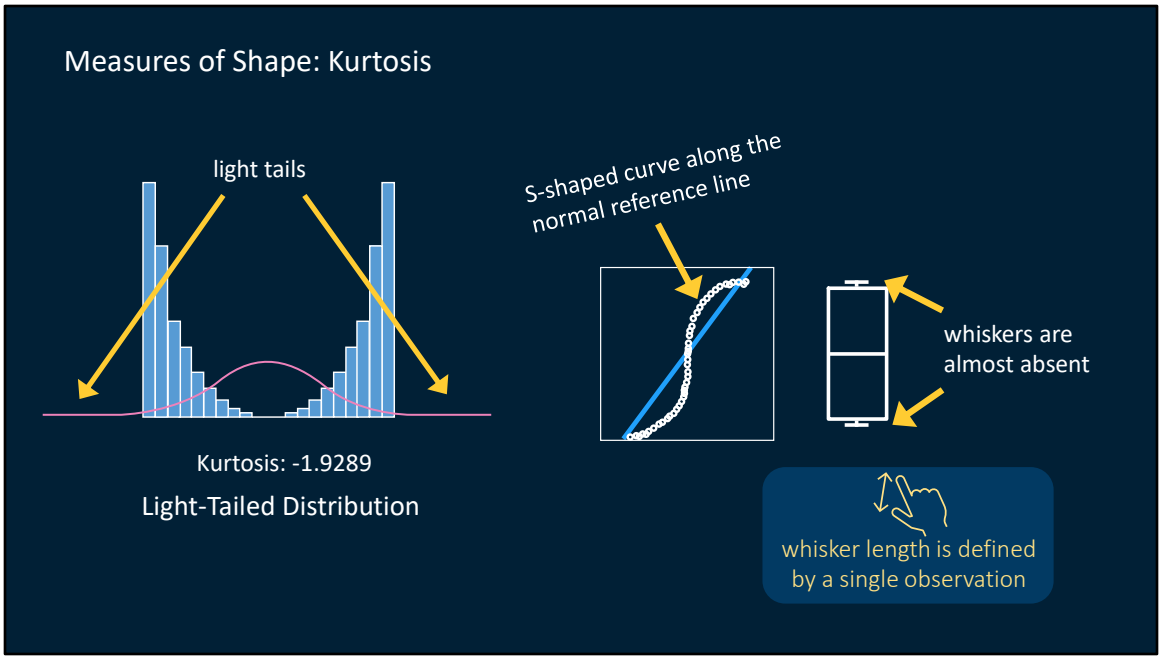


Here are examples of two excess kurtosis distributions. The graph on the left is an example of a positive-excess kurtosis distribution. The graph on the right is an example of a negative-excess kurtosis distribution. ***Kurtosis*** measures the tendency of your data to be concentrated toward the center or toward the tails of the distribution. The closer the kurtosis statistic is to 0, the closer the tails of your data resemble the tail thickness of the normal distribution. The pink curve on each of these graphs represents the shape of the normal distribution with the mean and variance estimated from the sample data.

Be aware that normal distribution actually has a kurtosis value of positive 3. SAS has standardized it to zero, so that when you are assessing normality, the values against which to compare the sample skewness and kurtosis are the same.



A positive kurtosis statistic means that the data has heavier tails and is more concentrated about the mean than a normal distribution. A negative kurtosis statistic means that the data has lighter tails and is less concentrated about the mean than a normal distribution. If you were to zoom in on the positive kurtosis histogram shown on the left, you'd see that the data extends well beyond two standard deviations of the mean, having more mass in the tails, which are "fatter" or "heavier." Similarly, on the negative kurtosis histogram shown on the right, you'd see less mass in the tails, which are "thinner" or "lighter."

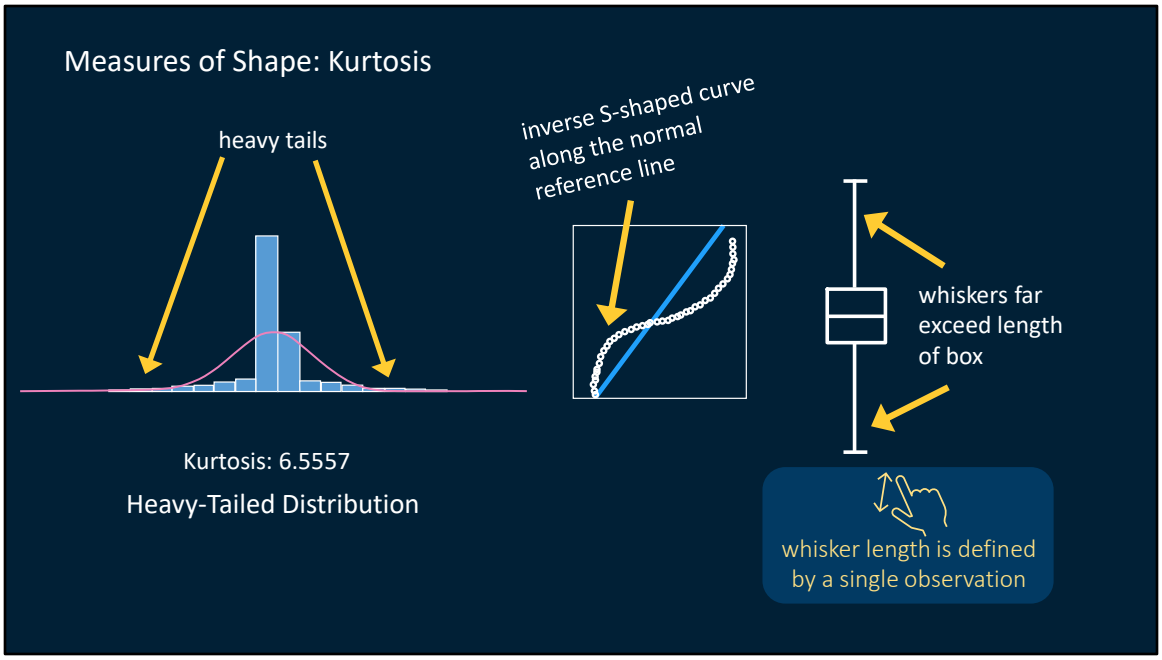


Let's see the shape of the distribution using normal probability plots and box plots.

The histogram shows a light-tailed distribution.

The typical normal probability plot of such a negative-excess kurtosis distribution shows an S-shaped curve along the normal reference line. Remember that the vertical axis is your ordered response values, so the mean of your data is a horizontal line running approximately through the middle of the plot. If you examine your data points paying particular attention to the tails, the data in the tails looks like it is being squashed from the top and the bottom, toward the mean. In other words, the data points are closer to the mean than they should be if they were normal, meaning that the tails of the distribution are light.

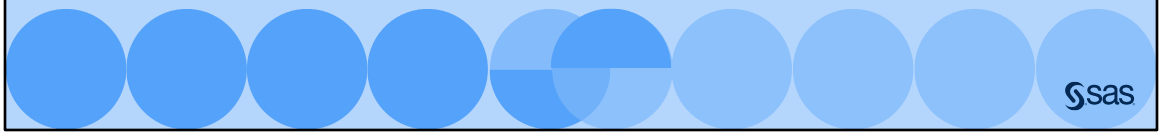
The box plot of a negative-excess kurtosis distribution has extremely short tails (actually a U-shaped population, with a dip in the middle rather than a hump). The whiskers are almost absent. However, it is advised not to put too much emphasis on whisker length because the length is defined by a single observation.



The histogram shows a heavy-tailed distribution. The typical normal probability plot of such a positive-excess kurtosis distribution shows an inverse S-shaped curve along the normal reference line. If you examine your data points paying particular attention to the tails, the data in the tails looks like it is being stretched in the vertical direction, away from the mean. In other words, the data points are farther from the mean than they should be if they were normal, meaning the tails of the distribution are heavy. The box plot of a positive-excess kurtosis distribution has long tails. The length of the whiskers far exceeds the length of the box.

Platykurtic and Leptokurtic Distributions

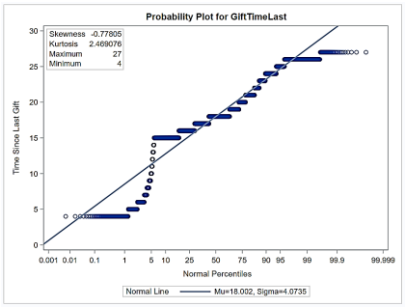
To explore this learning resource in greater detail, please visit the course page on learn.sas.com. There, you'll find comprehensive course materials, tutorials, and updates to support your learning journey.

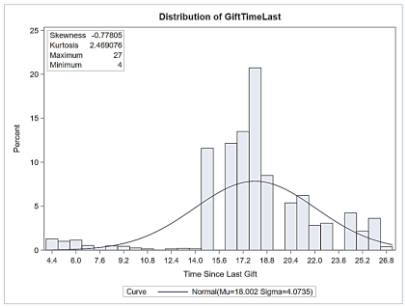


Question: Normal Distribution

Examine the normal probability plot and histogram of Time Since Last Gift. Do the plots indicate that the data is approximately normally distributed?

a. True
b. False



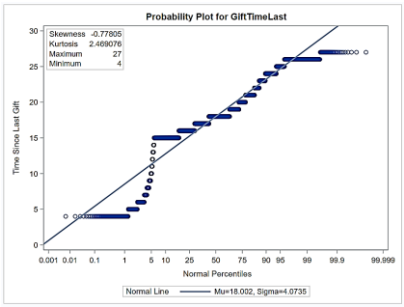


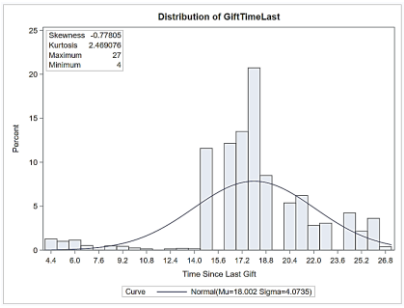
Answer: Normal Distribution

Examine the normal probability plot and histogram of Time Since Last Gift. Do the plots indicate that the data is approximately normally distributed?

a. True

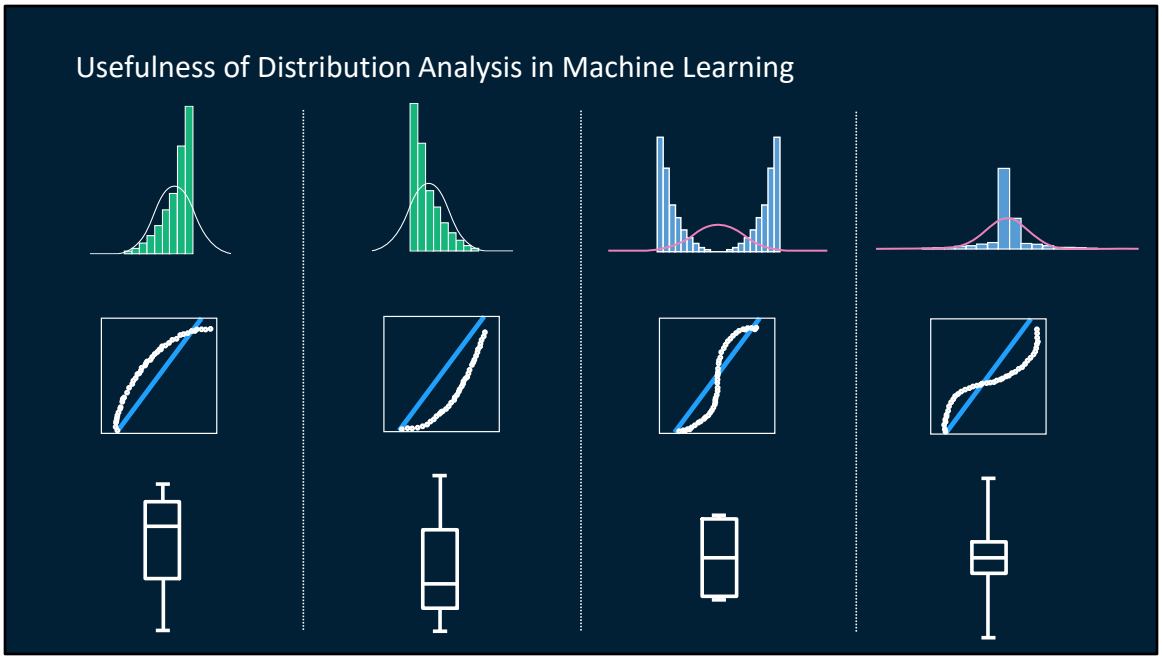
b. False





Correct answer: No

The points do not follow the diagonal reference line in the normal probability plot, indicating that there is departure from normality. The histogram does not follow the normal curve. The data has high kurtosis and is a little negatively skewed, so the data doesn't seem to be approximately normal.

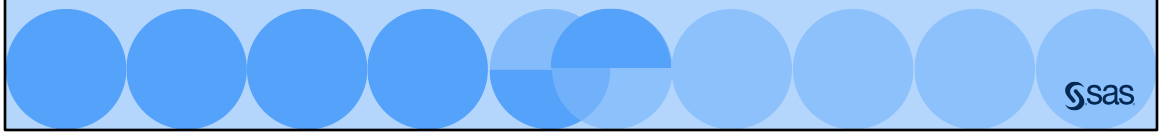


Real life data is usually dirty and noisy. Exploring the data can be one of the most important and time-consuming parts in a typical machine learning project. Common distribution plots and summary statistics that you have seen in this section can be used for exploring machine learning data.

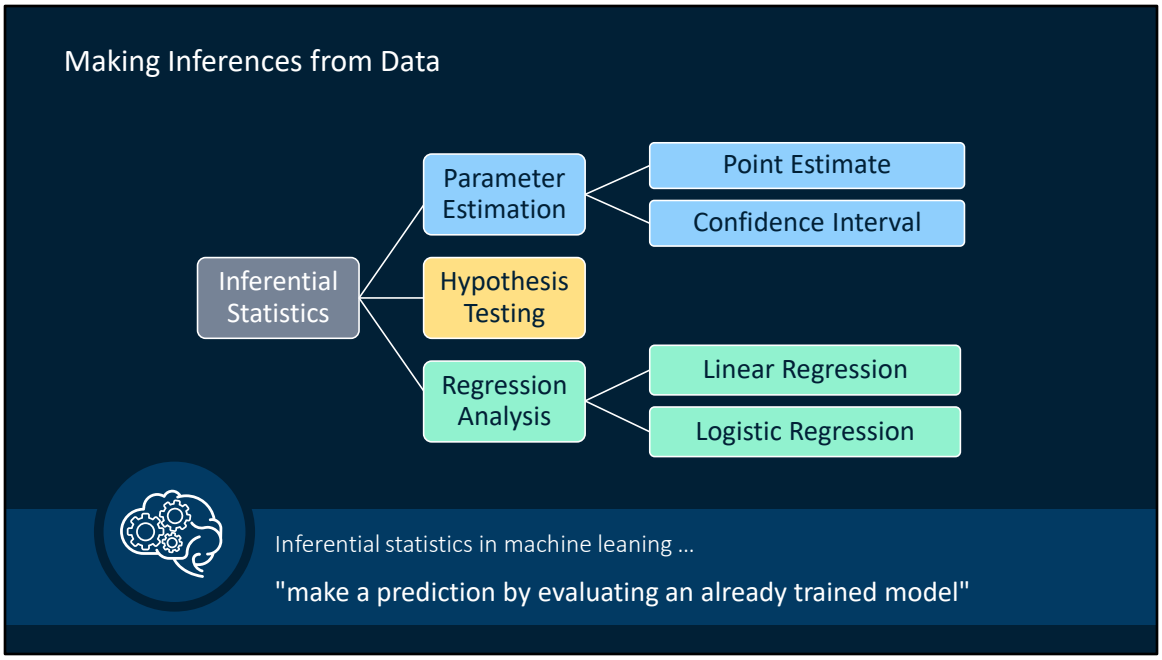
The main point here is that the non-normal data distributions in histograms, box plots, and normal probability plots are visually very obvious. Using measures of shape like skewness and kurtosis in conjunction with measures of central tendency and dispersion makes possible a near complete numerical description of the data.

Demo: Exploring Data

To explore this learning resource in greater detail, please visit the course page on learn.sas.com. There, you'll find comprehensive course materials, tutorials, and updates to support your learning journey.



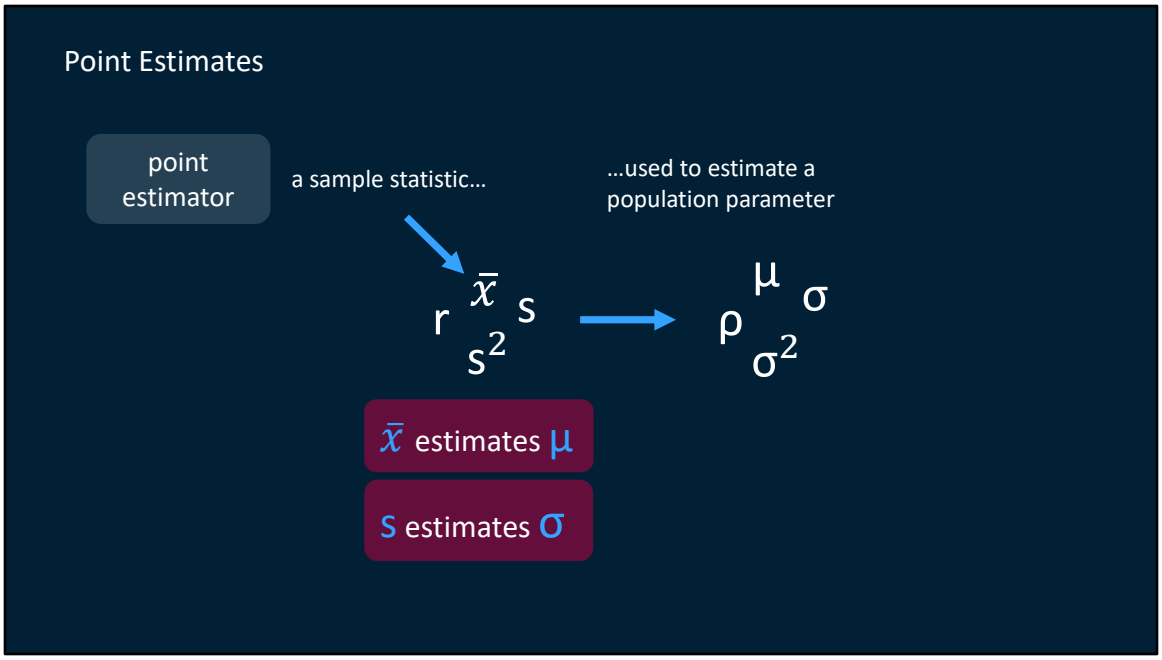
2.3 Inferential Statistics



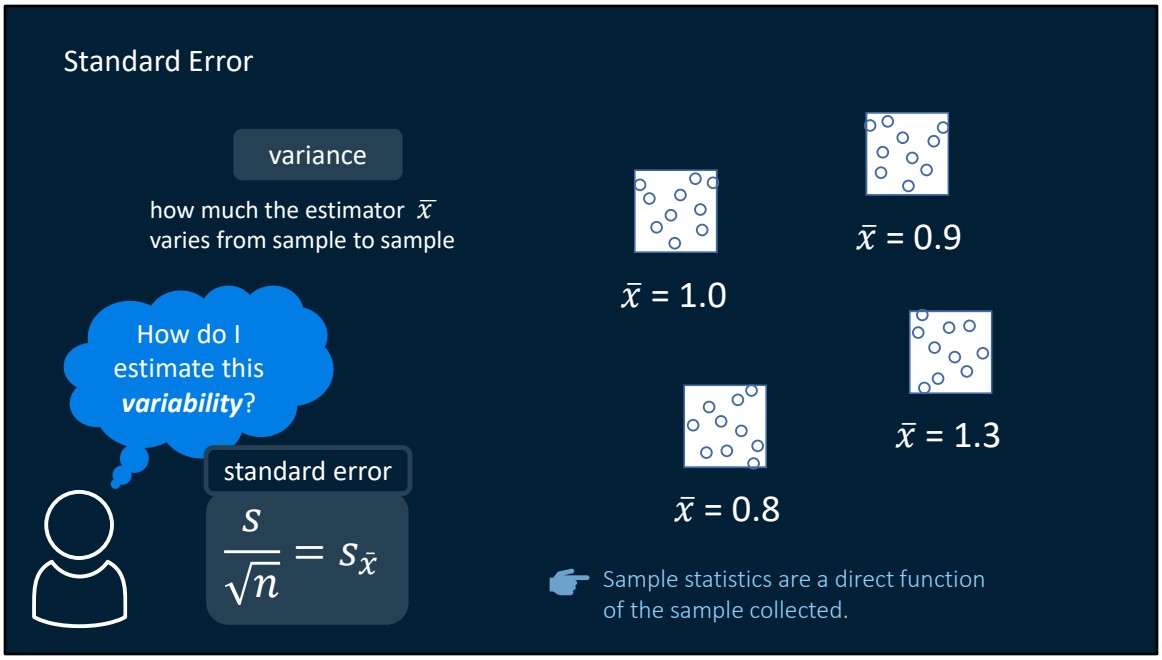
Inferential statistics help to suggest explanations for a situation or phenomenon. It enables you to draw conclusions based on extrapolations, and in that way it's fundamentally different from descriptive statistics, which merely summarize the data that has actually been measured.

Inferential statistics can be broadly classified into parameter estimation, hypothesis testing, and regression analysis. There are two types of parameter estimates: point estimates and interval estimates, also known as confidence intervals. You use hypothesis testing to test an assumption about a population parameter. Regression is a broad term. However, this course focuses on linear regression and logistic regressions. We will discuss them in upcoming lessons.

In machine learning, the term "inference" is sometimes used instead to mean "make a prediction by evaluating an already trained model". In this context, inferring properties of the model is referred to as "training" or "learning" (rather than "inference"), and using a model for prediction is referred to as "inference" (rather than "prediction").

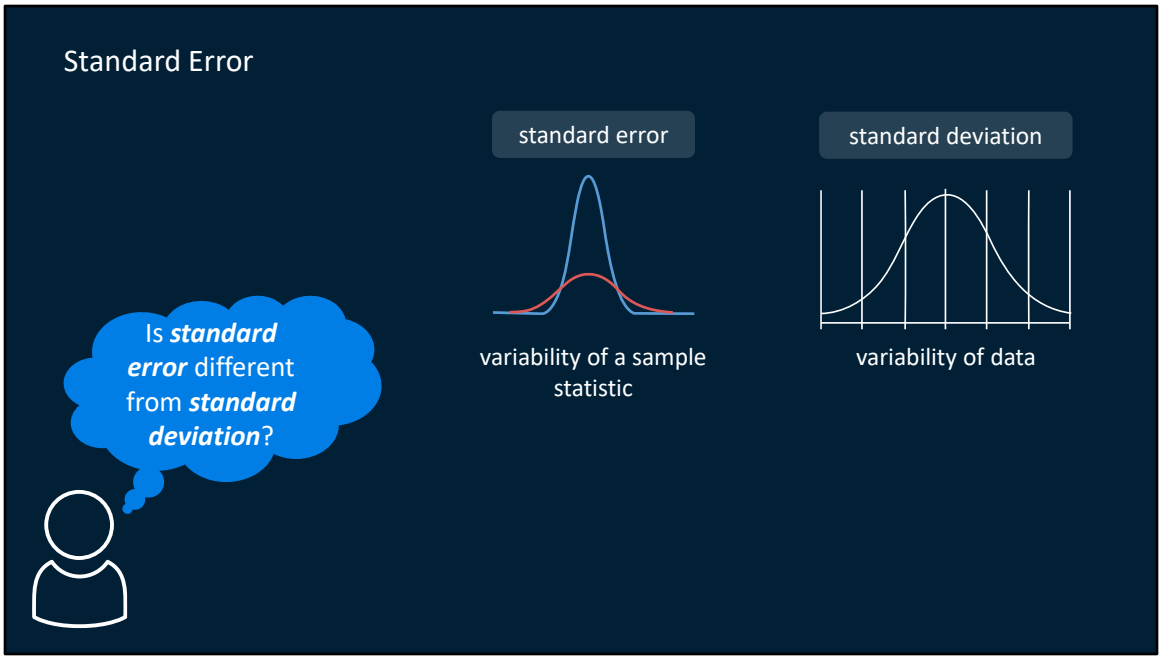


A point estimator is a sample statistic used to estimate a population parameter. You're already familiar with several point estimators. For example, you know that you can use X-bar, the sample mean, to estimate μ, the population mean. Similarly, you can use s, the sample standard deviation, to estimate σ, the population standard deviation.



Sample statistics are a direct function of the sample collected. This means that if a different sample was collected, the sample statistics would change. However, we collect only one sample due to time and resources. An estimator takes on different values from sample to sample, so it's important to know the variance of an estimator. For example, different samples yield different estimates of the mean for the same population. The variance of the sample mean refers to how much the value of the sample mean varies from sample to sample.

The key is to understand and estimate this variability. But how? A statistic that measures the variability of your estimator is the standard error. You calculate the standard error of the mean using the formula shown here, where *s* is the sample standard deviation and *n* is the sample size. The standard error of our estimate is the standard deviation of the sample data divided by the square root of the total number of sampled data points.



Standard error differs from the sample standard deviation because sample standard deviation deals with the variability of your data whereas standard error deals with the variability of your sample statistic.

Question: Variability in the Sample Statistic

If you draw different samples of the same size from a population, the sample statistic value would not change.

a. True
b. False

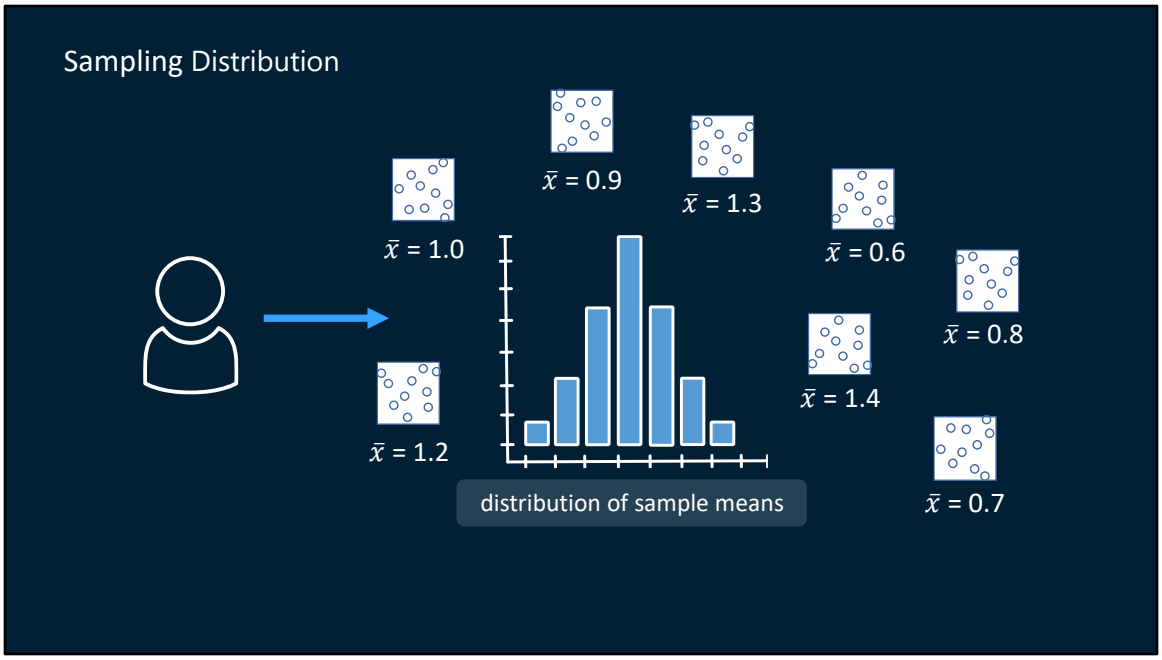
Answer: Variability in the Sample Statistic

If you draw different samples of the same size from a population, the sample statistic value would not change.

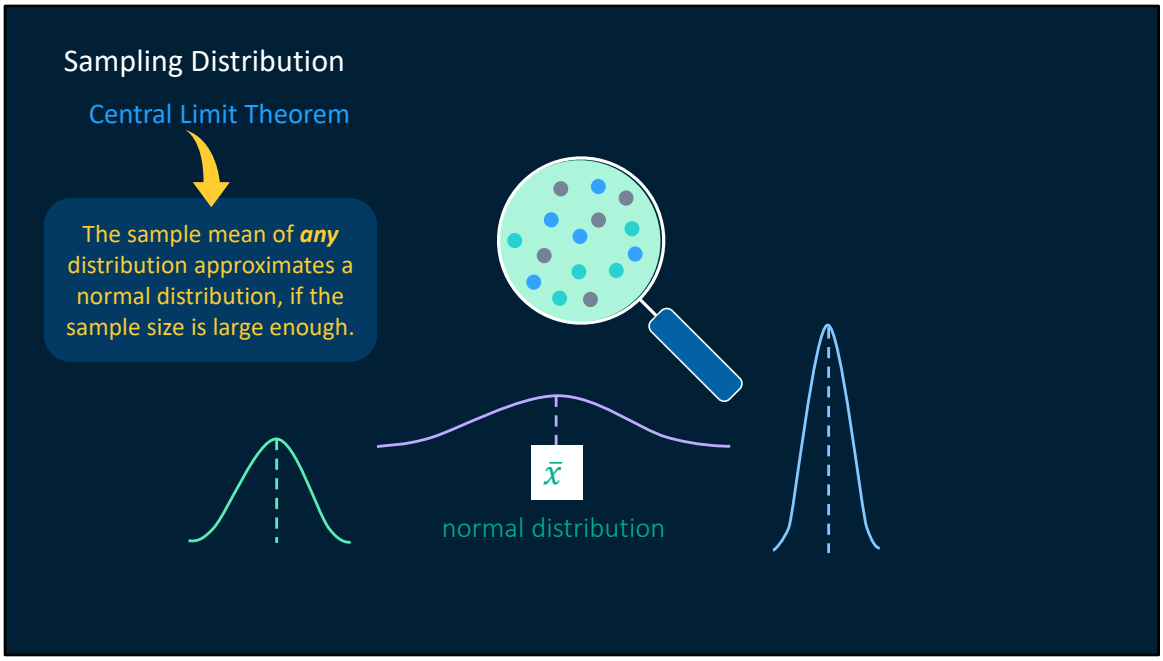
a. True
b. False

Correct answer: False

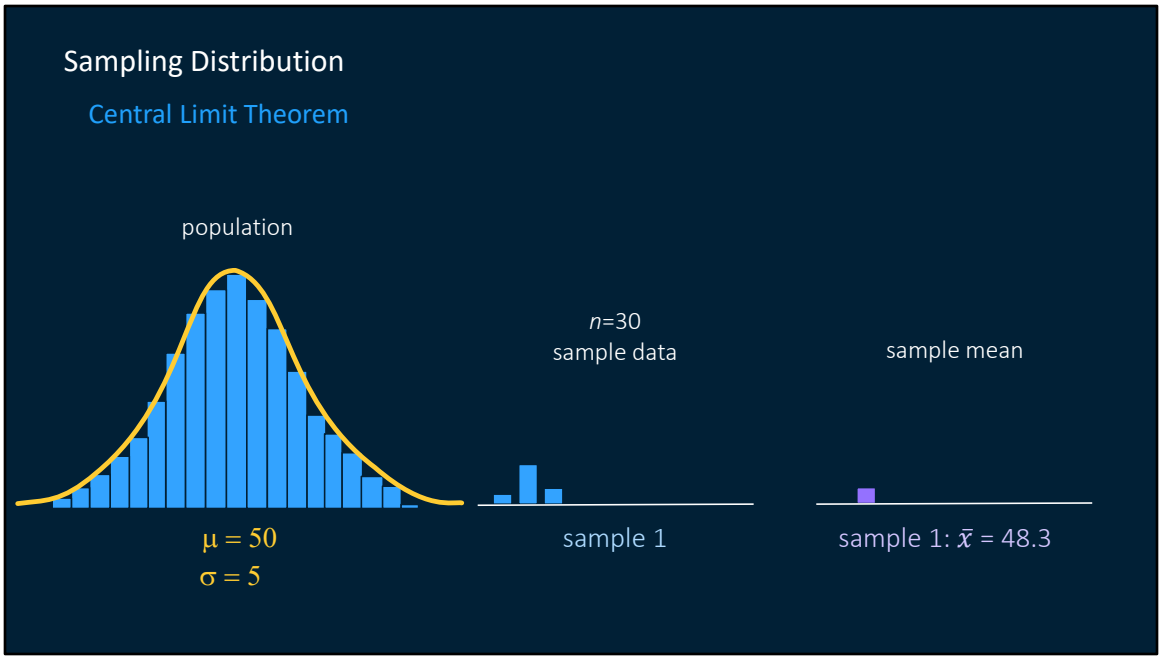
Sample statistics are a direct function of the sample collected. That means if a different sample was collected, the sample statistics would change. An example is the standard error of the mean, which measures the variability of your sample mean.



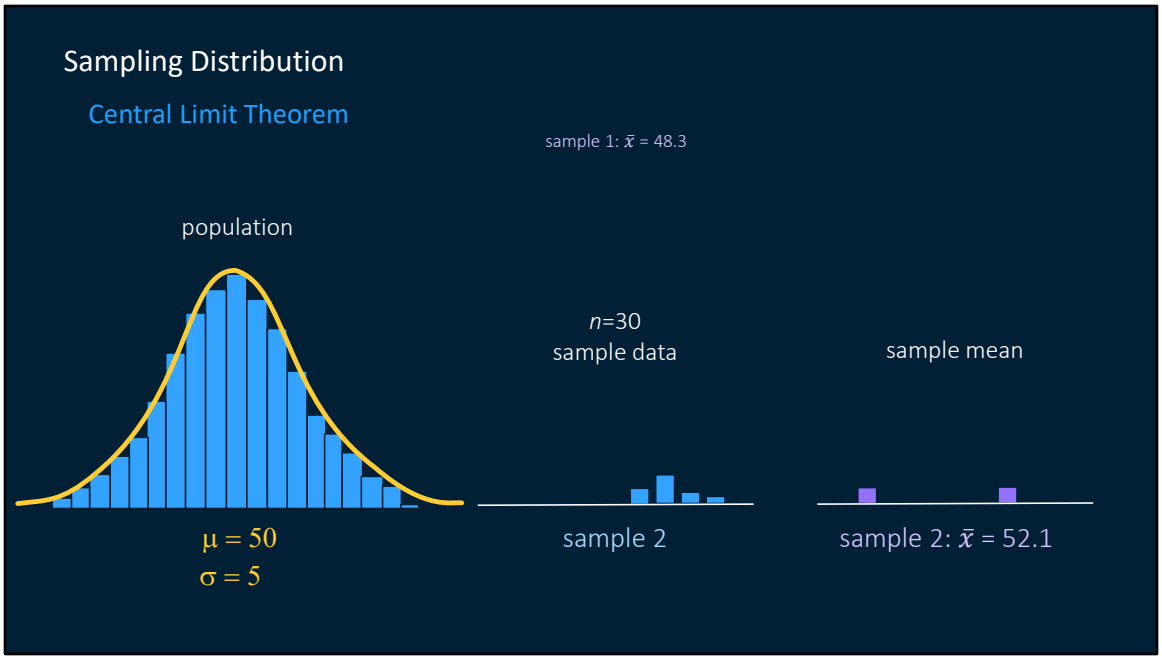
We discussed that sample means differ as samples differ. So how about looking at the distribution of all possible sample means of the population. Sampling distribution is another common concept discussed in the area of statistical inference. Sampling distribution is described as the distribution of sample means.



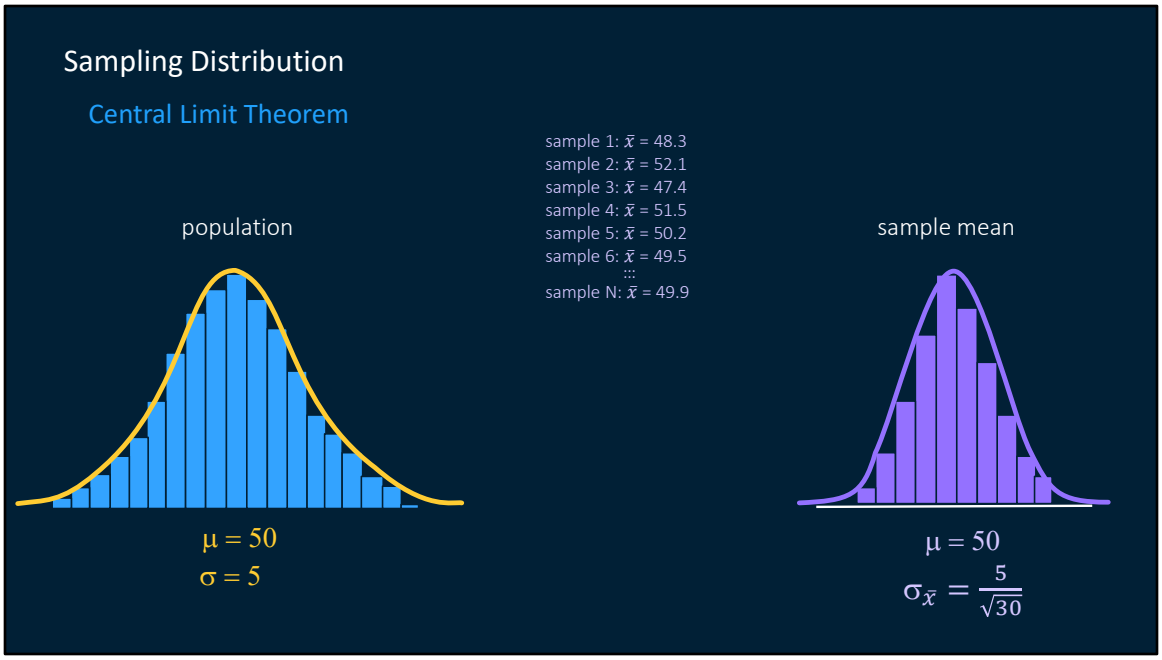
Often, the variability of sample statistics is approximately normal. Perhaps the most famous statistical theorem, the central limit theorem, indicates that the sample mean of ***any*** population distribution approximates a normal distribution, if the sample size is large enough.



For example, imagine that our population distribution is bell shaped, with a population mean, μ=50, and standard deviation, σ=5. Suppose we sample n=30 observations and the sample mean is 48.3.



Then we take another sample of 30 observations, and the sample mean is 52.1. Let's see what happens if we repeat this process many times.

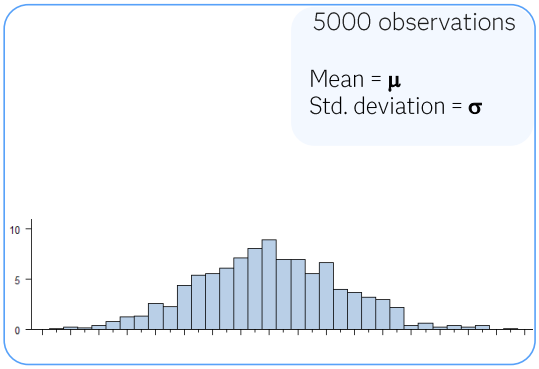


If we repeat this process many times, the distribution of the sample mean will follow a bell-shaped normal distribution with population mean μ=50 and standard error of σ divided by the square root of n=30.

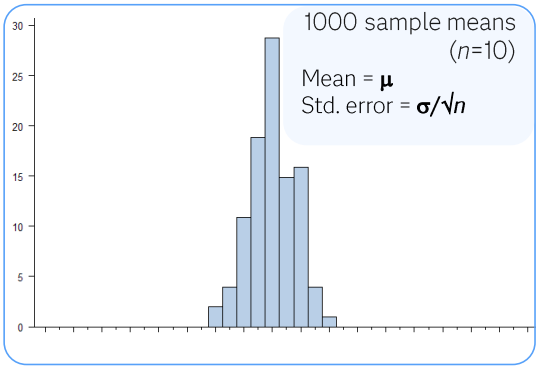
Question: Sampling Distribution

SAT Score is the variable of interest with 5000 observations. You have drawn 1000 samples of size 10. Which of the following is the sampling distribution?

a. Distribution of SAT Score



b. Distribution of Means of SAT Score

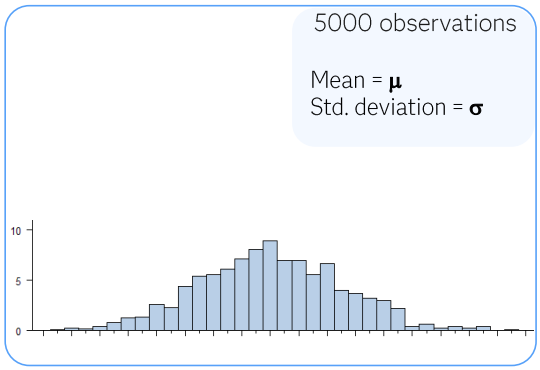


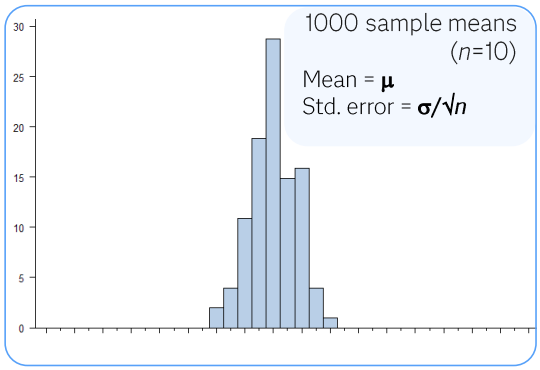
Answer: Sampling Distribution

SAT Score is the variable of interest with 5000 observations. You have drawn 1000 samples of size 10. Which of the following is the sampling distribution?

a. Distribution of SAT Score

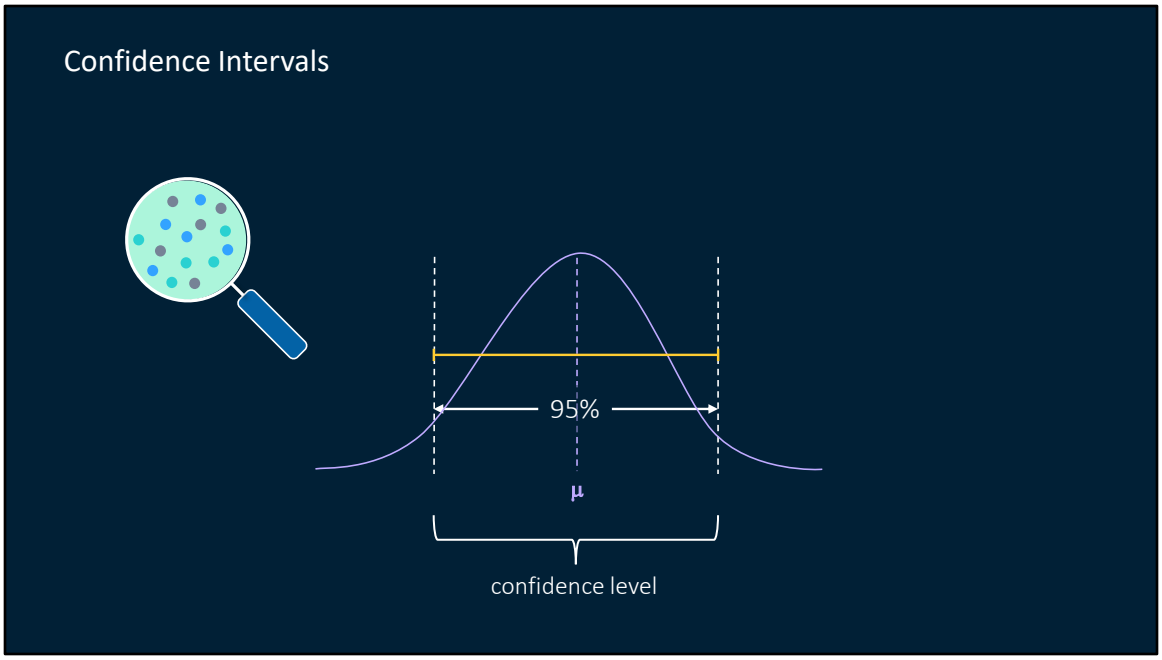
b. Distribution of Means of SAT Score



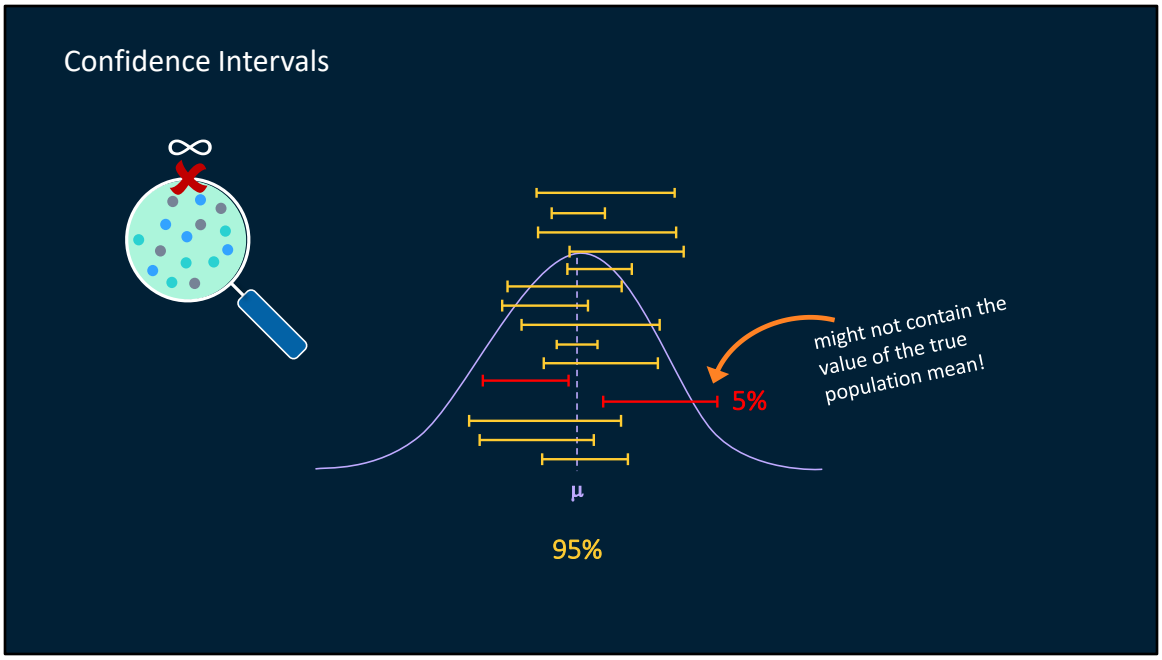


Correct answer: b

Option **a** shows the distribution of random variable **SAT Score** of all 5000 *observations*.
Option **b** represents the distribution of random variable **Means of SAT Score** of the 1000 *sample means*, all with the same sample size of 10. The variability of the distribution of sample means (**$\sigma/\sqrt{n}$**) is smaller than the variability of the distribution of the 5000 observations (**σ**). For example, it seems relatively likely to find one student with an SAT score of 1550 (out of a maximum of 1600), but not likely that a mean of a sample of 10 students would be 1550.



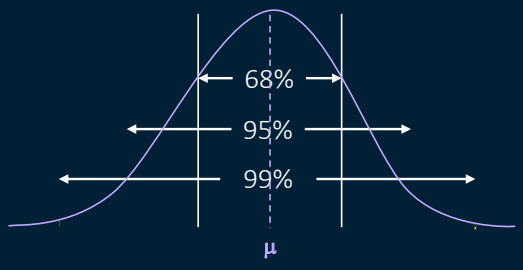
"Confidence intervals for the mean" are interval estimators of the population mean. They consider the variability of the sample statistic--in this case, the sample mean. You get to choose your desired degree of confidence, or confidence level. A typical confidence level is 95%, meaning that 95% of those theoretically infinite number of intervals would contain the true population mean, and 5% would not. The confidence interval shows a range of plausible values for the unknown population mean by reporting an upper and lower bound.



Here's a good way to think about confidence intervals. If you were to draw infinitely many samples and estimate your confidence interval exactly the same way each time, your confidence interval would represent the percentage of those intervals that would contain the true population mean. Of course, nobody draws infinitely many samples. It's rare that you would draw more than one.

An important thing to remember is that the confidence interval calculated for a particular sample might or might not contain the value of the true population mean.

Confidence Intervals

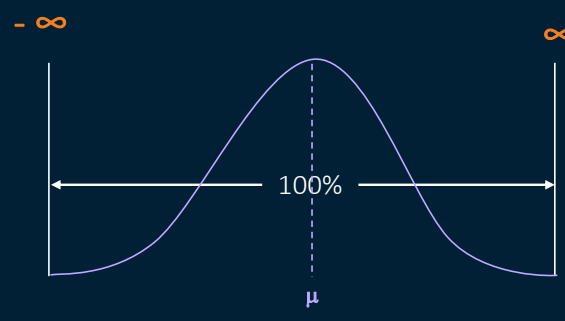


As you increase the confidence level, the width of the interval increases, making it less informative.

To construct a confidence interval for the mean, you must first select a significance level so that the appropriate value of t is used in the formula. Higher confidence levels are associated with larger t values, which in turn result in wider intervals. The formula calculates both the upper and lower bounds, equidistant from the sample mean estimate in the middle.

Why not raise the confidence level to 99.9%, so that any confidence interval you calculate contains the true value of the population mean? As you increase the confidence level, the width of the interval increases, making it less informative.

Confidence Intervals

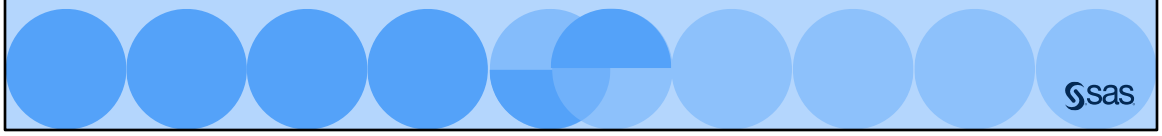


As you increase the confidence level, the width of the interval increases, making it less informative.

In the extreme, a 100% confidence interval for the mean for any sample ranges from negative infinity to positive infinity, which would tell you nothing useful about where the true population mean lies.

Confidence Interval for Mean

To explore this learning resource in greater detail, please visit the course page on learn.sas.com. There, you'll find comprehensive course materials, tutorials, and updates to support your learning journey.



Question: Interval Estimates

A researcher calculates 95% confidence interval for population mean as ($100, $130). What could be the most plausible value of the sample mean?

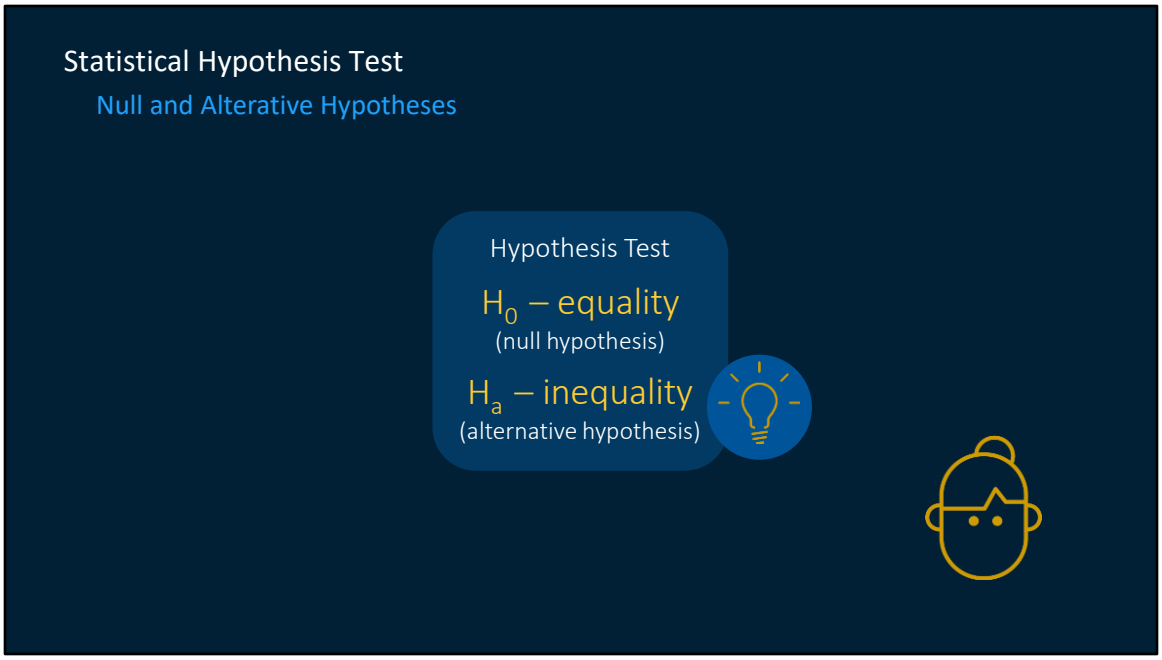
a. $100
b. $115
c. $125
d. $230

Answer: Interval Estimates

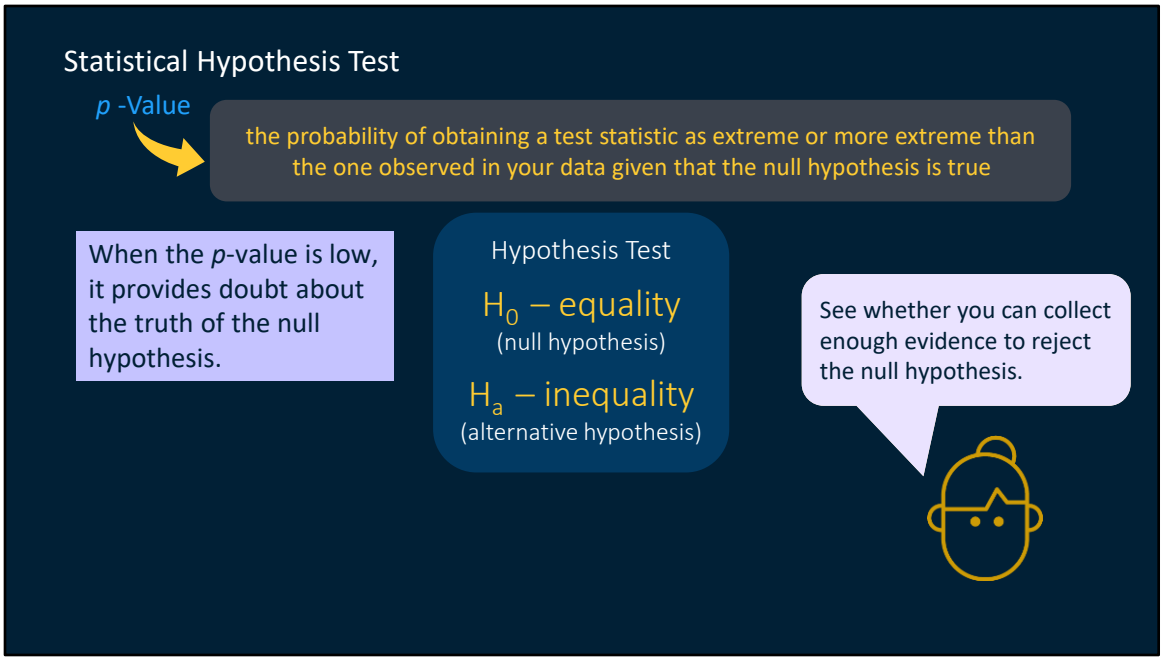
A researcher calculates 95% confidence interval for population mean as ($100, $130). What could be the most plausible value of the sample mean?

a. $100
b. $115
c. $125
d. $230

Correct answer : b
Recall the formula for calculating the confidence interval for the mean. It requires you to add and subtract (t times $S_{\bar{x}}$) from the sample mean $\bar{x}$ to obtain lower and upper bound of confidence interval.

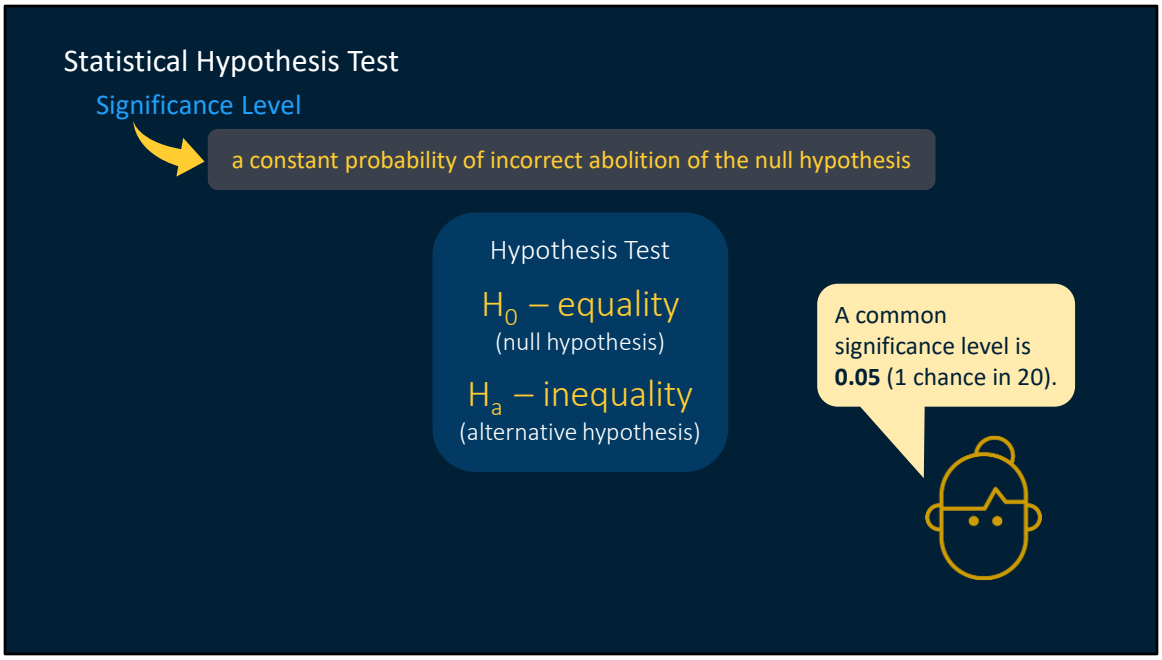


In a hypothesis test, you phrase questions as tests of hypotheses about population parameters. Your null hypothesis is usually one of equality, whereas the alternative hypothesis is one of inequality. The alternative hypothesis is typically what you suspect, or what you're attempting to demonstrate.

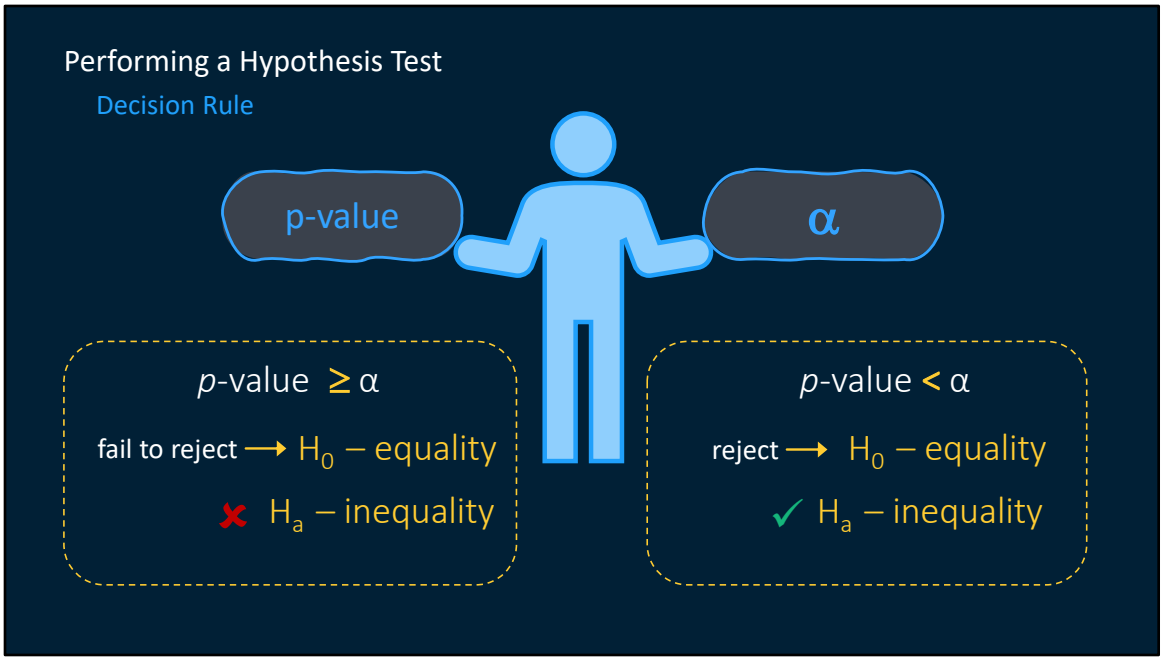


The conclusions reached from your hypothesis test are usually phrased in reference to the *p*-value, or the probability of obtaining a test statistic as extreme or more extreme than the one observed in your data given that the null hypothesis is true.

In a hypothesis test, you examine the data to see whether you can collect enough evidence to reject the null hypothesis. The *p*-value is an indication of the evidence collected. When the *p*-value is low, it provides doubt about the truth of the null hypothesis. But how low does the *p*-value need to be before you reject the null hypothesis completely? That depends on you.



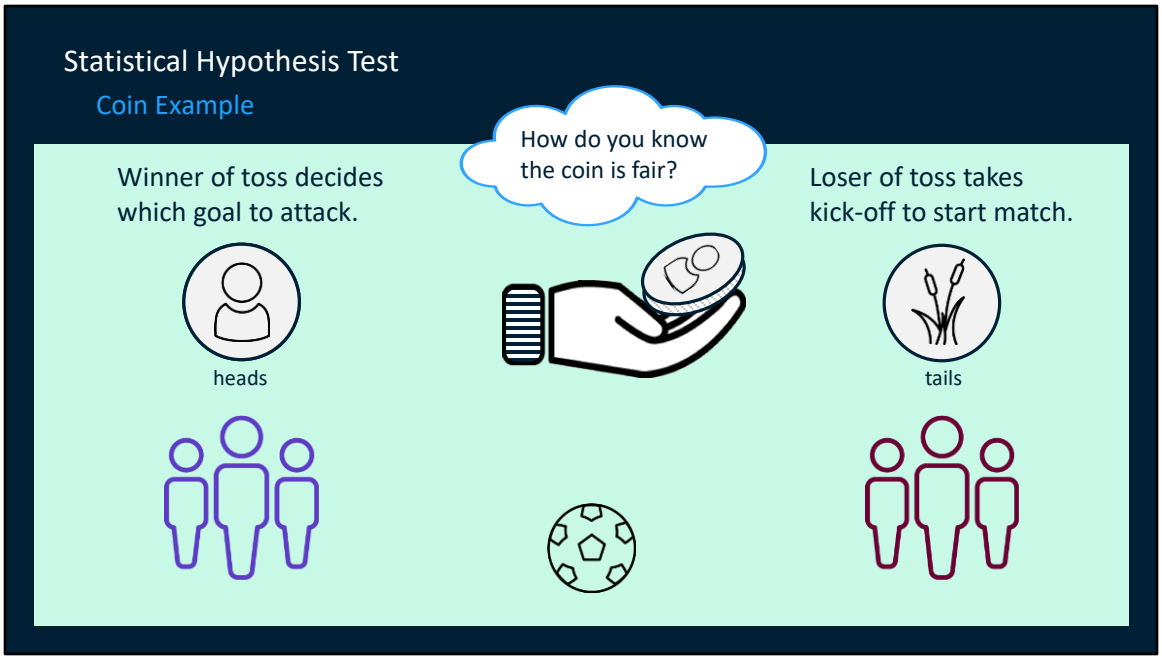
A common significance level is 0.05 (1 chance in 20). If you require a stricter cutoff, you might consider lowering your significance level when planning your analysis. The level of significance refers to a constant probability of incorrect abolition of the null hypothesis.



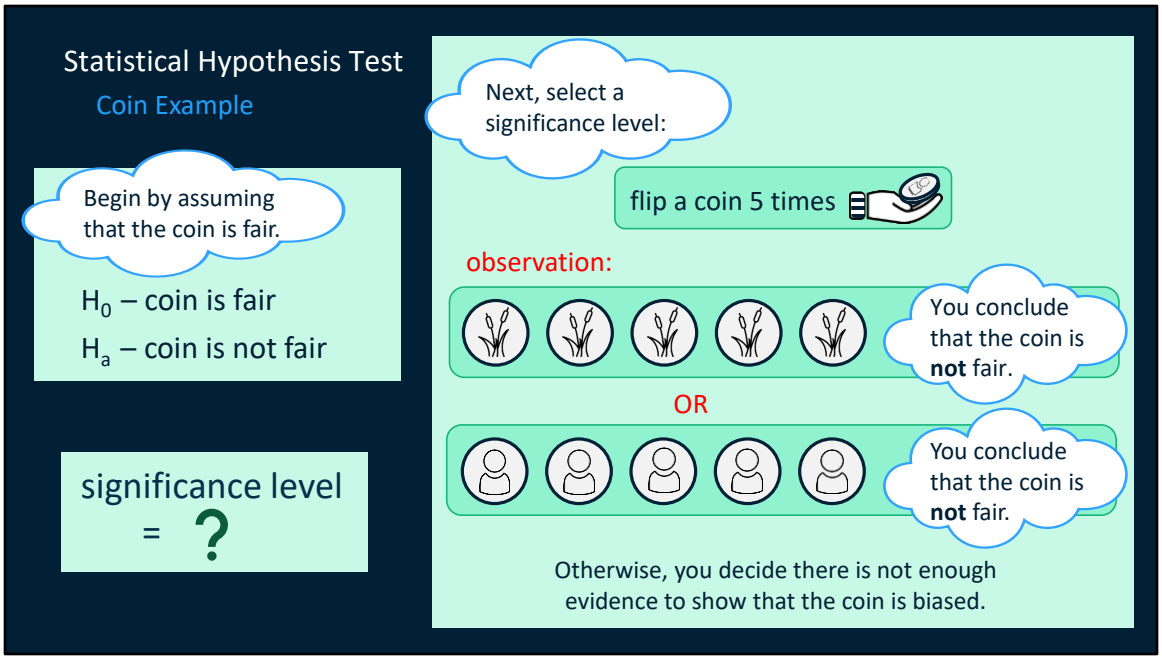
In practice, the level of significance, alpha, is stated in advance to compare it with the *p*-value.

If the *p*-value is less than alpha, you reject the null hypothesis in favor of the alternative. On the other hand, if your *p*-value is greater than or equal to alpha, evidence suggests that your null hypothesized value is statistically reasonable, so you fail to reject the null hypothesis.

Let's look at an example.

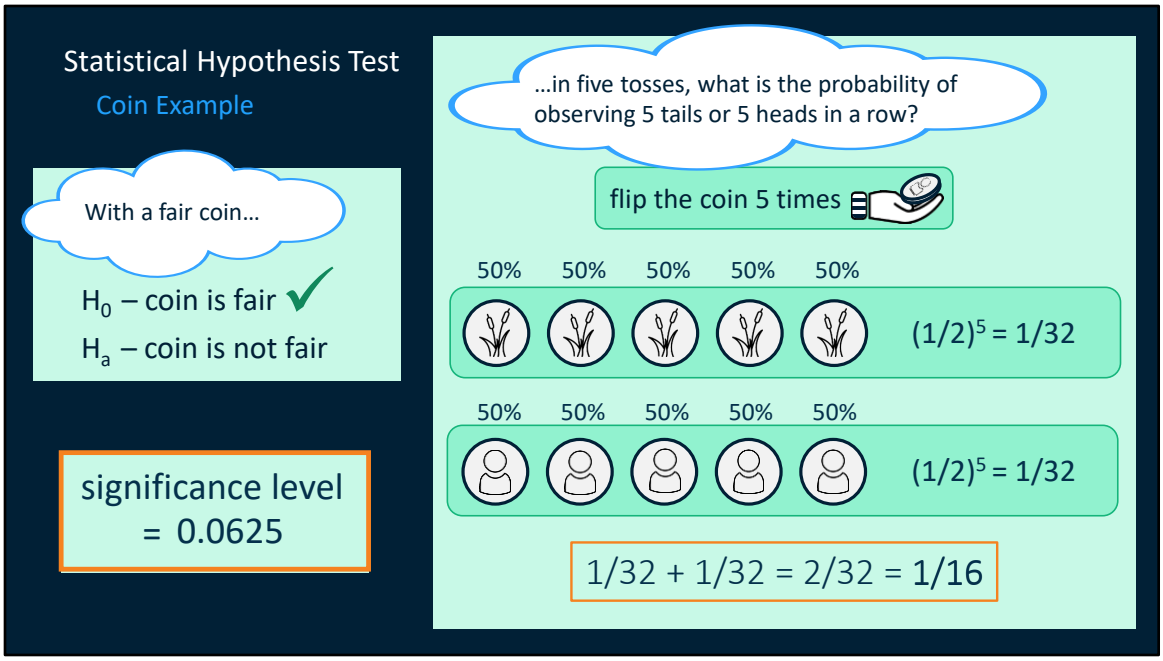


To begin each game of a worldwide soccer tournament, one of the teams chooses heads or tails, and the match referee tosses a coin. The team that wins the toss decides which goal it will attack, and the team that loses the toss takes the kick-off to start the match. But how do you know the coin is fair?



You might suspect that the coin is not fair (alternate hypothesis), but you begin by assuming that it is fair (null hypothesis).

Next, you select a significance level: if you flip a coin five times and observe five heads in a row or five tails in a row, you conclude that the coin is not fair. Otherwise, you decide that there is not enough evidence to show that the coin is biased.



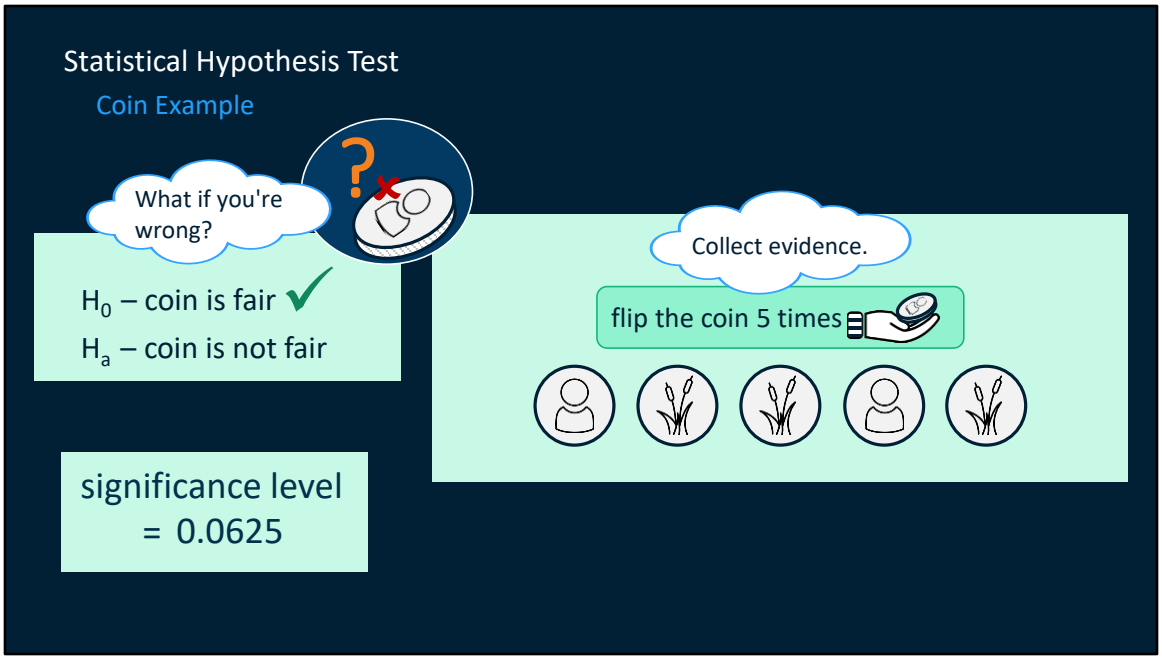
With a fair coin (a true null hypothesis), in five tosses (five trials), what is the probability of observing five tails or five heads in a row?

Each toss has a 50% probability of being tails. Tosses are independent, and therefore, the probability of five tails is (1/2) to the fifth power - or one out of 32.

In the same way, the probability of five heads is also one out of 32.

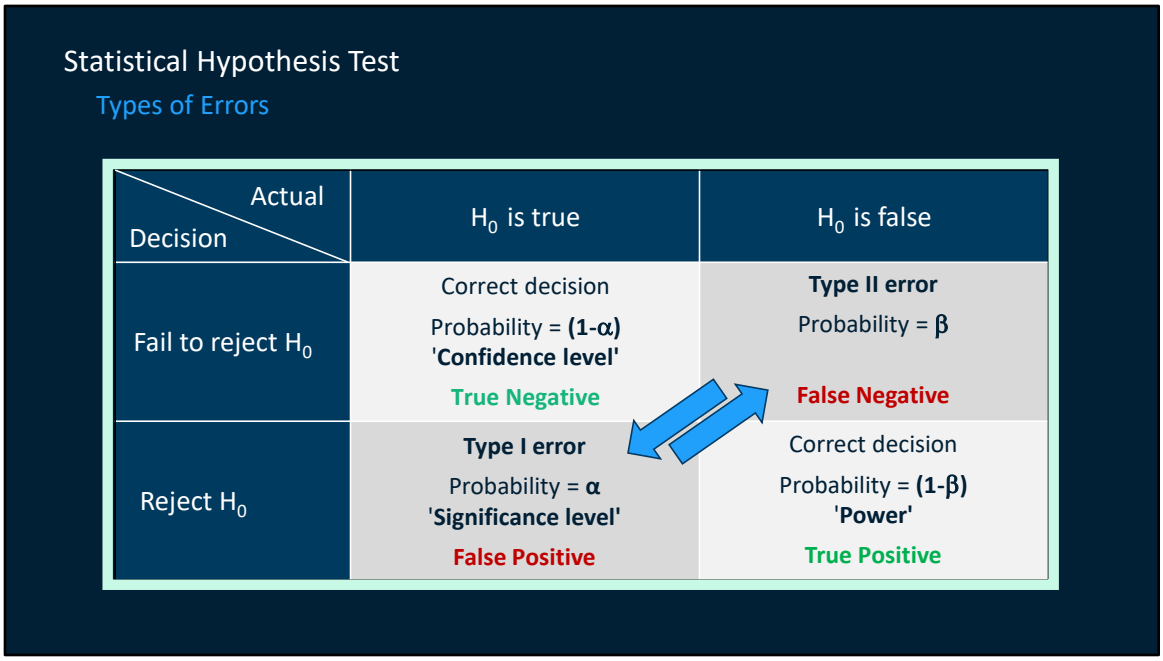
These probabilities can be added together to give the probability of five tails or five heads as two out of 32, or one out of 16.

So, the significance level is 1/16, or 0.0625.



To collect evidence, you flip the coin five times and count the number of heads and tails. Finally, you decide either that there is enough evidence to reject the assumption that the coin is fair (either all trials are heads, or all trials are tails), or that there is not enough evidence to reject the assumption that the coin is fair (meaning that not all trials are either heads or tails). So, in this example, we say that "we fail to reject the null hypothesis that the coin is fair".

So, you performed a hypothesis test and used a decision rule to decide whether the coin was fair or not. But was your decision correct? You began by assuming that the null hypothesis is true: that the coin is fair. But what if you're wrong?



If you reject the null hypothesis when it's actually true, you've made a Type I error.

The probability of committing a Type I error is α. α is the significance level of the hypothesis test. In the coin example, it's the probability that you conclude that the coin is not fair when it is fair. In machine learning terminology, it is often referred to as a *false positive* situation.

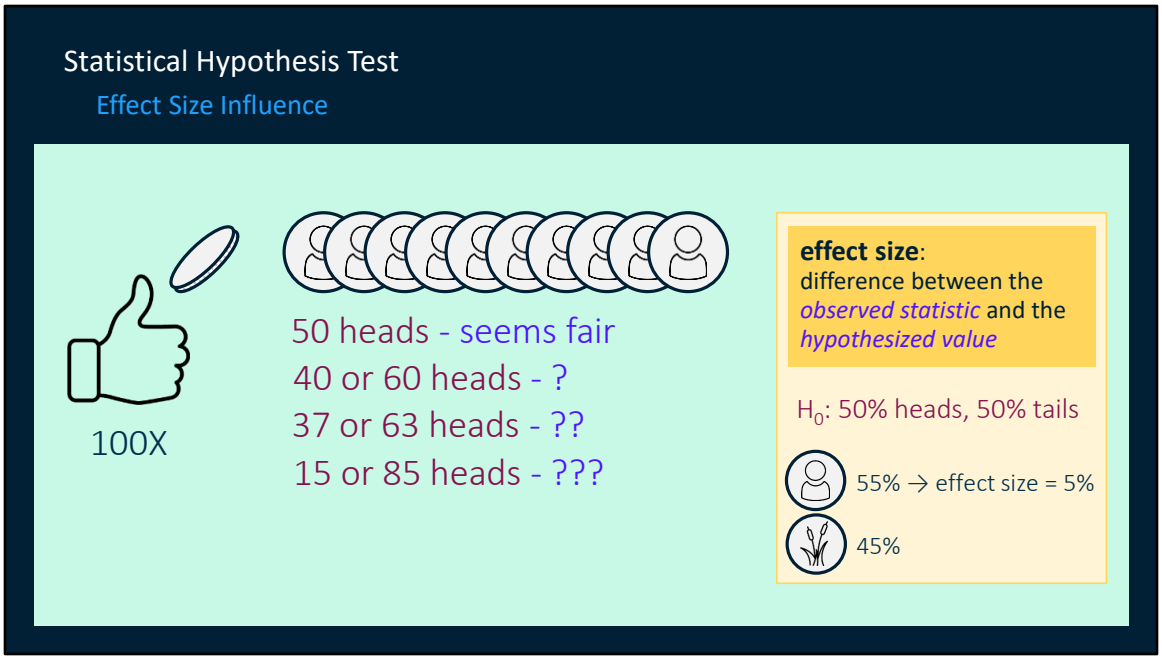
A *Type II error*, is when you fail to reject the null hypothesis and it's actually false. The probability of committing a Type II error is β. In the coin example, it's the probability that you fail to find that the coin is not fair when it is in fact biased. In machine learning terminology, it is often referred to as a *false negative* situation.

Type I and Type II errors are inversely related. As one type increases, the other decreases.

You always want to make the correct decisions.

The probability that you correctly reject the null hypothesis is equal to (1 – β), where again, β is Type II error rate. This is called the *power* of a statistical test.

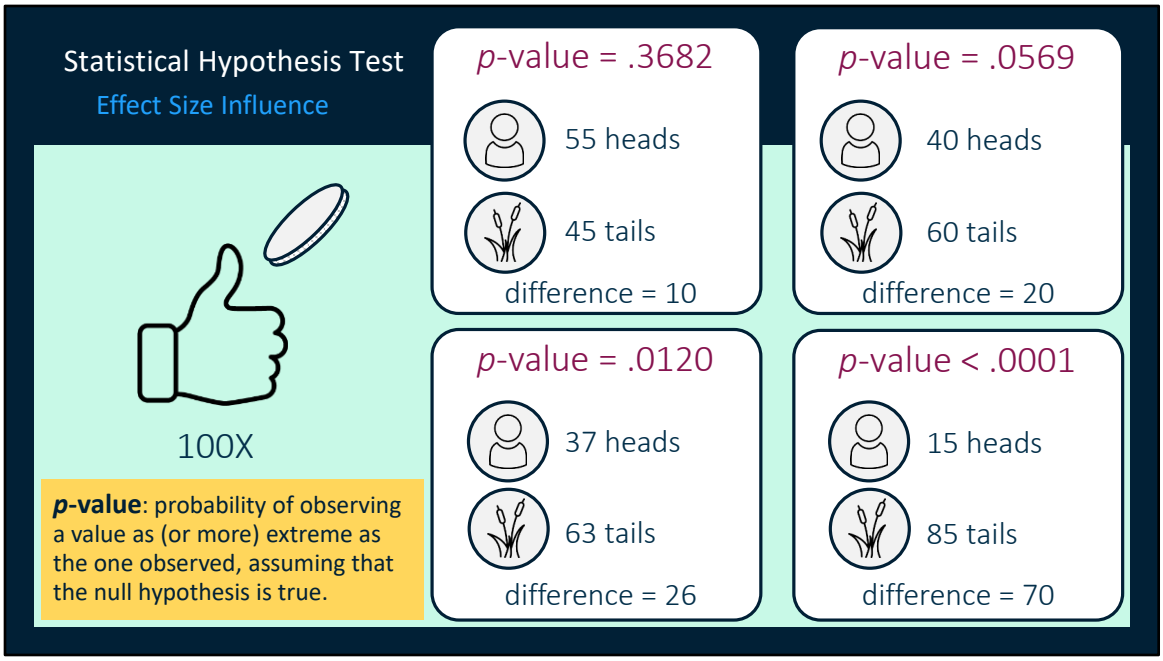
The probability that you correctly do not reject the null hypothesis is equal to (1 – α), where again, α is Type I error rate. This is called the *confidence level*.



If you flip a coin 100 times and observe 50 heads, you wouldn't doubt that the coin is fair.

But you might be skeptical if you observe 40 or 60 heads. You'd be more skeptical if you observe 37 or 63 heads, and you'd be highly skeptical if you observe 15 or 85 heads. As the difference between the number of heads and tails increases, you have more evidence that the coin is not fair. Statisticians refer to the difference between the observed statistic and the hypothesized value as the effect size.

The null hypothesis of a fair coin suggests 50% heads and 50% tails. If the coin was actually weighted (warped or modified in some manner) to give 55% heads, the effect size would be 5%.



A *p*-value measures the probability of observing a value as extreme as the one observed or more extreme, assuming the null hypothesis is true.

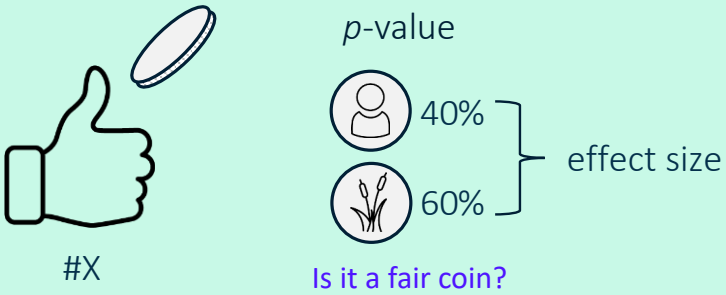
Suppose you flip the coin 100 times, and you observe 55 heads and 45 tails. The difference of 10 is associated with a *p*-value of .3682. *p*-values this large are often seen in experiments with a fair coin.

As the difference between heads and tails gets larger (for example, 20, 26, or as high as 70), the corresponding *p*-values get smaller.
You would rarely see a small *p*-value (for example, less than .0001) with a fair coin. In the case of 15 heads and 85 tails, you have evidence that the coin is not fair. The chance of getting 85 tails or even more extreme results is less than one in 10,000 if the null hypothesis is true. So, either the null hypothesis is false or something extremely unlikely happened. The typical interpretation is that the null is false. This can be verified with further experiments.

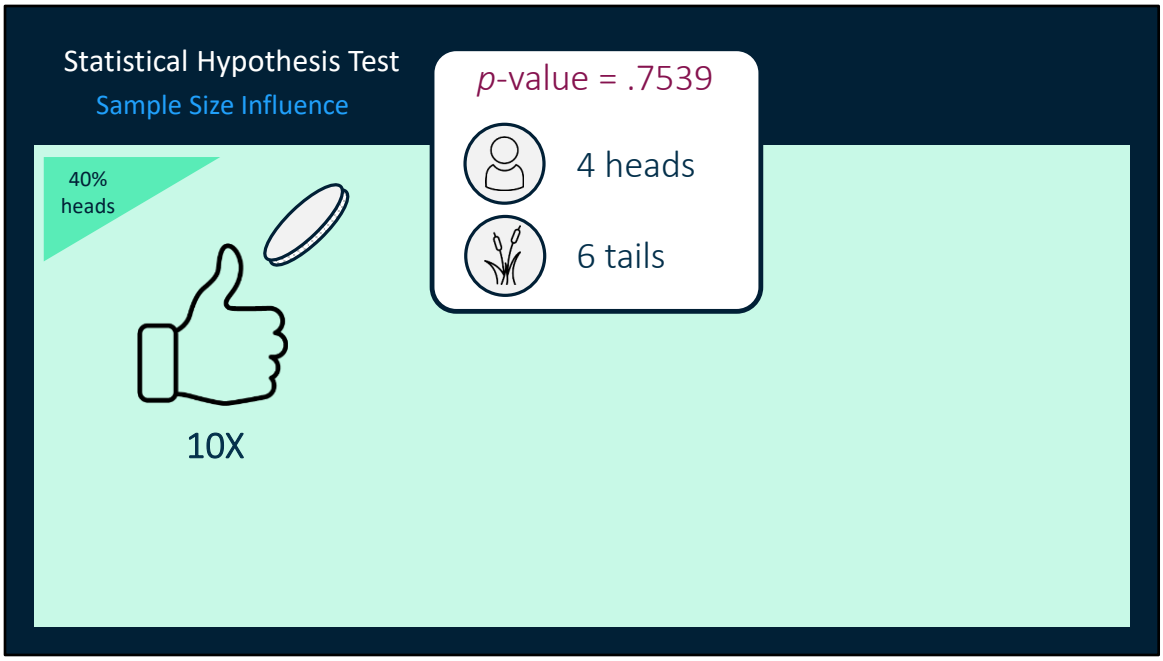
Consequently, the *p*-value is used to determine statistical significance. It helps you assess whether you should reject the null hypothesis.

Statistical Hypothesis Test

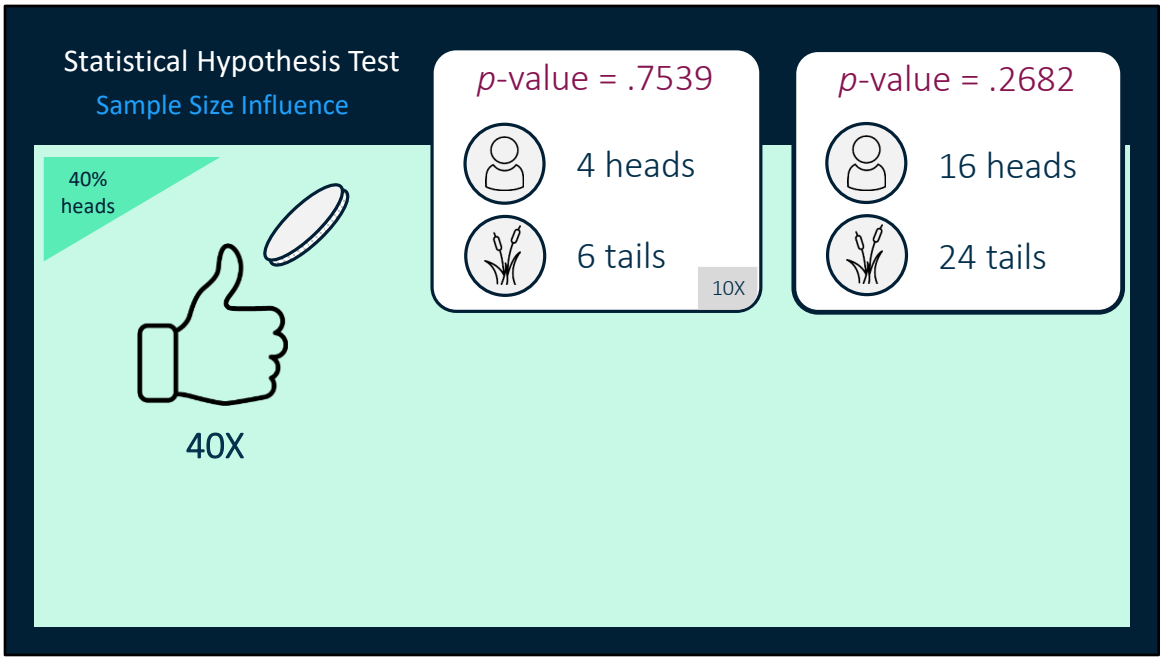
Sample Size Influence



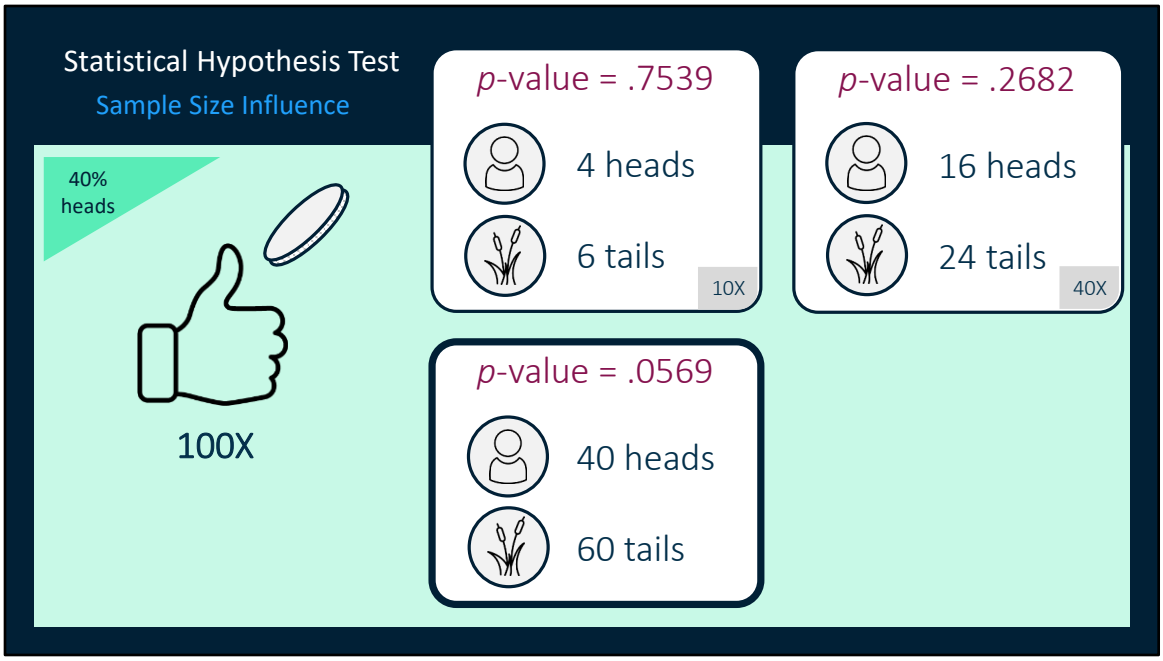
A *p*-value is not only affected by the effect size (in this case, the observed proportion of heads). It's also affected by the sample size (in this case, the number of coin flips). For a fair coin, you'd expect 50% of the flips to be heads. What if you get 40% heads instead of the 50% you expect? Is it a fair coin?



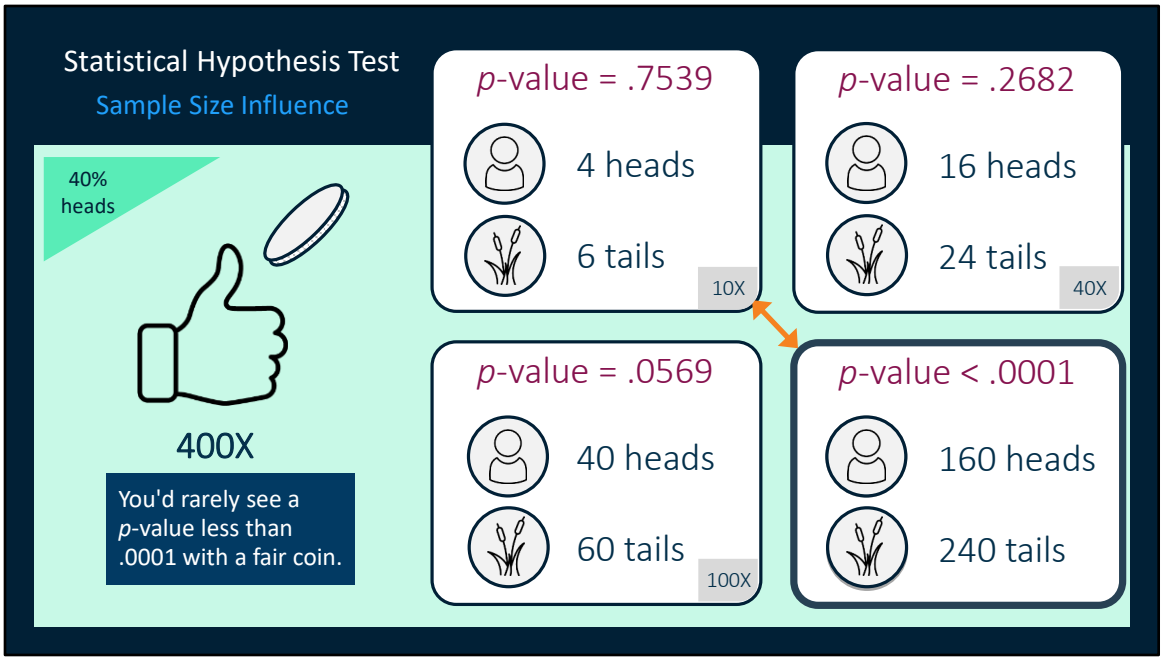
Let's say that you flip the coin 10 times and observe 40%, or four heads. The resulting *p*-value is .7539.



When you flip the coin 40 times and observe 40%, or 16 heads. The resulting *p*-value has decreased from .7539 to .2682.

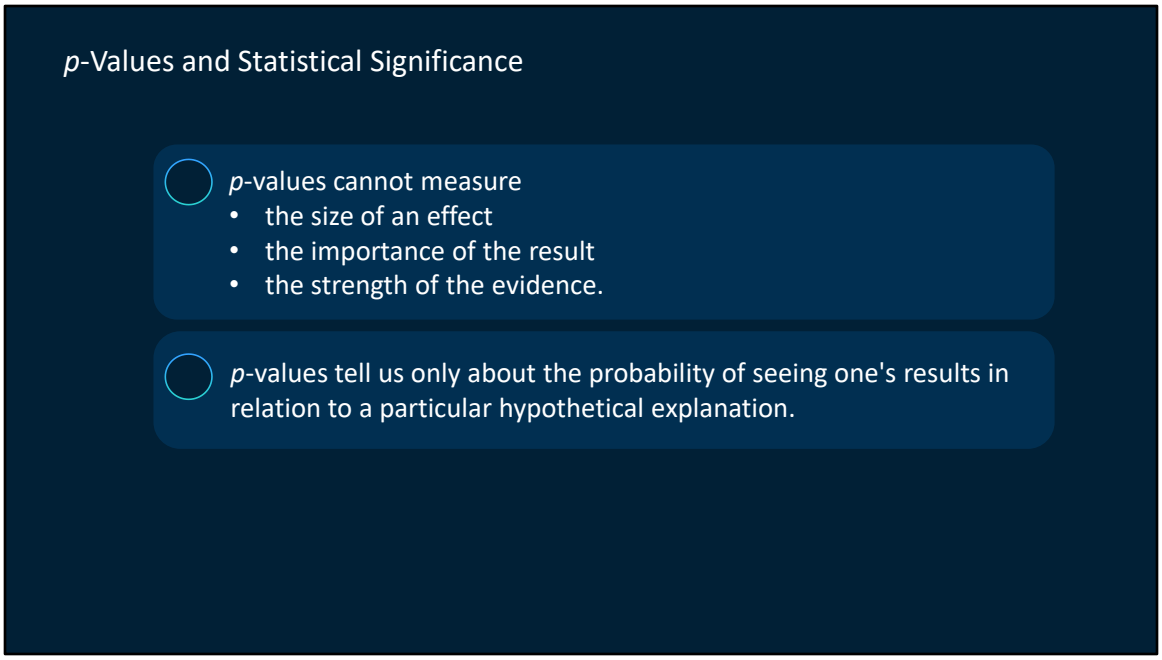


When you flip the coin 100 times and observe 40%, or 40 heads. The resulting *p*-value is further decreased to .0569.



What if you flip the coin 400 times and you observe 40% heads? The evidence becomes stronger, and the *p*-values become smaller, as the number of trials increases. As you saw when we talked about confidence intervals, the variability around the mean estimate gets smaller as the sample size gets larger. For larger sample sizes, you can measure means more precisely. Therefore, 40% heads out of 400 flips makes you more confident that this was not just a chance difference, when compared to 40% heads out of only 10 flips.

The smaller *p*-values reflect this confidence. The *p*-value here, less than .0001, assesses the probability that this difference from 50% occurred purely by chance. Remember, as you saw earlier, you'd rarely see a *p*-value less than .0001 with a fair coin.



p-values and statistical significance are often misunderstood and misused. Be wary that *p*-values cannot measure the size of an effect or the importance of the result or the strength of the evidence. However, it continues to be a tool that separates true findings from false ones.

p-values (significance probabilities tell us only about the probability of seeing one's results in relation to a particular hypothetical explanation.

Question: Hypothesis Testing Terminology

Which of the following terms can be defined as the difference between the observed statistic and the hypothesized value?

a. effect size
b. type I error
c. alpha
d. power

Answer: Hypothesis Testing Terminology

Which of the following terms can be defined as the difference between the observed statistic and the hypothesized value?

a. effect size
b. type I error
c. alpha
d. power

The correct answer is a.
Effect size refers to the difference between the observed statistic and the hypothesized value. As the difference increases, you have more evidence against the null hypothesis.
A Type I error occurs if you reject the null hypothesis when it's actually true.
Alpha is the probability of committing a Type I error.

Question: Type I and Type II Errors

Keeping the sample size constant, which of the following statements are true regarding type of errors? (Select all that apply)

a. The Type I and Type II error rates do not influence each other.
b. Type I and Type II errors are inversely related.
c. A higher confidence level leads to increasing the power of the test.
d. By setting the Type I error rate, you indirectly influence the size of the Type II error rate as well.

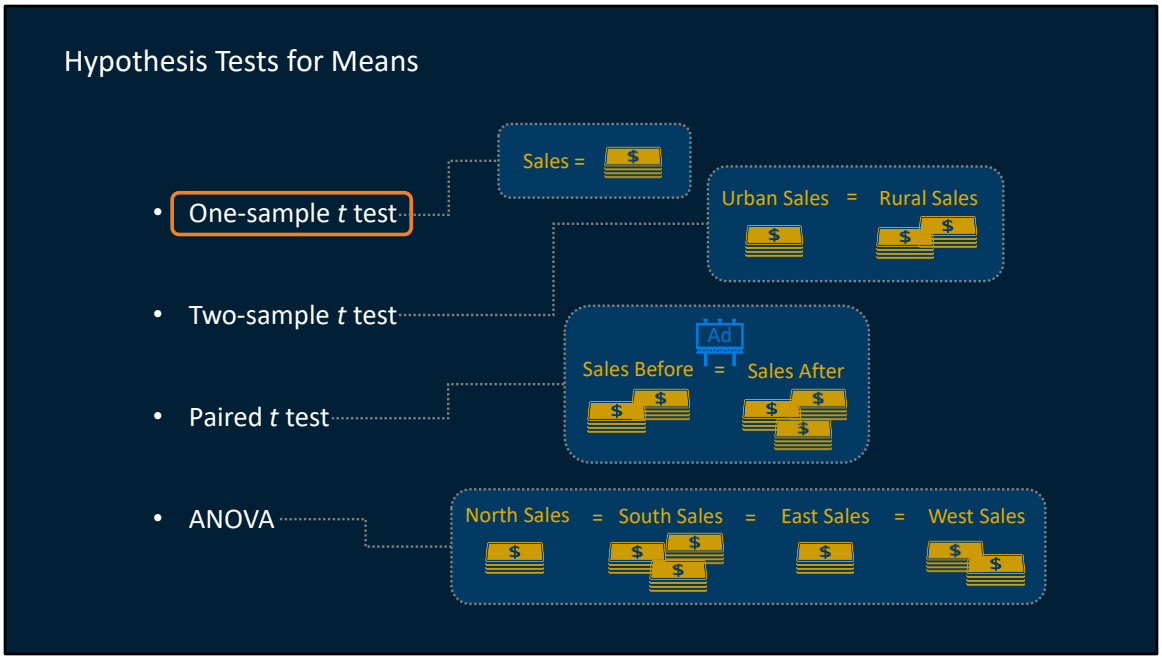
Answer: Type I and Type II Errors

Keeping the sample size constant, which of the following statements are true regarding type of errors? (Select all that apply)

a. The Type I and Type II error rates do not influence each other.
b. Type I and Type II errors are inversely related.
c. A higher confidence level leads to increasing the power of the test.
d. By setting the Type I error rate, you indirectly influence the size of the Type II error rate as well.

The correct answer is b and d.

Type I and Type II errors are inversely related. As one type increases, the other decreases.
A higher confidence level is a resultant of setting a lower significance level. Consequently, the Type II error rate would increase and leads to lowering the power of the test.



There are several hypothesis tests for means, including one-sample t test, two-sample t test, paired t test, and ANOVA.

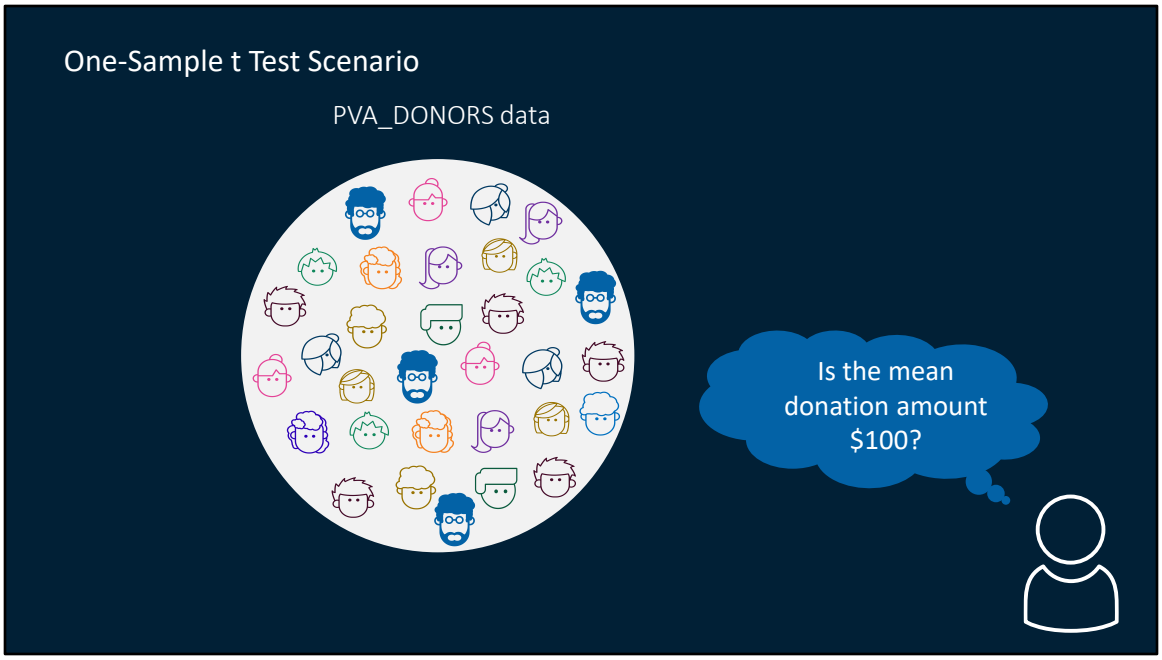
A *one-sample* t *test* compares the mean calculated from a sample to a hypothesized value. For example, sales are equal to some dollar amount.

If you want to compare the means of two different groups, you can conduct a *two-sample* t *test*. For example, rural sales are equal to urban sales.

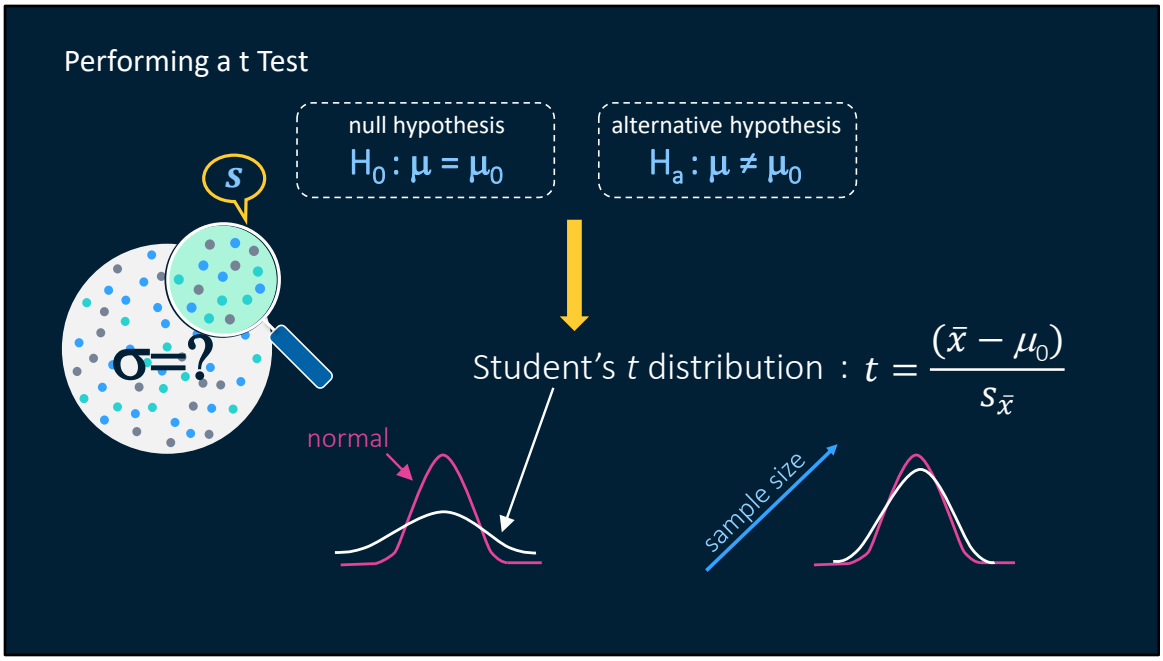
For many types of data, repeat measurements are taken on the same subject throughout a study. For example, sales before advertising and sales after advertising. The simplest form of this study is often referred to as the *paired* t *test*.

Analysis of Variance (ANOVA) is a statistical technique used to compare the means of two or more groups. For example, sales of north, south, east and west regions are equal.

In this course, we focus on the one-sample *t* test.



Let's look at another example of hypothesis testing. The veterans' organization receives different amounts of donations from different individuals. The organization believes that the mean donation amount is $100, and you'd like to verify this.



You're testing the null hypothesis, μ is equal to μ0, against the alternative hypothesis, μ is not equal to μ0. When you don't know the true population standard deviation, σ, you must estimate it from the sample. Then you must also use Student's t distribution, rather than the normal distribution, for calculating *p*-values and confidence limits. Student's t distribution is a probability distribution that is used to estimate population parameters when the true population variance is unknown or the sample size is appropriately small, or both.

Student's *t* distribution is similar to the normal distribution, but it has more probability in the tails and is not as peaked as the normal distribution.

Student's *t* distribution approaches the normal distribution (Z statistic) as the sample size increases. SAS always reports the t distribution. You calculate the value of Student's *t* statistic using the equation t = (xbar-mu0) all divided by the standard error.

Performing a t Test

PVA_DONORS data

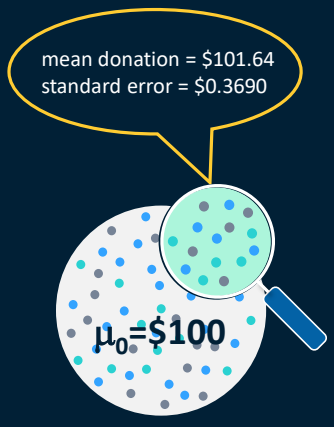
$$t = \frac{(\bar{x} - \mu_0)}{s_{\bar{x}}}$$

$$t = \frac{(101.64 - 100.00)}{0.3690}$$

$$= 4.44$$

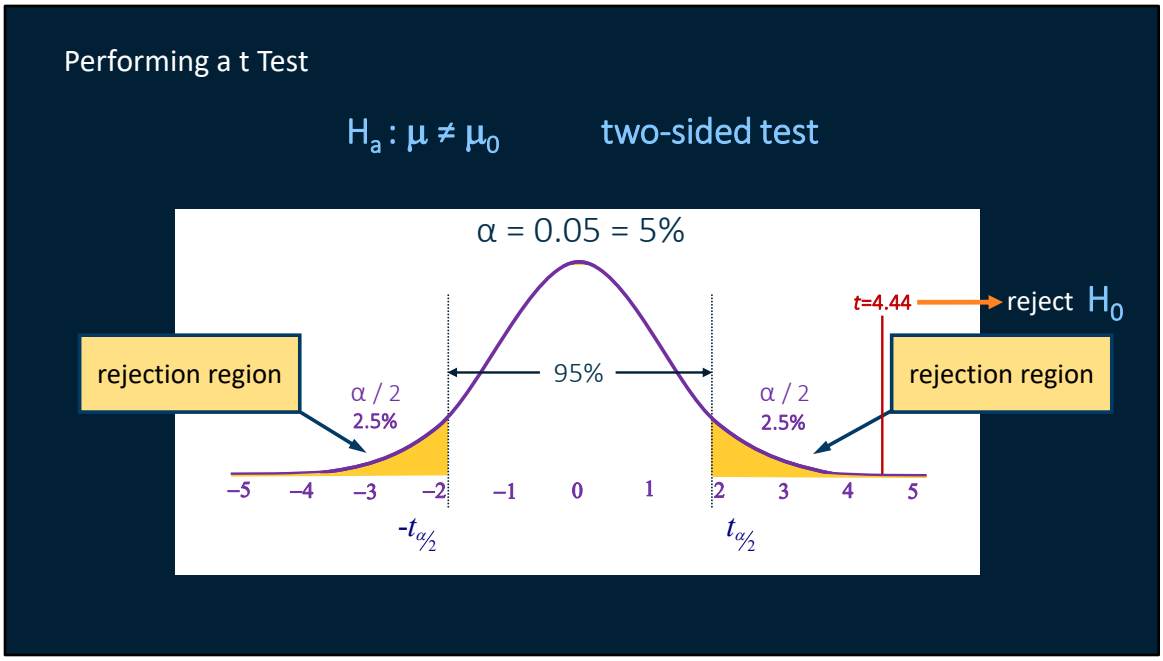
likelihood of obtaining the sample mean

p-value ⬅ probability



Let's calculate Student's *t* statistic for the PVA_DONORS data to test our null hypothesis. Mu0 is the hypothesized value of $100.00, xbar is the sample mean of **Donation Amount**, $101.64, and s sub xbar is the standard error of the mean, $0.369. The resulting *t* value is 4.44.

Before you can decide about your hypothesis, you need to know the probability of observing a test statistic of t=4.44 or more extreme, given that the null hypothesis is true. This probability is the *p*-value. It quantifies how likely we are to obtain the sample mean of $101.64 when the null hypothesis of population donation amount $100 is true.



Let's look at the *t* distribution. Our alternative hypothesis is an inequality, so this is a two-sided test. The shaded area in the graph is the rejection region. For a two-sided test, the rejection region is contained in both tails, and the area in each tail corresponds to alpha/2, or in this case, 2.5%. If the *t* value falls in the shaded region, then you reject the null hypothesis. Otherwise, you fail to reject it.

Our *t* value, 4.44, falls inside the rejection region, so we reject our null hypothesis. The alpha and *t* distribution mentioned here are directly related to those in confidence intervals. In our example, alpha is 0.05, or 5%. The area outside the confidence interval corresponds to the shaded region in the *t* distribution, or the rejection region.

Question: t Distribution

As the sample size increases, the t-distribution becomes more similar to the normal distribution.

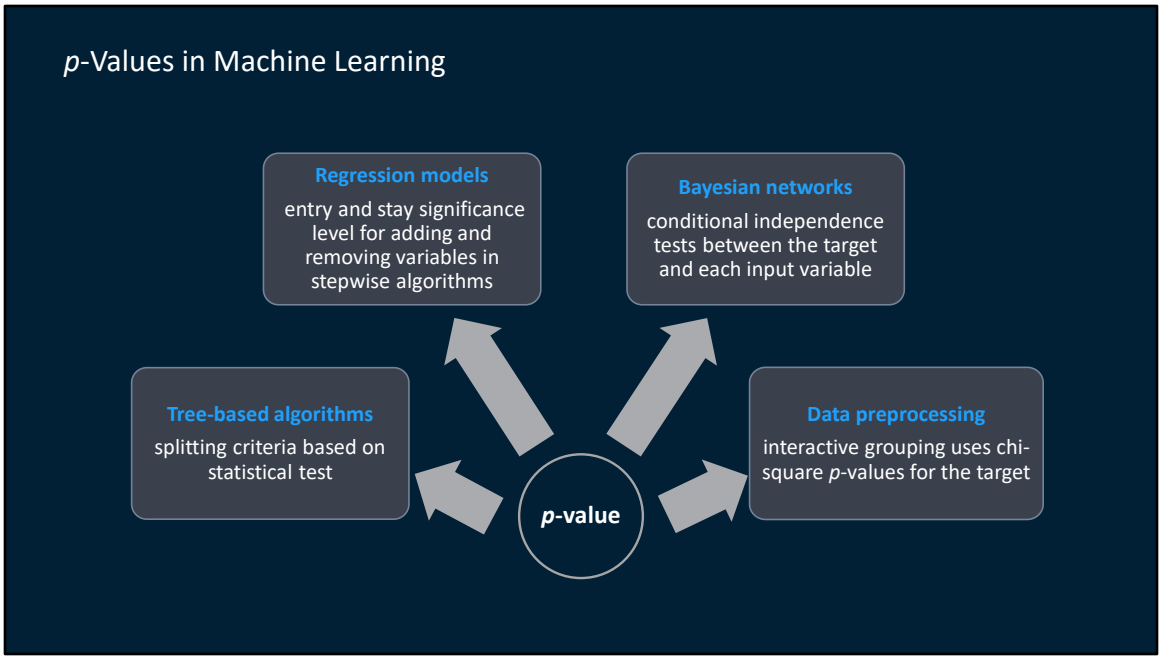
a. True
b. False

Answer: t Distribution

As the sample size increases, the t-distribution becomes more similar to the normal distribution.

a. True
b. False

Correct answer : True
The shape of t-distribution is affected by the sample size. If the sample size increases, then t-distribution gets closer to the normal distribution.



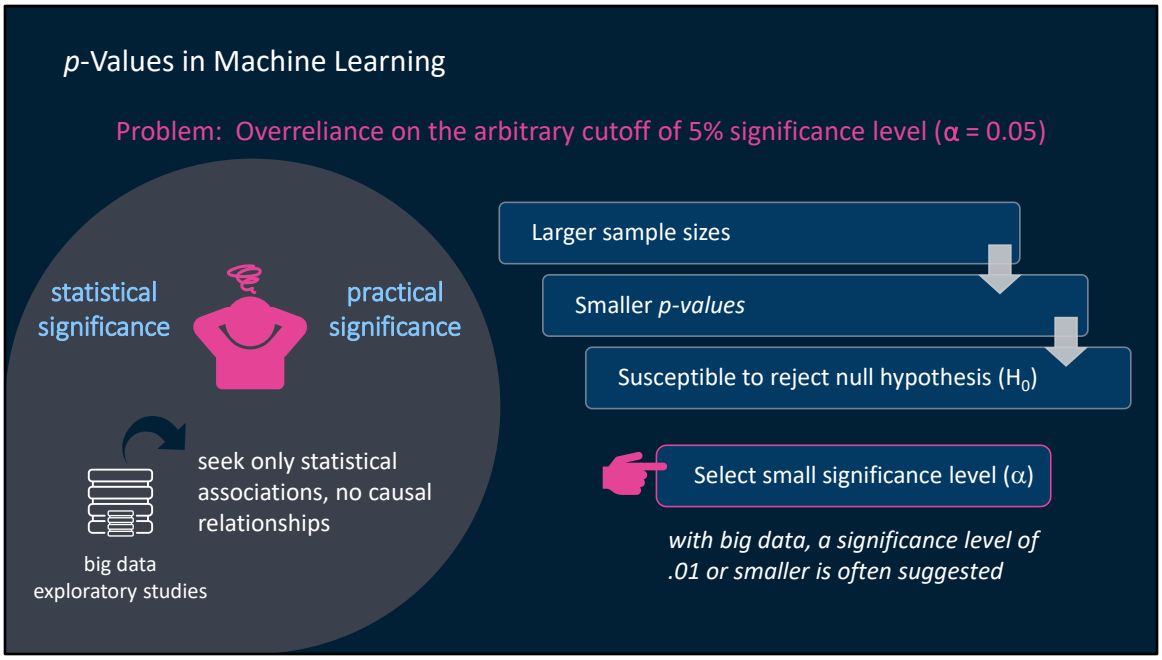
Statistical concepts including *p*-values go hand-in-hand with machine learning.

For example, tree-based algorithms have several splitting criteria based on statistical tests that use *p*-value approach. You need to specify the significance level for the splitting criteria like CHAID, chi-square, and F tests. Regression models like linear, logistic, quantile, and GLM use entry and stay significance level for adding and removing variables in the forward, backward, and stepwise sequential methods.

In a Bayesian network, you specify the significance level (*p*-value) for independence testing using Chi-Square and G-Square. The smaller the value you specify, the fewer variables are selected.

Interactive grouping is one more example where chi-square uses the *p*-value of the Pearson chi-square statistic for the target. So, in machine learning context, *p*-values are predominantly used for feature selection at the data preprocessing level.

In statistics, hypothesis tests are conducted to check whether the inference made for the population holds good or not in an experiment. However, you should not rely too much on *p*-values in machine learning experiments. In machine learning, our primary focus is predictive modeling and not inferential modeling. Machine learning projects generally take care of all the statistics before they are deployed into practice. There are a host of data preprocessing techniques such as feature engineering among others that take care of assumptions in data for machine learning. Incorporating statistical concepts such as hypothesis testing and *p*-values on an already well-set machine learning model might lead to increased complexity in making inferences based on *p*-values. Ultimately, *p*-values show their significance only when there are fewer parameters or variables involved in the experiments or projects.

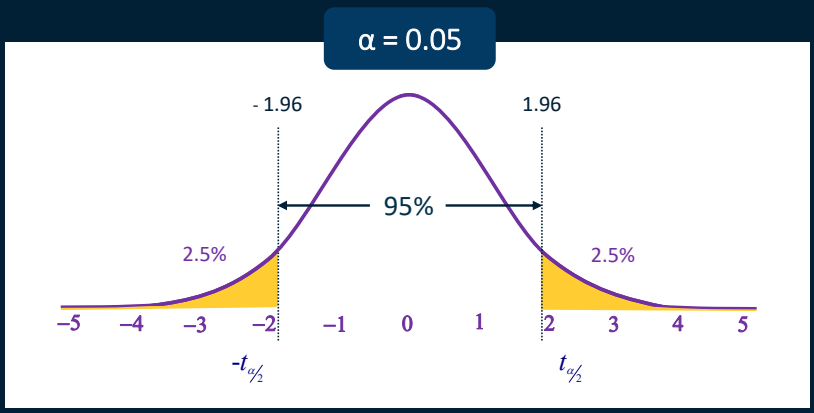


Overreliance on the arbitrary cutoff of 5% significance level can result in a failure to understand the difference between statistical significance and practical significance. This problem is particularly acute in the context of big data exploratory studies, where you seek only statistical associations rather than causal relationships.

Large sample sizes result in smaller *p*-values, in which case you tend to reject the null hypothesis almost every time. So, it's better to select a very small significance level cutoff while using a *p*-value approach in a machine learning world. With big data, a significance level of .01 or smaller is often suggested.

p-Values in Machine Learning

Impact of Level of Significance

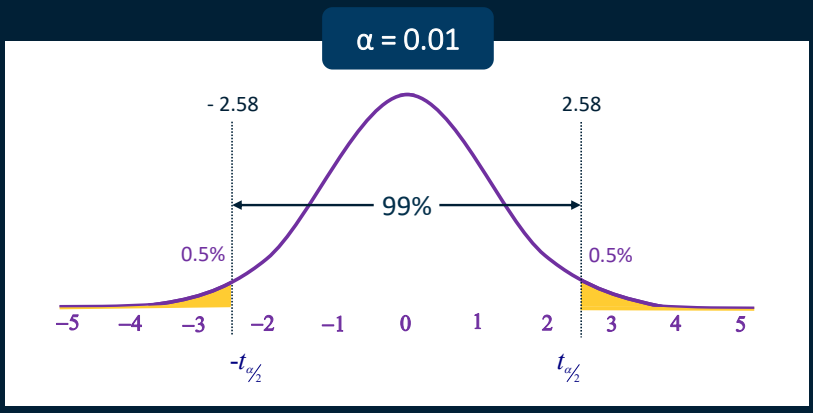


Let's return to the PVA_DONORS example. For the given donation amount, a critical value of +/- 1.96 is shown.

A critical value is a point on the t-distribution that is compared to the test statistic to determine whether to reject the null hypothesis.

p-Values in Machine Learning

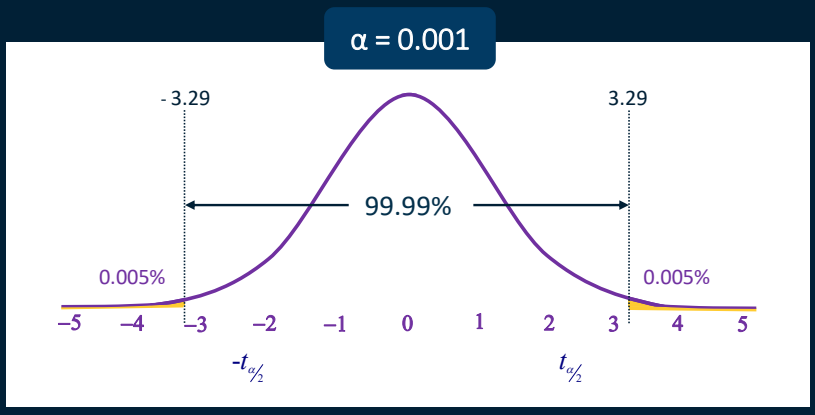
Impact of Level of Significance



As the level of significance decreases, the rejection region also shrinks. The critical value shown here is now +/- 2.58.

p-Values in Machine Learning

Impact of Level of Significance



As the level of significance further decreases, the rejection region shrinks more. The critical value shown here is now +/- 3.29.

Demo: Testing a Hypothesis Using One-Sample t Test

To explore this learning resource in greater detail, please visit the course page on learn.sas.com. There, you'll find comprehensive course materials, tutorials, and updates to support your learning journey.

Question: Confidence Interval

A 95% confidence interval for mean of **Gift Amount Average 36 Months** is ($14.67, $15.08). From this, what can you conclude at alpha=0.05?

a. The true mean of **Gift Amount Average 36 Months** is significantly different from $15.

b. The true mean of **Gift Amount Average 36 Months** is not significantly different from $15.

c. The true average **Gift Amount Average 36 Months** is less than $15.

d. None of the above. You cannot determine statistical significance from confidence intervals.

Answer: Confidence Interval

A 95% confidence interval for mean of **Gift Amount Average 36 Months** is ($14.67, $15.08). From this, what can you conclude at alpha=0.05?

a. The true mean of **Gift Amount Average 36 Months** is significantly different from $15.

b. The true mean of **Gift Amount Average 36 Months** is not significantly different from $15.

c. The true average **Gift Amount Average 36 Months** is less than $15.

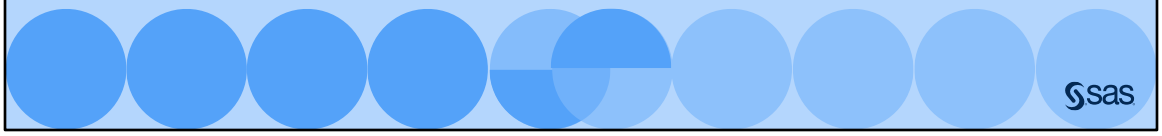
d. None of the above. You cannot determine statistical significance from confidence intervals.

b is the correct answer.

The interval $14.67 to $15.08 contains $15. If your interval contains the hypothesized value, the true mean is not significantly different.

Lesson Summary

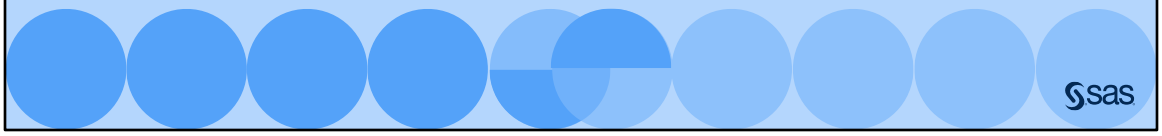
To explore this learning resource in greater detail, please visit the course page on learn.sas.com. There, you'll find comprehensive course materials, tutorials, and updates to support your learning journey.



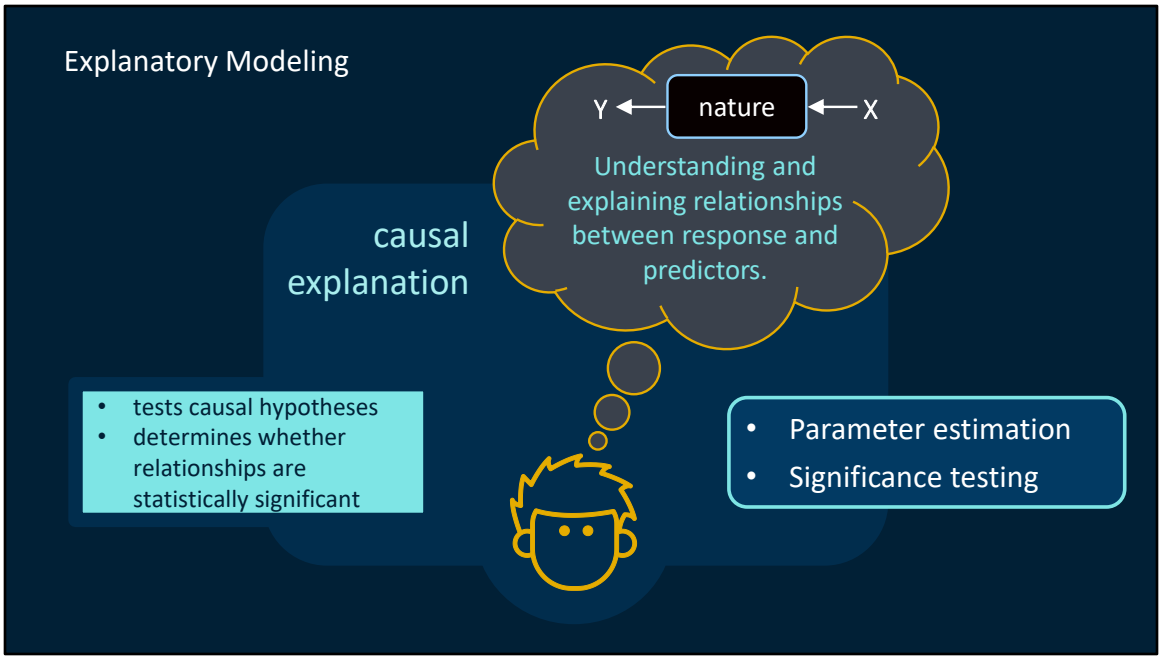
Lesson 3: Explanatory Modeling Using Linear Regression

Lesson Overview

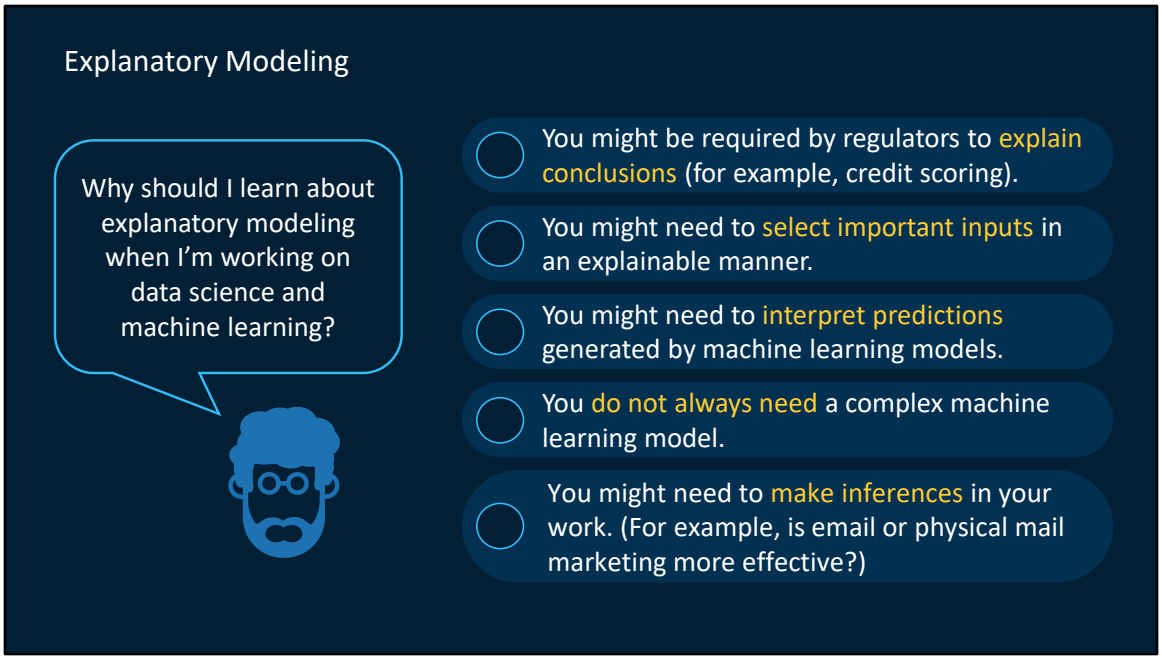
To explore this learning resource in greater detail, please visit the course page on learn.sas.com. There, you'll find comprehensive course materials, tutorials, and updates to support your learning journey.



3.1 Correlation and Simple Linear Regression



Explanatory modeling is focused on understanding and explaining the relationships between a response variable and a set of predictors. In explanatory modeling, statistical models are applied to data in order to test causal hypotheses. In such models, a set of underlying factors that are measured by predictors are assumed to cause an underlying effect, measured by a response variable. A primary goal of explanatory modeling is to estimate parameters and test hypotheses about these parameters. The primary goal is to test the hypotheses, so there is an emphasis on both theoretically meaningful relationships and determining whether each relationship is statistically significant.



The data scientist and predictive modelers might wonder, "Why should I learn about explanatory modeling?" There are many reasons to understand explanatory modeling even if you are primarily a machine learning modeler.

In many areas, regulators require an explanation of the model results, which is greatly facilitated by using tools and methodology of explanatory modeling. For example, in credit scoring, loan decisions need to stand up to regulators' scrutiny, and a logistic regression model is commonly used. Financial institutions might need to explain why a loan was approved or denied.

Explanatory modeling is one of the important approaches for variable selection. Stepwise and LASSO are common variable selection regression methods used in the machine learning world, and details about why particular variables were selected are provided.

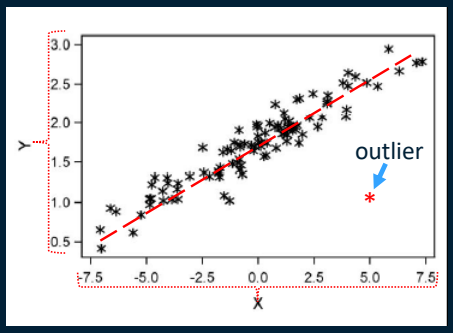
Most of the machine learning algorithms focus on predictive accuracy rather than interpretability of the model. In those situations, explanatory modeling comes to your rescue by running regression models in order to provide prediction explanations.

In many cases (for example, when using smaller sets of data), running a simple explanatory model like regression is sufficient even for prediction purposes, and you do not need to run a complex machine learning model at all. Another reason is that, despite a focus on prediction, there can be situations in which you need to make inferences in your work. Marketers might need to not only predict who is likely to buy a product, but also inform their department whether email marketing or physical mail marketing tends to get better responses.

Explore Your Data before Regression Modeling

scatter plot

continuous



continuous

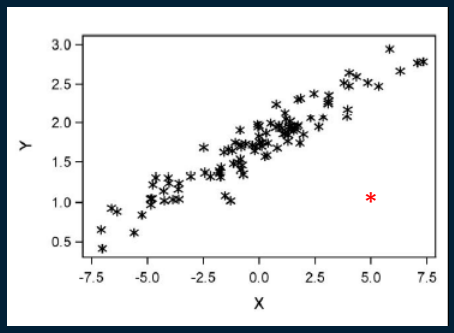
Scatter plots:
- identify **relationships**
- show **range** of values
- help **outliers** stand out

In this lesson, we focus on determining the relationships between continuous variables. A good way to start exploring these relationships is through creating a scatter plot. Scatter plots can help identify the relationship between X and Y. They show the range of X and the range of Y. Often, unusual observations stand out better in these bivariate plots than in graphs and analyses of single variables.

Explore Your Data before Regression Modeling

scatter plot

continuous



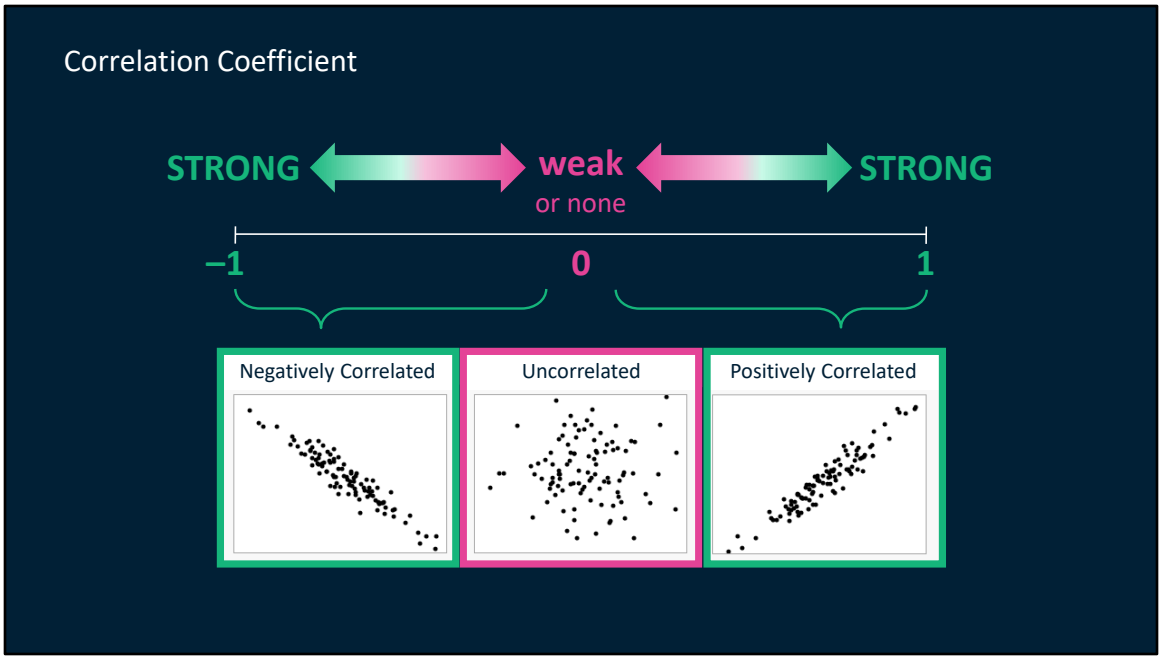
continuous

Linear association:
The general shape of the scatter plot is a **straight line**.

How can we quantify the **strength** of this relationship?

Pearson correlation coefficient

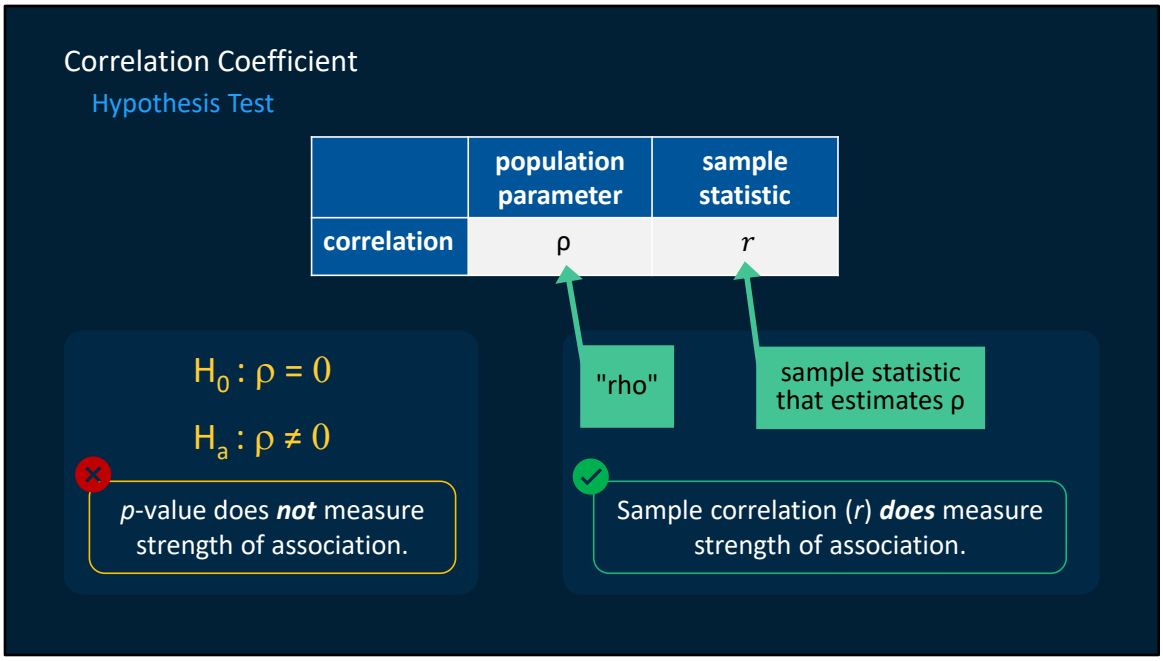
In this picture, the data points line up. We could describe the relationship pictured as a positive linear association between continuous variables. Are we seeing a strong or weak relationship? How can we quantify the strength of this relationship? A statistic that describes the strength of a linear association is the Pearson correlation coefficient.



There are several statistics that measure the correlation between variables. The most commonly used is Pearson's correlation coefficient. This correlation measures the strength of the linear association between continuous variables. It is a unitless measure that can take values between negative 1 and positive 1.

If one variable tends to increase in value as the other variable increases in value, this is a positive correlation. If one variable tends to decrease in value as the other variable increases, the correlation is negative. If no linear relationship exists between the two variables, the correlation is zero. (That is, they are uncorrelated.) Stronger relationships are closer to +1 or -1, and weaker relationships are closer to zero.

Keep in mind that continuous variables can have strong associations that are nonlinear in nature. For this reason, it is important to look at scatter plots when interpreting a correlation coefficient.



The population parameter that represents a correlation is ρ, and the correlation coefficient *r* is the sample statistic that estimates ρ. The null hypothesis for a test of a correlation coefficient is that ρ equals 0, and the alternative hypothesis is that ρ is not equal to 0. Rejecting the null hypothesis means that evidence suggests that the true population correlation is statistically different from 0.

Keep in mind that a *p*-value doesn't measure the strength of the association. The sample correlation, *r*, ***does*** measure the strength of association. To determine the strength of the association between the two variables, you need to focus on *r* to see whether it's meaningfully large. As with many statistics, very large sample sizes can result in small *p*-values. In practice, with a large enough sample size, you would almost always reject the hypothesis that ρ is equal to 0, even if the value of your correlation is small for all practical purposes.

Question: Scatter Plots

Scatter plots are useful to accomplish which of the following?
(Select all that apply)

a. explore the relationships between two variables
b. locate outlying or unusual values
c. identify range of values of the two variables
d. quantify the association

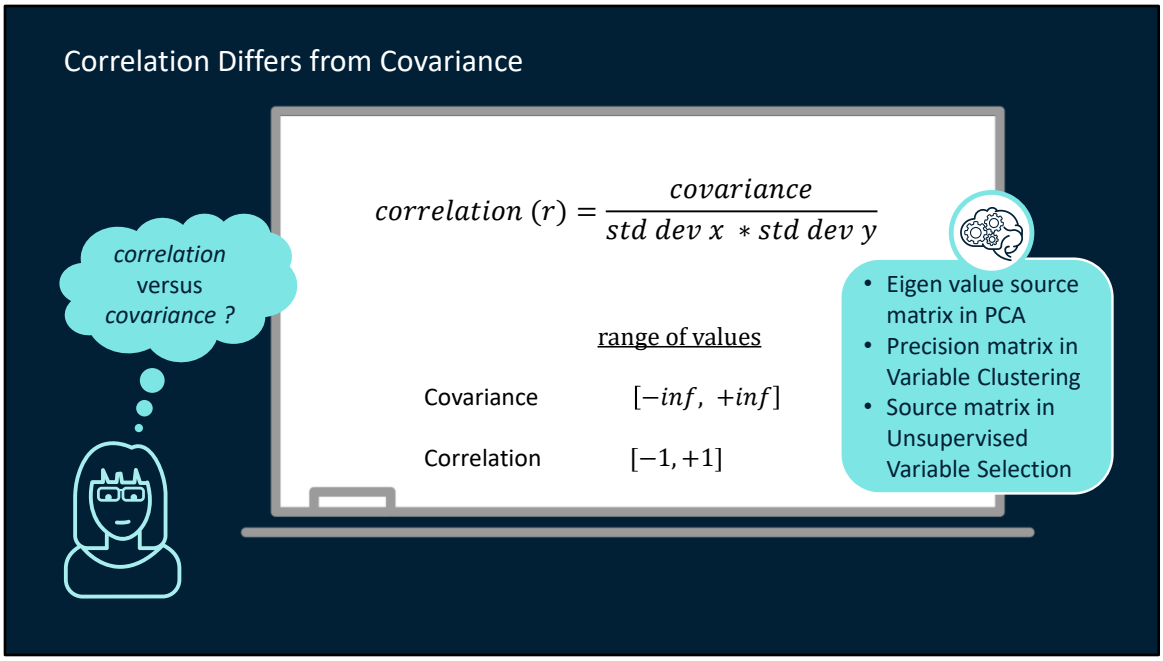
Answer: Scatter Plots

Scatter plots are useful to accomplish which of the following? (Select all that apply)

a. explore the relationships between two variables
b. locate outlying or unusual values
c. identify range of values of the two variables
d. quantify the association

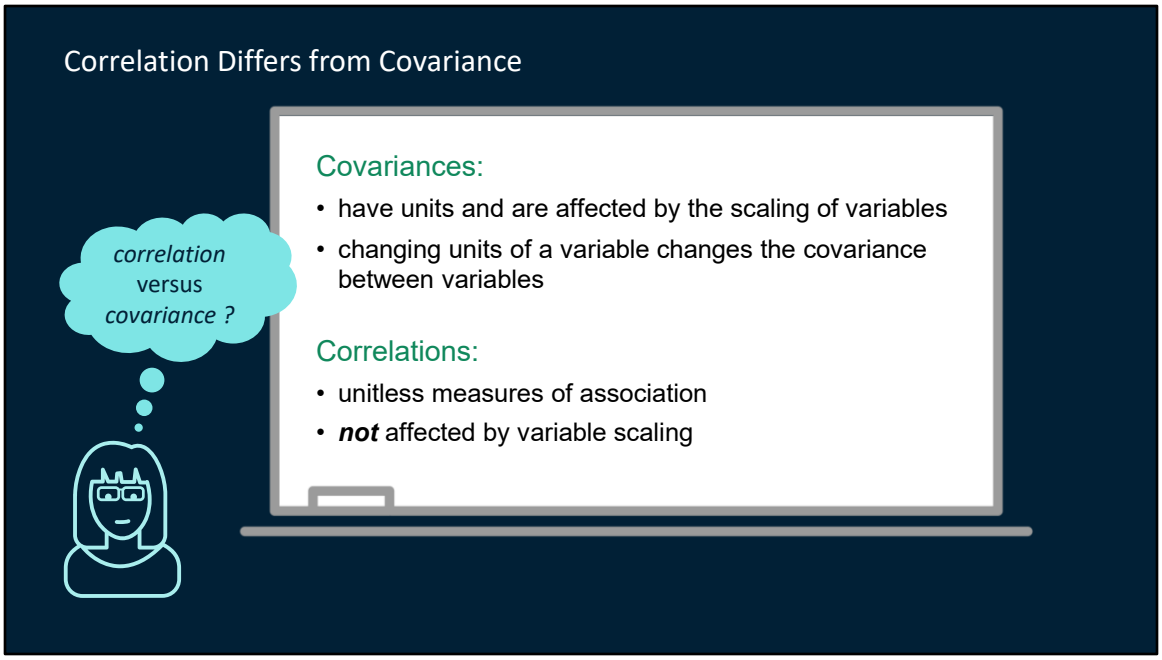
Correct answer: a, b, and c.

Scatter plots are useful to explore the relationships between two continuous variables, locate outlying or unusual values, and identify a basic range of Y and X values. They are also useful for identifying possible trends and communicating data analysis results. However, scatter plots cannot quantify the association between two continuous variables; instead, the Pearson correlation coefficient can be used.

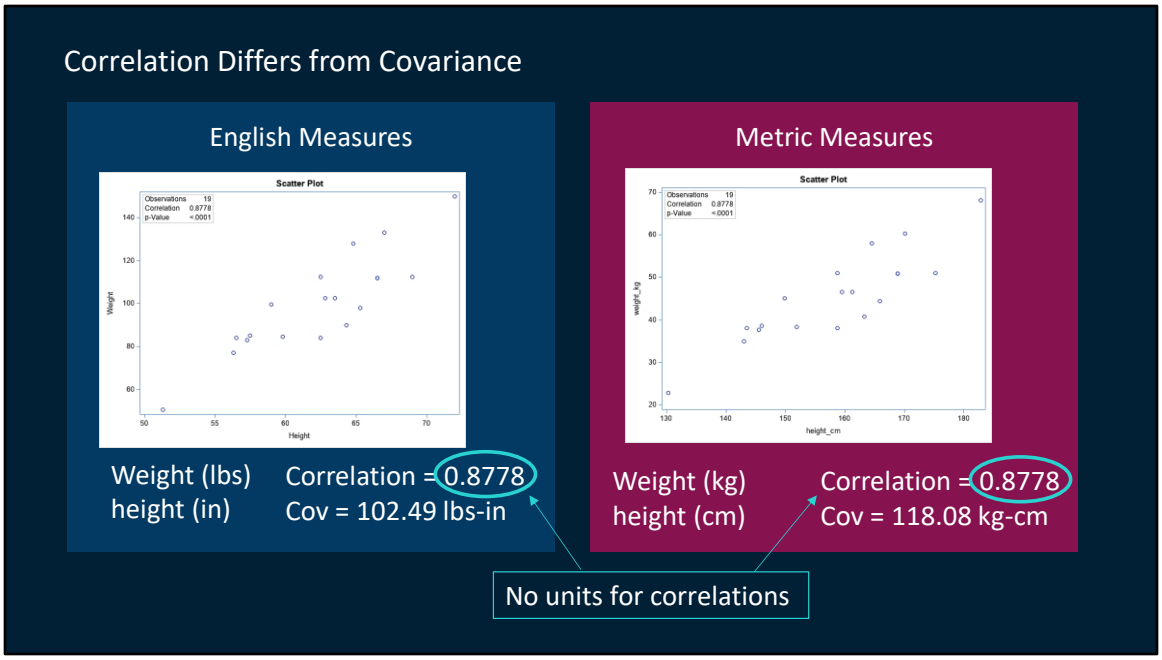


Students sometimes confuse correlation with covariance because both describe an association between continuous variables that can take on positive or negative values. **At times you need to make a choice between a correlation matrix and a covariance matrix.** In machine learning, for example, you need to select Eigen value source matrix in Principal Component Analysis (PCA). Likewise, you are required to specify the maximum number of coordinate descent iterations for estimating the sparse precision covariance matrix in Variable Clustering, and you need to specify the source matrix for running an unsupervised variable selection.

The correlation is actually a standardized version of a covariance. It is the covariance divided by the product of the standard deviations of X and Y. So whereas a covariance can take on values from negative infinity to positive infinity, the correlation will have values between negative and positive 1.



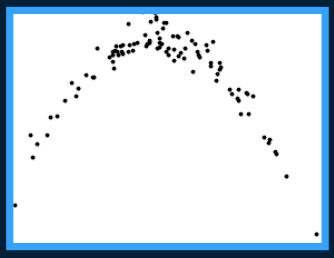
Another difference between covariance and correlation is that covariances have units and are affected by the scaling of variables. Changing the units of a variable changes the covariance between variables. Correlations are unitless measures of association and are not affected by variable scaling.



For example, let's consider this data set of 19 students' heights and weights. Note that the correlation is 0.8778, regardless of whether the measurements are in inches and pounds or centimeters and kilograms. When the scale changes from English measurements to metric, the covariance between height and weight changes from 102 to 118.

The invariance to scaling, and the more intuitive boundaries (-1, +1), make correlations easier to use for communicating relationships.

Pearson Correlation Is Inappropriate for Some Data



Strong curvilinear relationship is not captured by correlation.

Correlation would be near zero, despite strong cyclical relationship.

Presence of one unusual observation would cause a strong correlation.

Correlation measures linear association only.

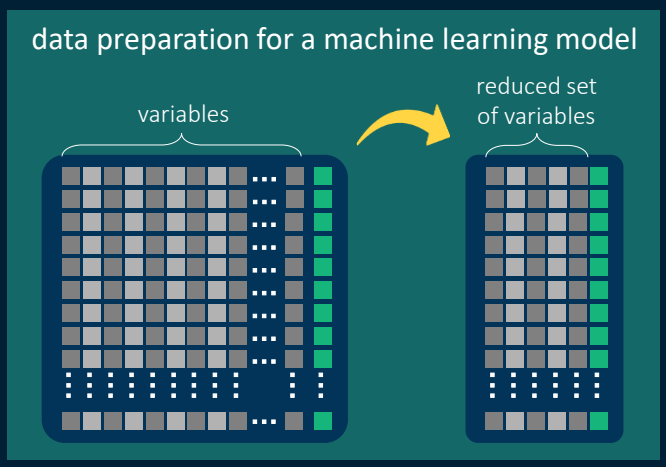
Several relationships between continuous variables are not linear in nature and thus would be inappropriate to quantify with correlation. One such relationship is shown on the left. This picture shows a strong curvilinear relationship between X and Y. Although it is possible to calculate a correlation (which would be near zero), the correlation would be misleading. Without a picture, an analyst cannot be sure that the correlation coefficient is the right tool for the job.

The correlation coefficient for the middle graph would also be near zero, despite the strong cyclical relationship.

A correlation coefficient would also provide a misleading description of the data on the right. Here, there is essentially no relationship between X and Y. But the presence of one unusual observation would cause a strong positive X, Y correlation. Without more data to fill the gap in information between the outlier and the rest of the observations, the analyst cannot conclude that there is a linear relationship at all.

Graphing your data is a crucial step in correlation and, in fact, in all statistical analyses.

Using Correlation for Variable Screening

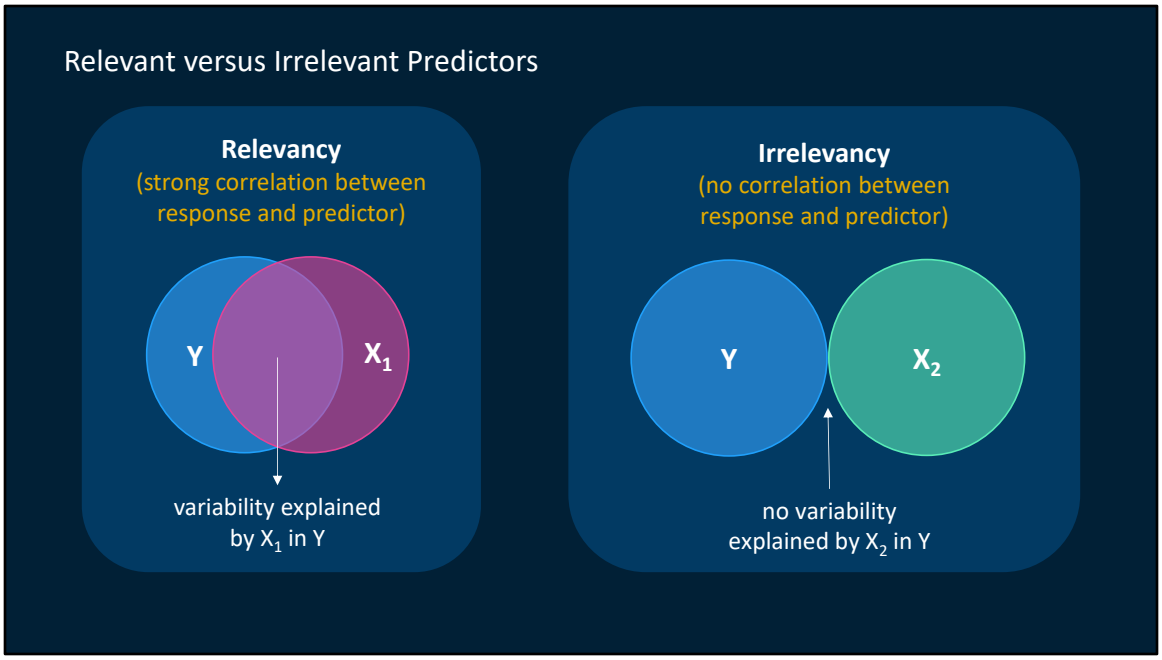


Variable screening identifies ***irrelevant*** predictors.

Variable screening identifies ***redundant*** predictors.

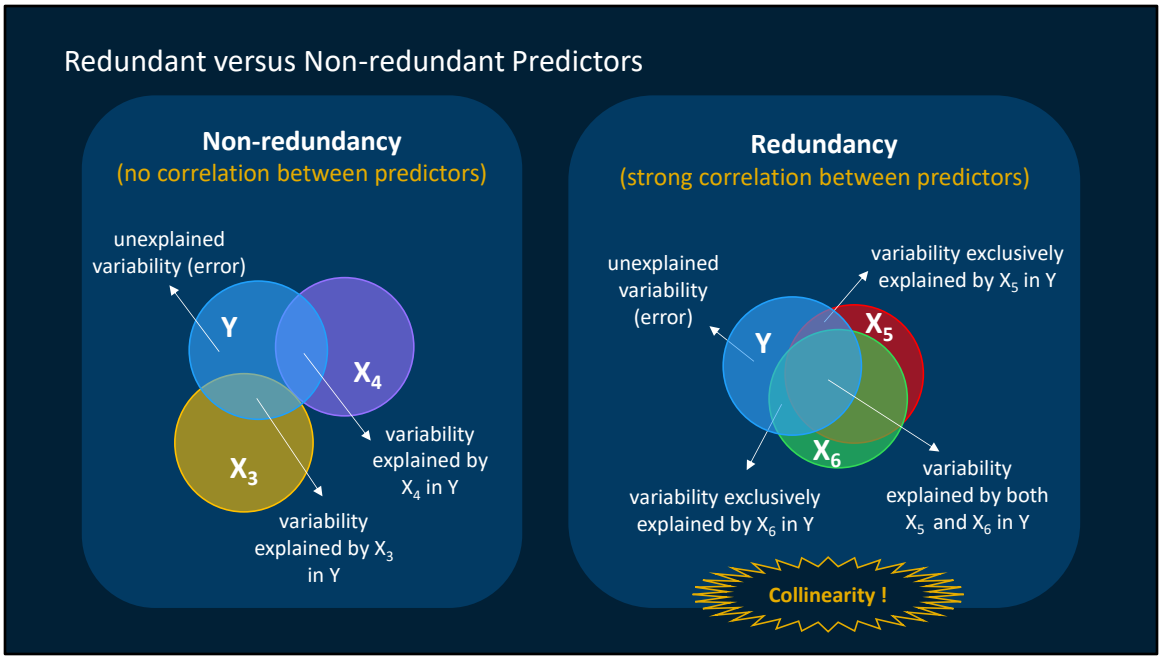
Correlations are not only useful for explanatory modeling. They are often used during data preparation before building a machine learning model as part of variable screening. When the researcher has many predictors, variable reduction can be important, particularly for models that have no built-in variable reduction methods (such as neural networks).

Correlations can be used for variable screening in two general approaches: based on correlations between predictors and the response or based on correlations among predictors. Variable screening based on correlations with the response variable identifies irrelevant predictors. Variable screening based on correlations among predictors identifies redundant predictors. Irrelevant and redundant predictors can be discarded prior to fitting machine learning models. These two approaches are called supervised and unsupervised variable selection.



A predictor is said to be *relevant* when it has a strong correlation with the response variable, like $\mathbf{X_1}$ as shown in the illustration on the left. A relevant predictor will then explain a greater amount of variability in the response variable (**Y**).

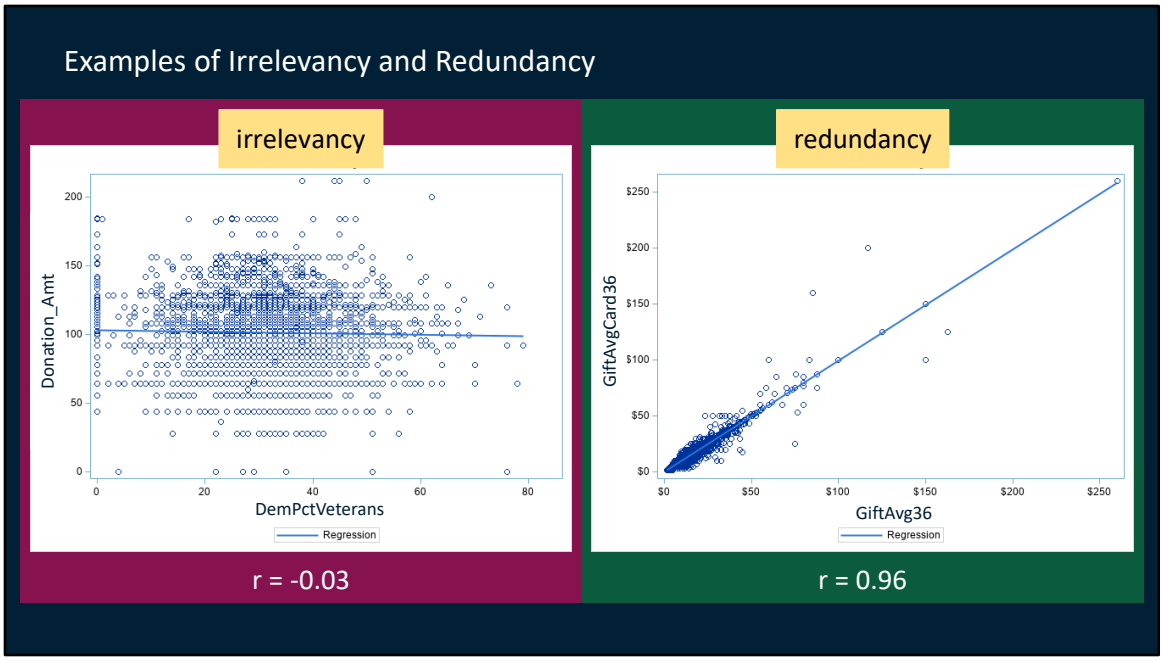
A predictor is said to be *irrelevant* when it does not have a strong correlation with the response variable, like $\mathbf{X_2}$ as shown in the illustration on the right. An irrelevant predictor will not explain, or barely explain, any amount of variability in the response variable (**Y**).



Two predictors are said to be *non-redundant* when they do not have a strong correlation with each other, like $\mathbf{X_3}$ and $\mathbf{X_4}$ as shown in the illustration on the left. Non-redundant or *independent* predictors do not poorly affect the variability explained by them in the response variable. In fact, predictors that are independent are preferred because they do explain mutually exclusive variability in the response variable (**Y**).

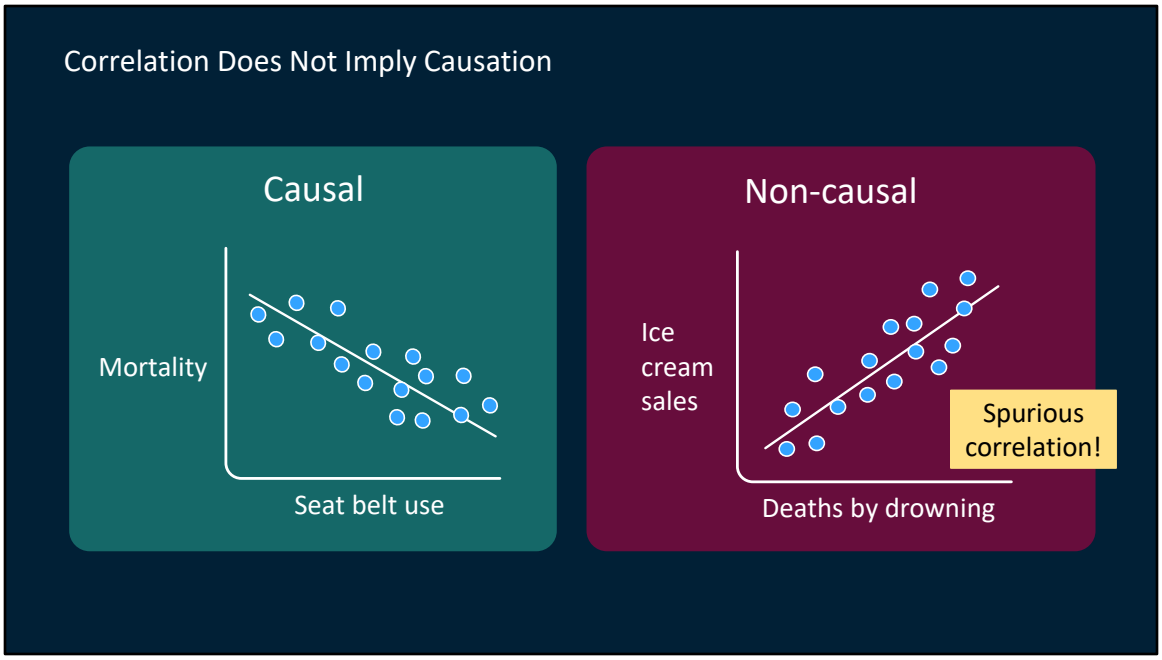
Two predictors are said to be *redundant* when they have a strong correlation with each other, like $\mathbf{X_5}$ and $\mathbf{X_6}$ as shown in the illustration on the right. Redundant or *dependent* predictors might poorly affect the variability explained by them in the response variable. In fact, predictors that are dependent are not preferred because a sizeable amount of variability in the response variable (**Y**) is commonly explained by them. Therefore, a very minimal amount of variability is explained by them exclusively. This phenomenon is commonly known as *collinearity*, We discuss collinearity in detail later in this lesson.

Examples of Irrelevancy and Redundancy



In the first graph shown here, the predictor **DemPctVeterans** has a near zero correlation with the target variable, **Donation_Amt**. This predictor could be considered irrelevant for prediction and discarded from analyses.

In the second graph, **GiftAvg36** and **GiftAvgCard36** are redundant. They essentially convey the same information. With the high correlation between them of 0.96, it is reasonable to use one predictor and discard the other from analyses.



Some correlations are due to a causal relationship between variables. The negative correlation between seat belt use and mortality in vehicular accidents is a causal relationship. But many other associations are not due to causal relationships. For example, a strong correlation exists between ice cream sales and deaths by drowning. This is a spurious correlation! The increase in drowning deaths is obviously caused by more exposure to activities in water, not by ice cream. People tend to both eat more ice cream and swim more in warm weather.

Question: Pearson Correlation

Which of the following statements is true?

a. A negative Pearson correlation indicates a low degree of linear association.
b. The *p*-value for a Pearson correlation can measure the strength of relationship.
c. A random cloud of data implies a negative correlation.
d. The Pearson correlation coefficient is a measure of linear association.

Answer: Pearson Correlation

Which of the following statements is true?

a. A negative Pearson correlation indicates a low degree of linear association.

b. The *p*-value for a Pearson correlation can measure the strength of relationship.

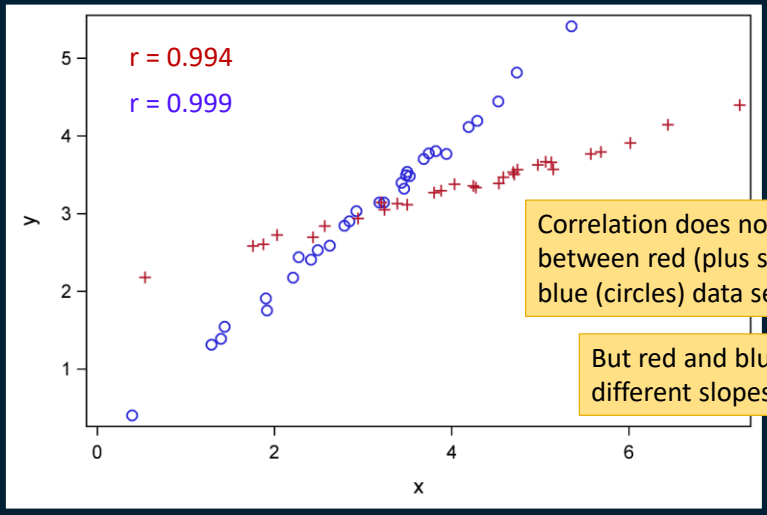
c. A random cloud of data implies a negative correlation.

d. The Pearson correlation coefficient is a measure of linear association.

The correct answer is d.

The Pearson correlation coefficient, r, measures the strength of the linear association between continuous variables and can take values between negative 1 to positive 1. Stronger relationships are closer to +1 or -1, and weaker relationships are closer to zero. A *p*-value doesn't measure the strength of the association, and a random cloud of data indicates that the variables are uncorrelated.

Correlation versus Regression

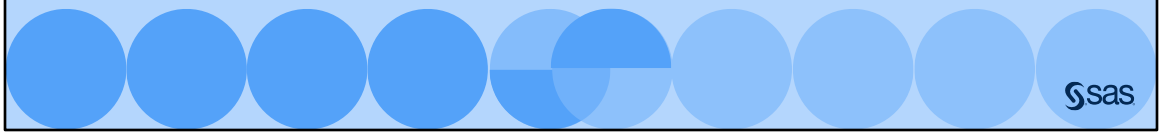


What is the use of regression when we already have correlation to quantify the relationship between two continuous variables? Regression gives us additional information about the relationship.

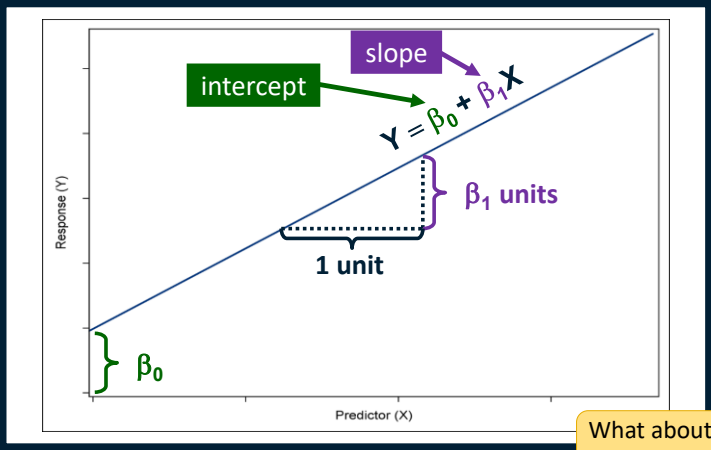
For example, two data sets are pictured here, one in red (plus signs) and the other in blue (circles). The correlation between X and Y for the blue data is virtually identical to the X, Y correlation for the red data. But clearly there are very different linear relationships in these data. Y starts smaller but increases at a faster rate with X for the blue data set. Regression modeling can distinguish between these data where correlation would not. So how do we describe these relationships using linear regression?

History of the Term Regression

To explore this learning resource in greater detail, please visit the course page on learn.sas.com. There, you'll find comprehensive course materials, tutorials, and updates to support your learning journey.



Simple Linear Regression



$$Y = \beta_0 + \beta_1 X + \varepsilon$$

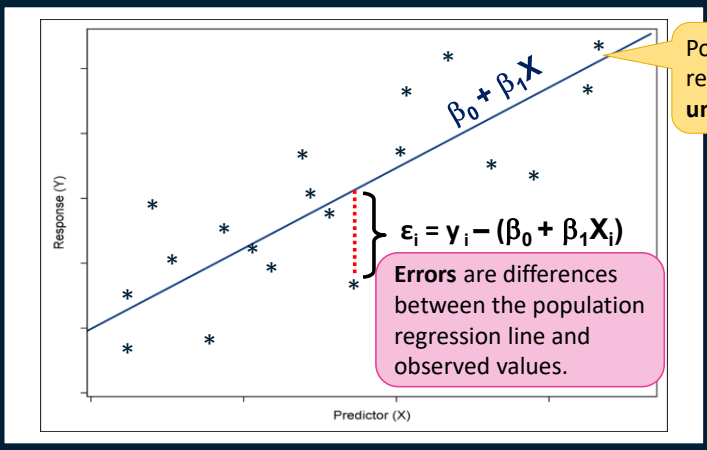
Here's a picture of the simple linear regression model.

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

The term *simple* means that it involves only one predictor. Two or more predictors would make this a multiple regression model.

In the population regression model pictured, the intercept (β_0) describes the average value of Y when X = 0. The regression coefficient (β_1) describes the average change in Y when the predictor X increases by 1 unit. What about the error term, epsilon? To visualize the error term, we need to add observations to this picture.

Simple Linear Regression



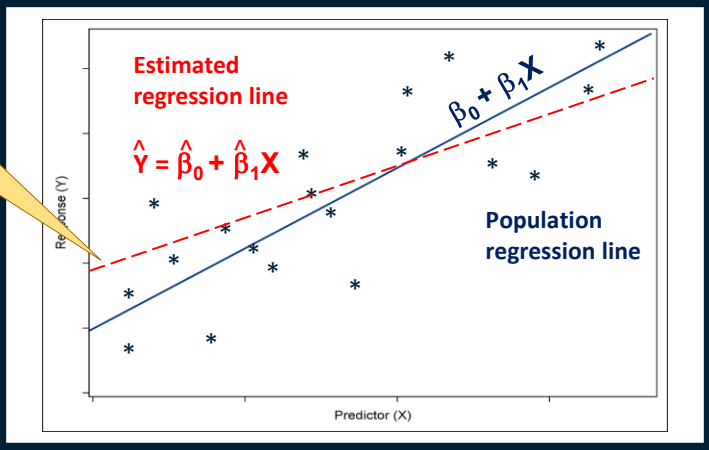
$$Y = \beta_0 + \beta_1 X + \varepsilon$$

Now that we have data points, we can visualize the errors. The errors can be visualized as the vertical distance between a data point and the population regression line. In other words, errors are the difference in Y between an observation and the predicted value for that observation from the population regression equation.

Errors are population parameters. Because we never know the true values of the population parameters β_0, β_1, we never know the true values of the errors. The parameters β_0, β_1, and ε need to be estimated from a sample.

Simple Linear Regression

How do we estimate the regression line?

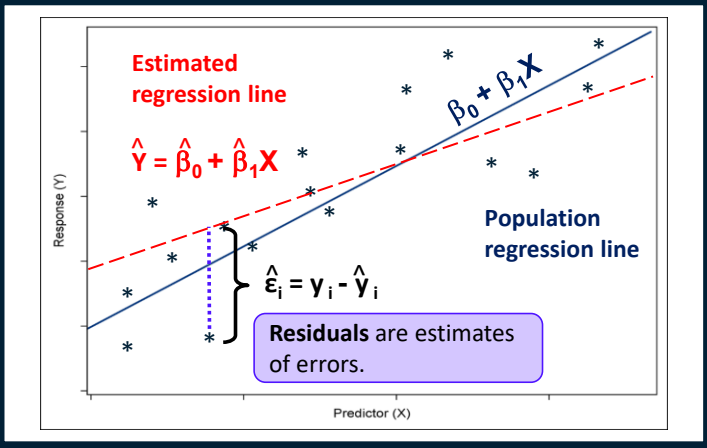


$$Y = \beta_0 + \beta_1 X + \varepsilon$$

When a sample is collected from the population, a regression line can be estimated. If the sample is representative and large enough, and important predictors are measured, the estimated regression equation will likely be a good approximation to the population regression equation. How is this "best fit" regression line estimated?

Simple Linear Regression

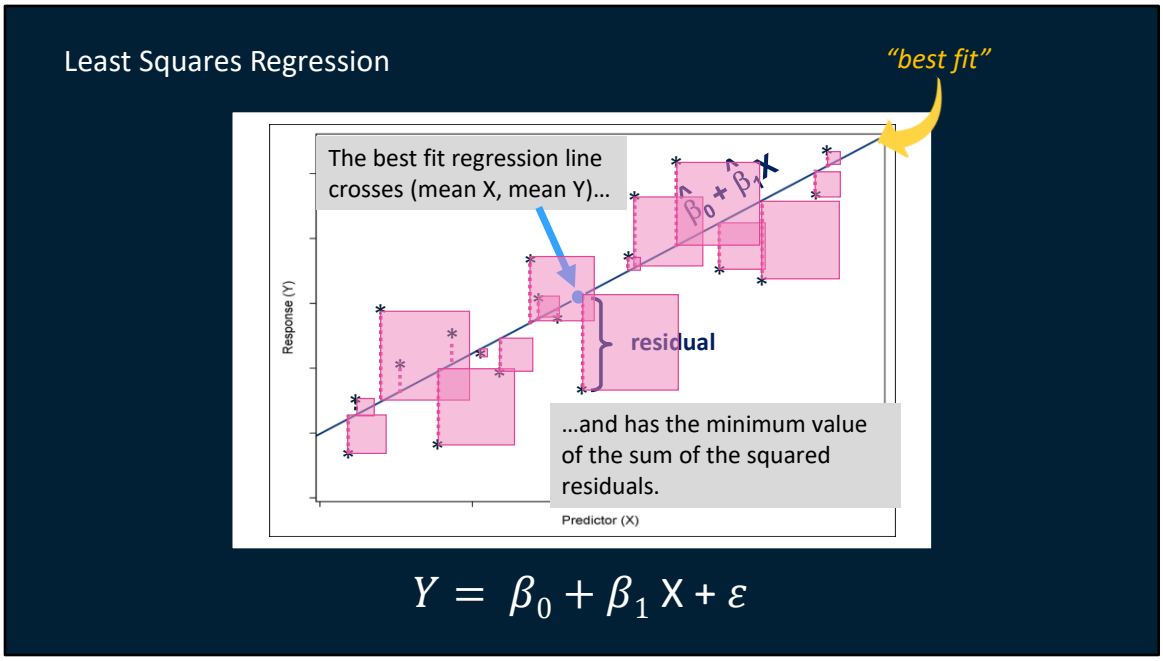
Estimating and assessing regression involves residuals.



$$Y = \beta_0 + \beta_1 X + \varepsilon$$

A commonly used approach to estimating the regression line is to use ordinary least squares. The process involves the residuals, the difference between predicted values and actual values of the response variable Y. Residuals are estimates of the population errors and thus can be observed, whereas population errors are never observed directly. Many of the assumptions for linear regression involve the error term. When we check assumptions of regression models that involve the error term, we will use the residuals as a proxy.

It's worth noting that statistical literature often distinguishes between errors and residuals, but machine learning literature often uses the term *error* to mean what a statistician would call a *residual*: the difference between the observed values and predicted values from a model fitted to a sample.

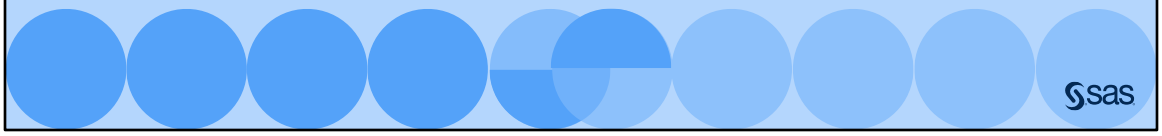


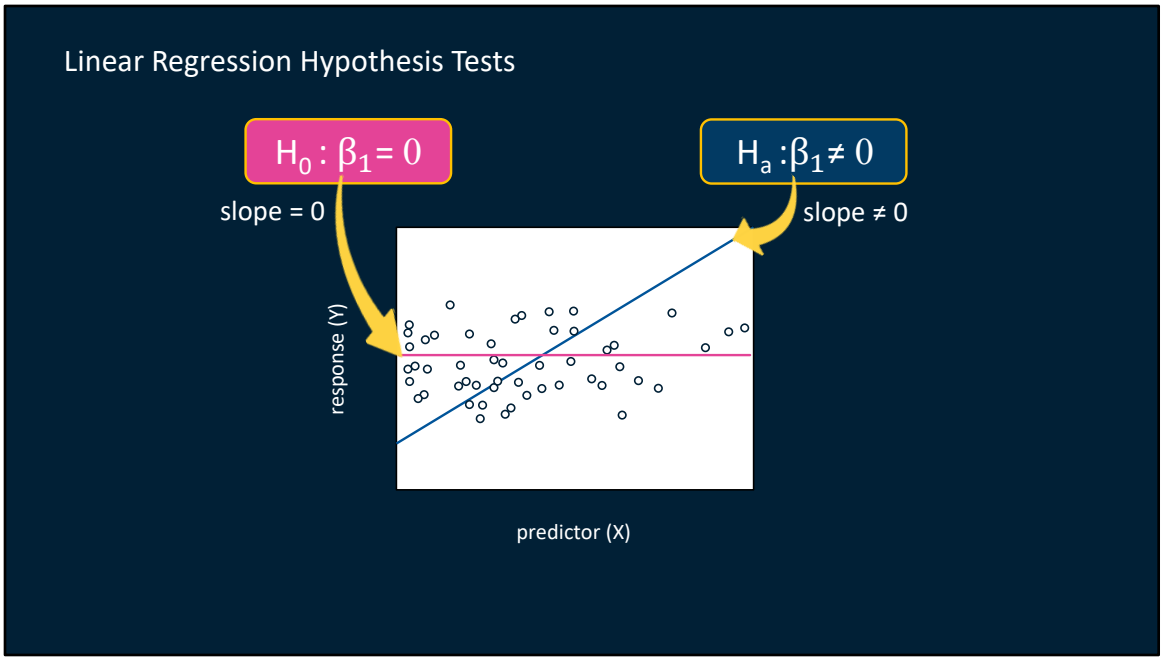
In order to estimate an unknown relationship between continuous variables in a population, you collect a sample and attempt to estimate the relationship using ordinary least squares. How does ordinary least squares produce what is commonly called a "best fit" regression line? The process involves the residuals.

The best fit regression line is the line that crosses the point (mean X, mean Y) that has the minimum value of the sums of squared residuals. Again, the residuals are the differences between the observed values and the values predicted from a model. They can be visualized as the vertical distance from the regression line to each data point. These residuals are squared and then summed to produce the sums of squared residuals. This is where the term least squares regression comes from.

OLS Regression Parameter Estimates

To explore this learning resource in greater detail, please visit the course page on learn.sas.com. There, you'll find comprehensive course materials, tutorials, and updates to support your learning journey.





The hypothesis test for a simple linear regression model is a test of whether the slope (β_1) is significantly different from zero. A zero slope for a regression corresponds to a horizontal line where Y equals the mean value of Y. A flat regression line indicates that the mean value of Y does not depend on X. X gives us no information about Y.

If the null hypothesis of $\beta_1 = 0$ is rejected, then the best fit regression line has a nonzero slope, and the best estimate of the mean of Y depends on the value of X.

This statistical test is valid only when the assumptions of regression are met. These assumptions are described later this lesson.

As with correlation, it is important to distinguish between statistical significance (the *p*-value for the test) and the effect size (the magnitude of the parameter estimate β_1). A slope that is significantly different from zero might still be too small to be meaningful to the researcher.

Question: Simple Linear Regression

In a simple linear regression, if the parameter estimate (slope) of X1 is 0, then the best guess (predicted value) of Y when X1 = 15 is which of the following?

a. 15
b. the mean of Y
c. a random number
d. the mean of X1
e. 0

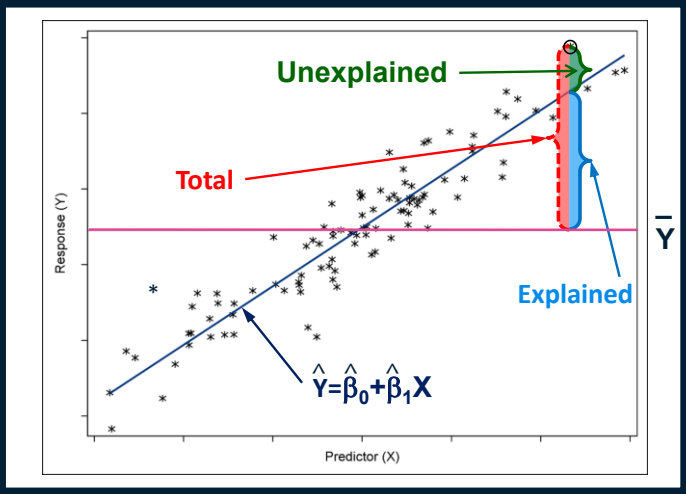
Answer: Simple Linear Regression

In a simple linear regression, if the parameter estimate (slope) of X1 is 0, then the best guess (predicted value) of Y when X1 = 15 is which of the following?

a. 15
b. the mean of Y
c. a random number
d. the mean of X1
e. 0

The correct answer is b.
In the simple linear regression model $Y = \beta0 + \beta1X + \varepsilon$, the intercept (β0) describes the average value of Y when X = 0. The regression coefficient (β1) describes the average change in Y when the predictor X increases by 1 unit. However, if β1, the parameter estimate (slope) of X1 is zero then the predicted value of Y takes the value equal to the intercept (β0), which is the mean of Y.

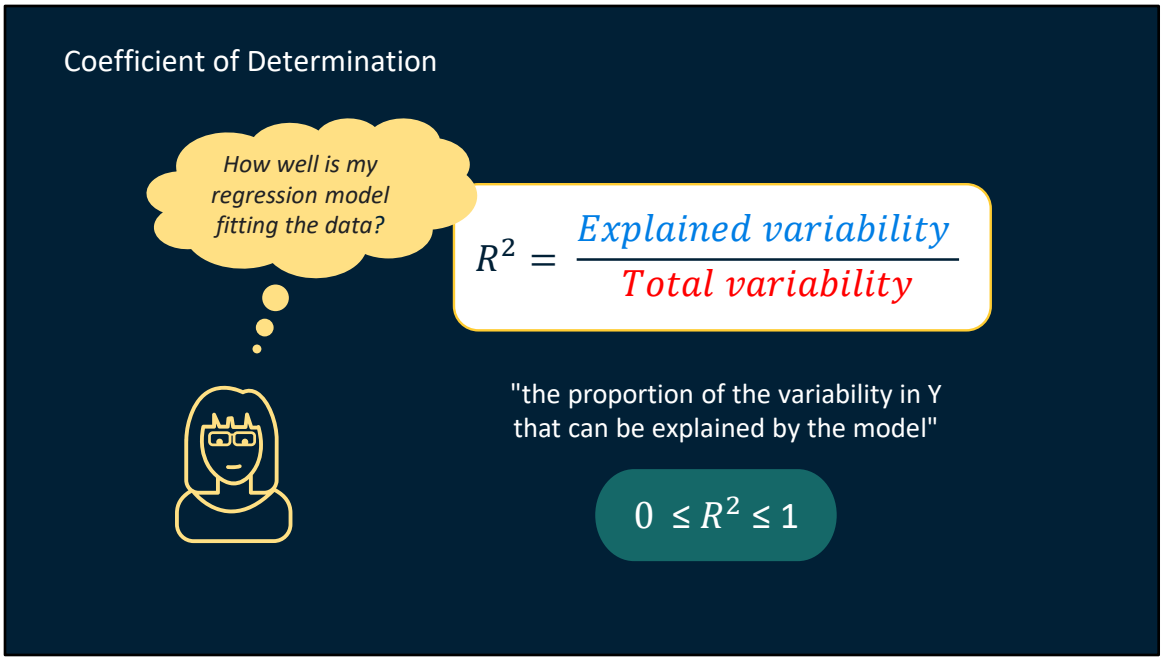
Explained versus Unexplained Variability



What proportion of the total variability is explained by the model?

In regression, the variability of the response variable Y can be partitioned into variability that can be explained by the model and unexplained variability. The explained and unexplained variability sum to the total variability in Y. The explained variability relates to the difference between the regression line and the overall mean value of the response. The unexplained variability relates to the difference between the observed values and the regression line.

Because we can partition variability into explained and unexplained, we often want to know what proportion of the total variability can be explained by the regression line.

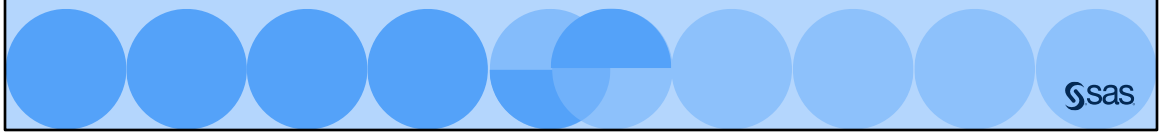


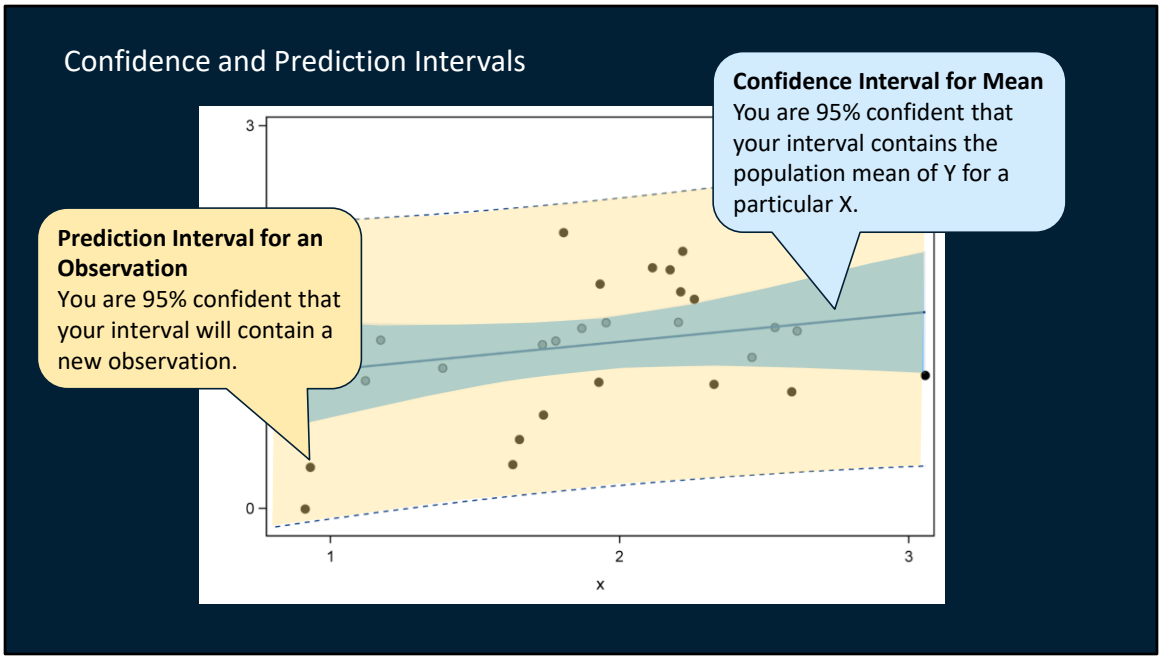
The *coefficient of determination*, R-square, is a measure of the proportion of variability in the response variable explained by predictor variables in the model.

The value of R-square can be between 0 and 1. Regression models with R-square close to 0 do not explain much variability in the data. Regression models with R-square close to 1 explain a relatively large proportion of variability in the data.

More about Least Square Regression

To explore this learning resource in greater detail, please visit the course page on learn.sas.com. There, you'll find comprehensive course materials, tutorials, and updates to support your learning journey.





Confidence intervals and prediction intervals can be constructed around the estimated regression line. Often these are 95% confidence limits or 95% prediction limits, but other values can also be used. The confidence limits for the mean enables you to say with a given level of confidence where you expect the true mean for a particular value of X to lie.

Sometime the mean Y is of less interest than the next observation of Y. In this case, prediction limits can be useful. Prediction limits describe an interval in which the next observation is likely to be found.

Question: Relationships between Variables

Which of the following is a true statement?

a. A correlation near -1 indicates a lack of association.
b. Scatter plots do not help interpreting correlation coefficients.
c. The coefficient of determination (R^2) can take values between -1 and +1.
d. Errors are population parameters, and residuals are estimates of the errors.

Answer: Relationships between Variables

Which of the following is a true statement?

a. A correlation near -1 indicates a lack of association.

b. Scatter plots do not help interpreting correlation coefficients.

c. The coefficient of determination (R^2) can take values between -1 and +1.

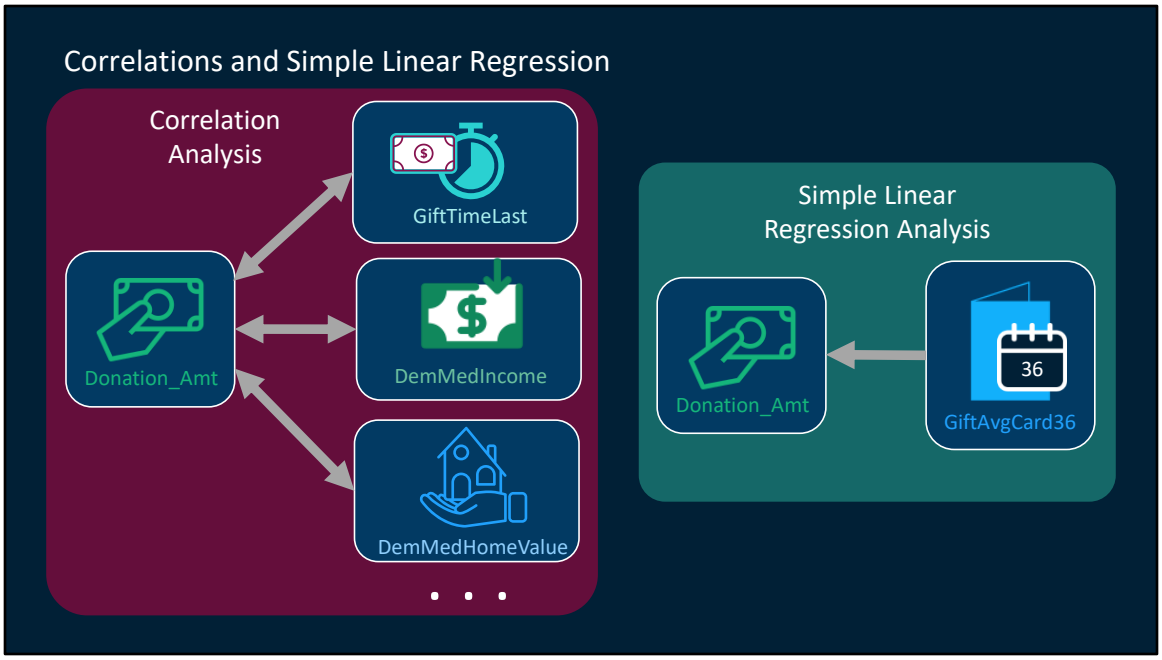
d. Errors are population parameters, and residuals are estimates of the errors.

The correct answer is d.

a. Is incorrect because a correlation near -1 indicates a strong negative association, not a lack of association.

b. Is incorrect because scatter plots are crucial for interpreting correlation coefficients. They can be used to detect nonlinear associations, which can be inappropriate to describe with a correlation coefficient.

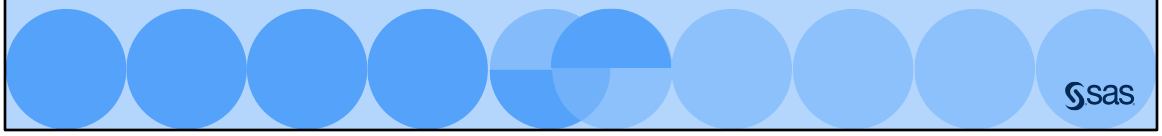
c. Is incorrect because the coefficient of determination (R-square) can take values between zero and 1.



Now we will explore relationships between predictors and the response through correlation analysis. We will estimate and test the significance of correlations between the continuous target **Donation_Amt** and each of the continuous predictors, such as **GiftTimeLast**, **DemMedIncome**, and **DemMedHomeValue**. After finding the predictor with the strongest correlation, we will regress **Donation_Amt** on **GiftCardAvg36**.

Demo: Correlation and Linear Regression

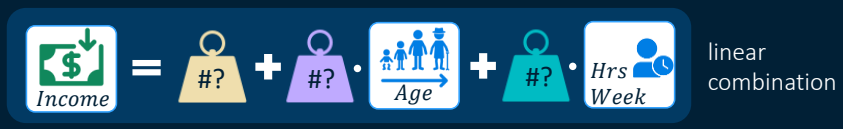
To explore this learning resource in greater detail, please visit the course page on learn.sas.com. There, you'll find comprehensive course materials, tutorials, and updates to support your learning journey.



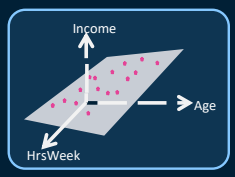
3.2 Multiple Regression and Model Selection

Multiple Regression

Recall this example:



Name	Age	Gender	HrsWeek	Income
John	24	M	60	241.89
Smith	18	M	30	162.31
Emma	31	F	45	220.04
…	…	…	…	…



In simple linear regression, you model the relationship between two variables with a line.

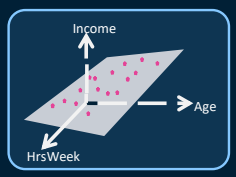
Recall an example we considered in lesson 1, a simple data set containing information on age, gender, working hours per week, and income. Where you have two predictors, the relationship between predictors and the response can be modeled with a plane. Here we see a multiple regression model with the dependent variable **Income** regressed onto the predictors **Age** and **HrsWeek**. The weights or parameters are estimated for the model using ordinary least squares.

Multiple Regression

Recall this example:

$$\widehat{Income} = \hat{\beta}_0 + \hat{\beta}_1 \, Age + \hat{\beta}_2 \, HrsWeek$$

Name	Age	Gender	HrsWeek	Income
John	24	M	60	241.89
Smith	18	M	30	162.31
Emma	31	F	45	220.04
…	…	…	…	…



$\hat{\beta}_1$ = effect of **Age** on **Income** while holding **HrsWeek** constant

$\hat{\beta}_2$ = effect of **HrsWeek** on **Income** while holding **Age** constant

These estimated weights are β_0 (the intercept), β_1 (the slope for **Age**), and β_2 (the slope for **HrsWeek**).

The multiple regression model is very similar to a simple linear regression model in that it has terms for an intercept and slope. But when there are two or more predictors, each has its own slope, and the meaning of the slope is different. Now, a slope describes the effect of a unit change in a predictor on the mean of the response ***while holding other predictors constant***.

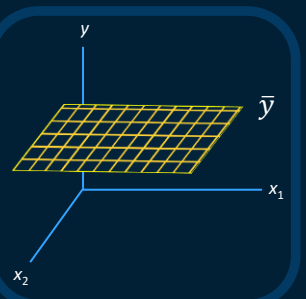
So β_1 indicates the average change in **Income** for a unit change in **Age**, while holding **HrsWeek** constant. And β_2 indicates the average change in **Income** for a unit change in **HrsWeek** while holding **Age** constant.

The regression coefficient for each predictor is adjusted for other variables in the model. Because of this, the estimated coefficients can change when other predictors are added or removed from the analysis.

Multiple Regression Hypothesis Test

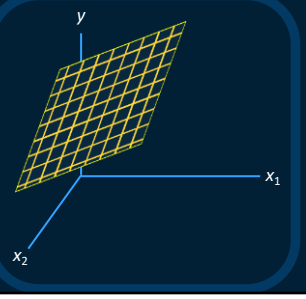
Null hypothesis: All the coefficients for predictors equal zero.

$$H_0: \beta_1 = \beta_2 = \cdots = \beta_k = 0$$



Alternative hypothesis: There is at least one nonzero coefficient.

$$H_a: \text{not}(\beta_1 = \beta_2 = \cdots = \beta_k = 0)$$



The hypothesis test for a multiple regression model is a test of whether the coefficients (that is the slopes or betas) are all equal to zero. A zero slope for each coefficient corresponds to a horizontal regression plane where the intercept equals the mean value of Y. A flat regression surface indicates that the mean value of Y does not depend on the X's. The predictors give us no information about Y.

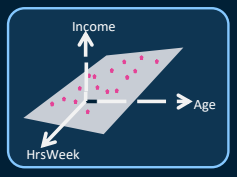
If the null hypothesis of all βs = 0 is rejected, then the best fit regression plane has at least one nonzero slope, and the best estimate of Y depends on the values of the X's.

Categorical Predictors in Regression

$$\widehat{Income} = \hat{\beta}_0 + \hat{\beta}_1 \cdot Age + \hat{\beta}_2 \cdot HrsWeek$$

Name	Age	Gender	HrsWeek	Income
John	24	M	60	241.89
Smith	18	M	30	162.31
Emma	31	F	45	220.04
…	…		…	…

How are these converted to a numeric variable?

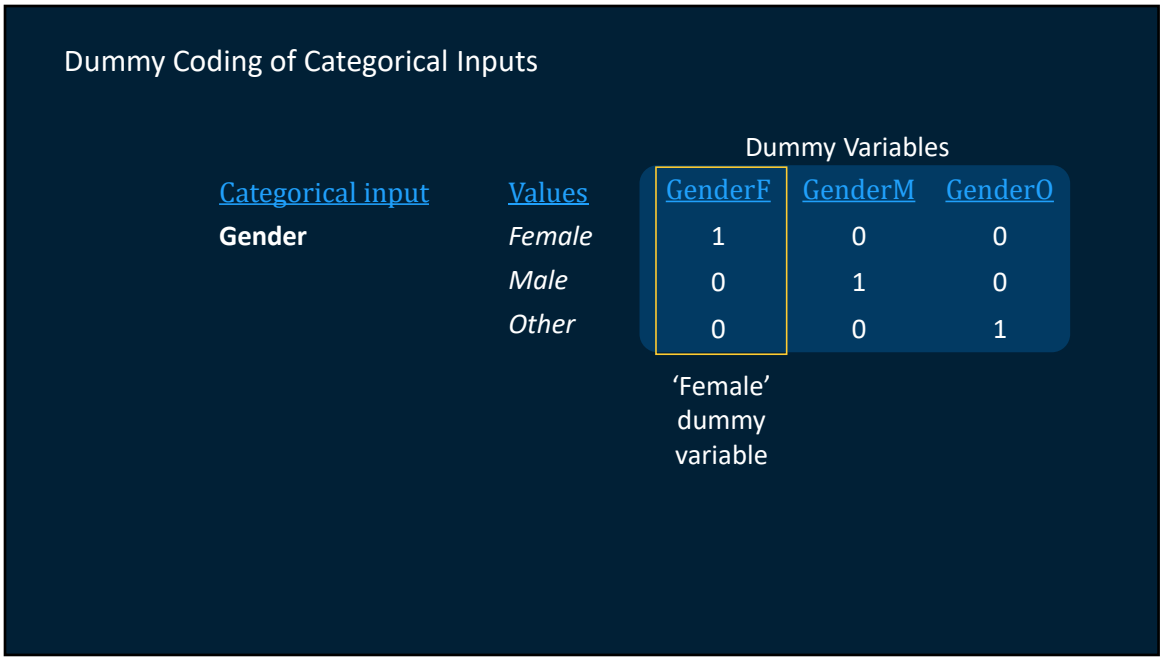


a regression model

The term *regression* originally referred only to modeling continuous predictors, but categorical predictors can be used as well. These predictors need to be converted into numeric variables that represent the different levels.

SAS does this conversion automatically for us. How does this conversion happen?

Dummy Coding of Categorical Inputs



Dummy variables (sometimes called *design variables* or *indicator variables*) are constructed numeric variables that represent the levels of a categorical predictor. These can be created in variety of ways, and one of the most common ways or parameterizations is shown here.

Let's suppose that the categorical predictor gender has three levels: *Male*, *Female*, and *Other*. In order to differentiate these levels or groups using numeric values, three dummy variables can be constructed. The first can be thought of as the Female dummy variable, because females are assigned a value of 1 and other individuals are assigned a value of 0.

Dummy Coding of Categorical Inputs

Categorical input	Values	Dummy Variables: GenderF	GenderM	GenderO
Gender	*Female*	1	0	0
	Male	0	1	0
	Other	0	0	1

'Male' dummy variable

In a similar manner, a dummy variable for males takes the value of 1 for people in the Male category and 0 for everyone else.

Dummy Coding of Categorical Inputs

Categorical input	Values	Dummy Variables: GenderF	GenderM	GenderO
Gender	*Female*	1	0	0
	Male	0	1	0
	Other	0	0	1

'Other' dummy variable

Finally, members of the Other category are assigned a value of 1 for the Other dummy variable, and everyone else gets a 0.

With three levels for gender, three dummy variables are created.

Multiple Regression with Categorical Predictors

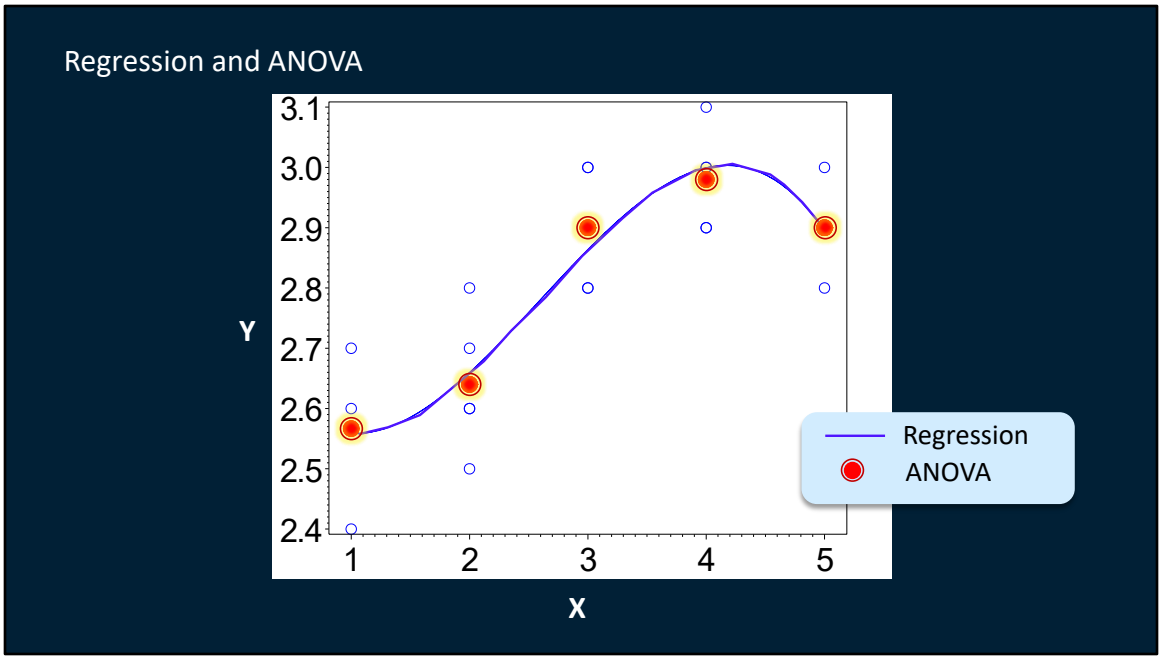
$$\widehat{Income} = \hat{\beta}_0 + \hat{\beta}_1 \cdot Age + \hat{\beta}_2 \cdot HrsWeek + \hat{\beta}_3 \cdot GenderF + \hat{\beta}_4 \cdot GenderM + \hat{\beta}_5 \cdot GenderO$$

three parameters for **Gender**

Name	Age	Gender	HrsWeek	Income
John	24	M	60	241.89
Smith	18	M	30	162.31
Emma	31	F	45	220.04
…	…	…	…	…

Regression with only categorical predictors is called analysis of variance (ANOVA).

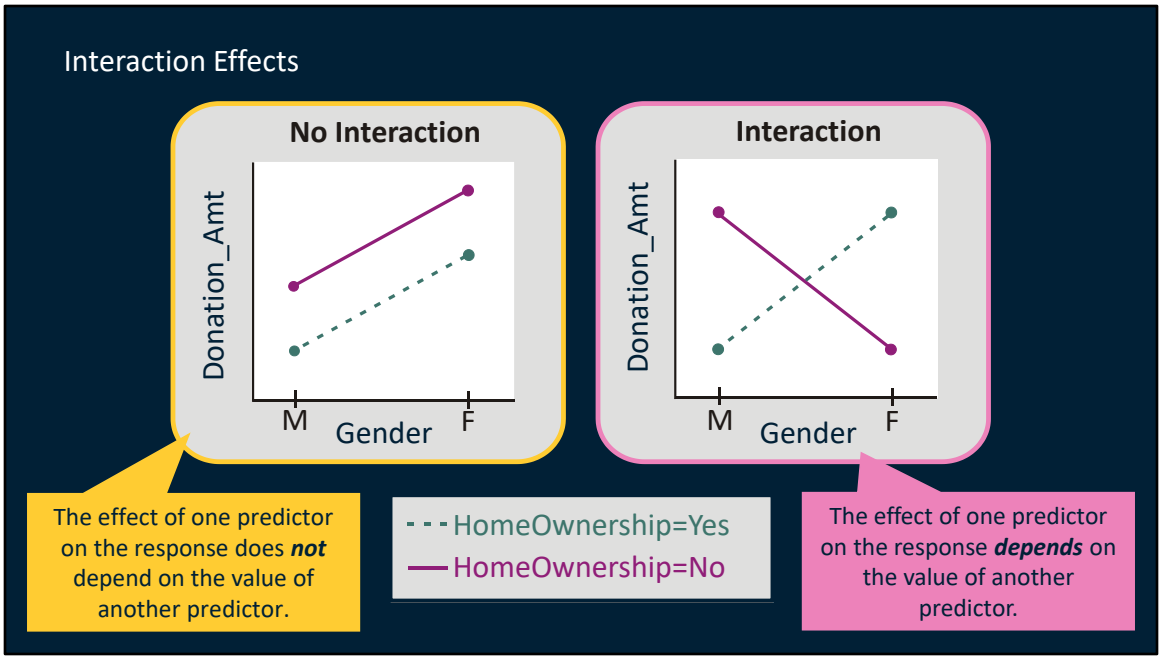
It is these three dummy variables and not the original variable **Gender** that would be used directly in a multiple regression model. Regression with only categorical predictors is called analysis of variance (or ANOVA).



There is a close proximity between regression and ANOVA (analysis of variance). In fact, both are part of the *general linear model* that uses the ordinary least squares method to fit a continuous response variable.

When a linear regression model is fit to the data, you assume that the mean of the response (**Y**) at each value of the predictor (**X**) follows a mathematical relationship. This relationship is being estimated and presented as a line. You are interested in estimating the nature of the relationship (the slopes), whether this relationship is significant (slopes equal to zero), and to what extent the variations in the data are explained by your model. Here a curve is used on the graph to represent a more complex relationship that reveals the notable difference between regression and ANOVA.

In ANOVA, the mean of the response (**Y**) at each value of the predictor (**X**) is also computed. But there is no mathematical relationship being defined between these average values and the values of the predictor (**X**). As a matter of fact, the predictor might not be on an interval scale. The predictors can be numeric or character variables, and they can be on a nominal or an ordinal scale. When this is the case, a regression model might not be appropriate because the mathematical relationship between the mean of the response and the predictors might not be defined. The predictors group your data into different groups, and the question of interest is whether the group means are equal to each other.



When there are multiple predictors in a regression model, there is the potential for predictors to interact. When interactions exist, modeling them can greatly increase the explanatory power and predictive ability of a model.

An interaction means that the effect of a predictor on the response variable depends on the value of another predictor. An interaction occurs when changing the level of one factor results in changing the difference between levels of the other factor.

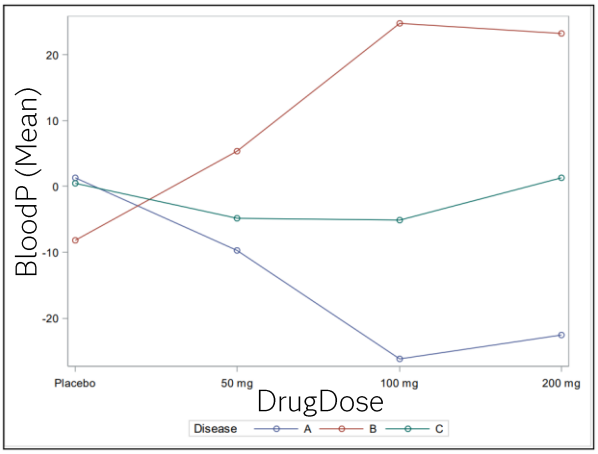
The plots displayed here are called *means plots*. The average **Donation_Amt** values over different levels of **Gender** were plotted and then connected for **HomeOwnership** *Yes* and *No*.

In the plot on the left, the two types of homeownership show the same change across different levels of **Gender**. However, in the plot on the right, as the **Gender** level changes, the average donation amount ***increases*** for donors who own a home and ***decreases*** for donors who do ***not*** own a home. This indicates an interaction between the variables **Gender** and **HomeOwnership**.

Question: Interaction

Examine this interaction plot. Which of the following statements is correct?

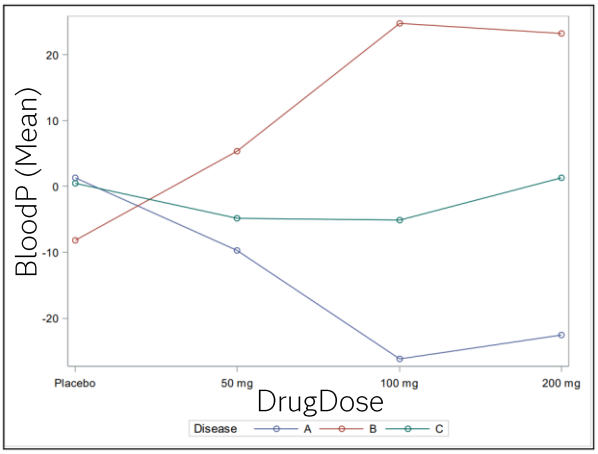
a. Drug dose does not affect blood pressure.
b. Disease does not affect blood pressure.
c. There is no interaction between drug dose and disease.
d. An interaction exists between drug dose and disease.



Answer: Interaction

Examine this interaction plot. Which of the following statements is correct?

a. Drug dose does not affect blood pressure.
b. Disease does not affect blood pressure.
c. There is no interaction between drug dose and disease.
d. An interaction exists between drug dose and disease.

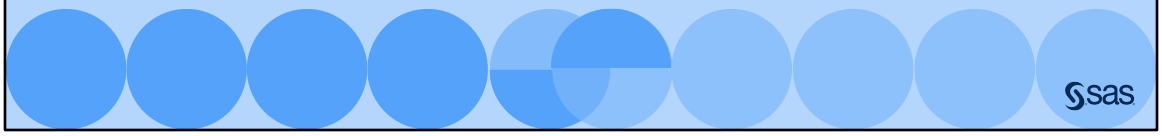


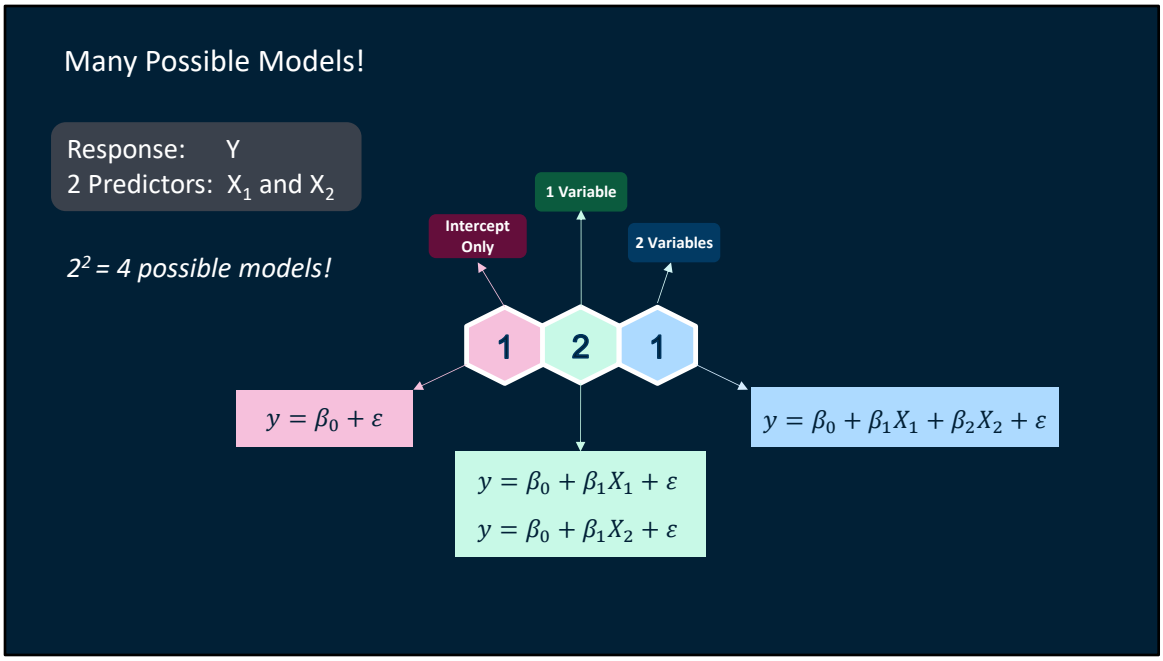
Correct answer: d.

Drug dose affects blood pressure, but that effect is not consistent across diseases. Higher doses result in increased blood pressure for patients with disease B, decreased blood pressure for patients with disease A, and little change in blood pressure for patients with disease C. Hence, an interaction exists between drug dose and disease.

Interactions

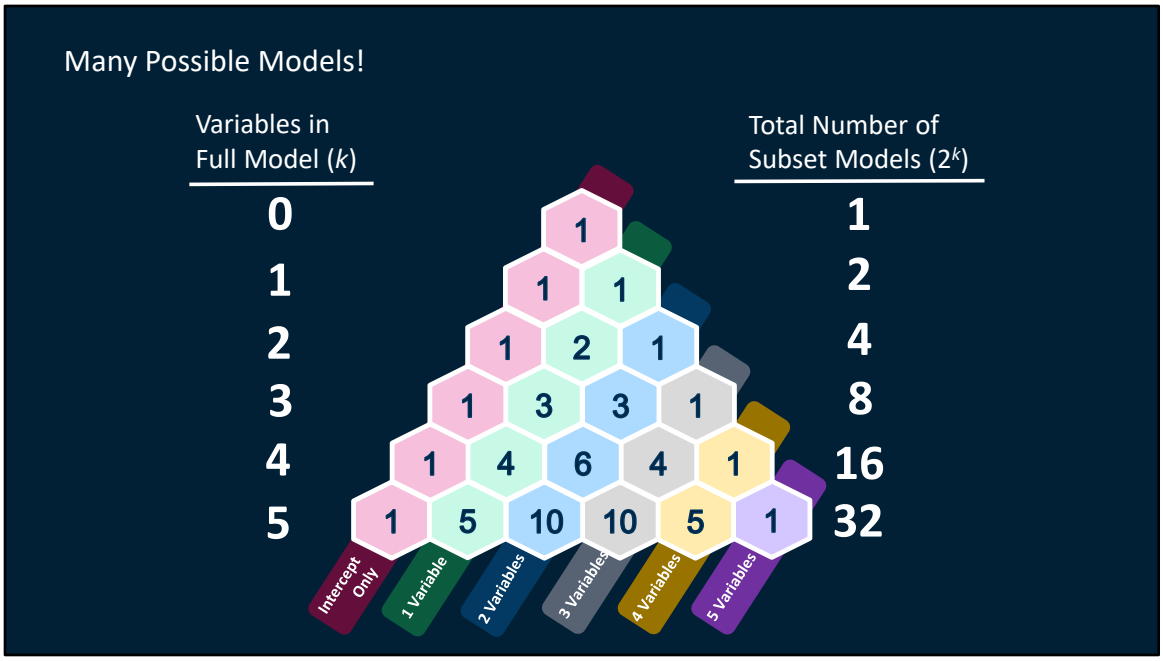
To explore this learning resource in greater detail, please visit the course page on learn.sas.com. There, you'll find comprehensive course materials, tutorials, and updates to support your learning journey.





For even small numbers of predictors, there are many possible regression models (and more so if interactions are included). For K predictors, there are 2 to K-power models. With 2 predictors, X1 and X2, there are 2^2 (4) possible models that could be fit to the data.

The four models include an intercept-only model, two models with 1 predictor (either X1 or X2), and one model with both X1 and X2.

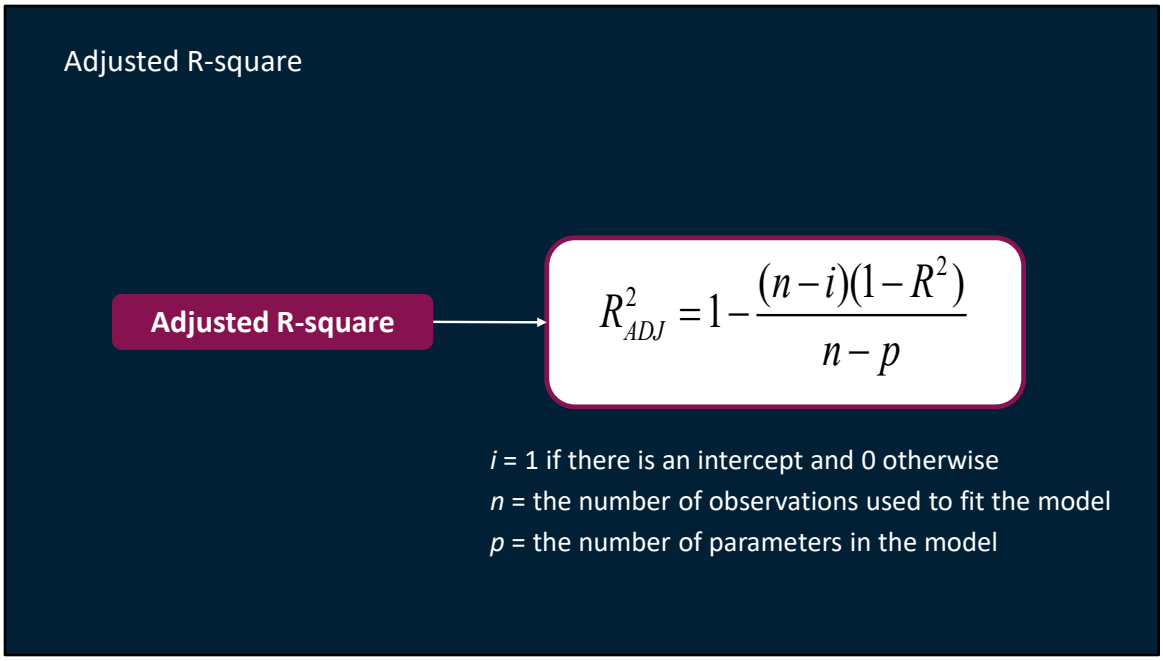


The number of models explodes quickly with increasing inputs and is even larger when considering polynomial and interaction effects.

How does one assess all these models to determine the best one for the data? There are a variety of fit statistics that are useful for comparing regression models.

Some of the fit statistics for comparing regression models include R-square, adjusted R-square, and various information criteria.

R-square, the coefficient of determination, was described in the previous section. So let's look at adjusted R-square and the information criteria.

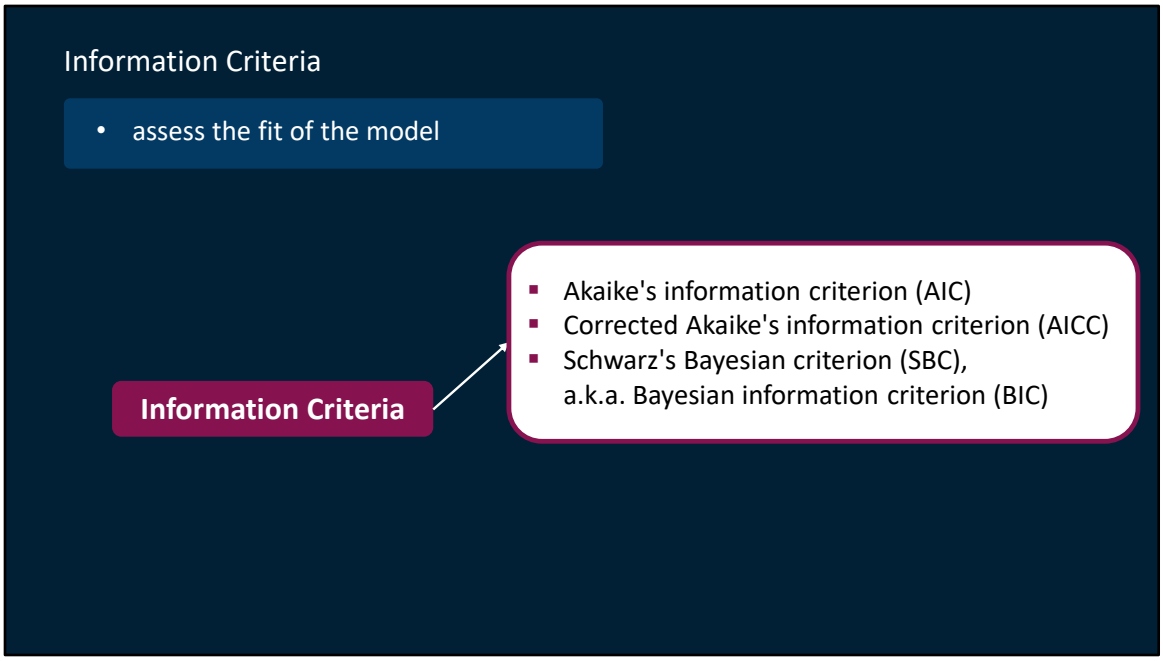


One statistic for comparing multiple regression models is the adjusted R-square. Adjusted R-square is a measure similar to R-square, but it takes into account the number of parameters in the model. It can be thought of as a penalized version of R-square, with the penalty increasing with each parameter added to the model.

As you can see from the equation, adjusted R-square is a function of R-square, the sample size, and the number of parameters in the model.

Why is this modified version of R-square useful? When comparing models with different numbers of effects, R-square is not an ideal choice. R-square increases as you include more terms in the regression model. A model with 10 predictors might have a higher R-square than a model with two predictors just because it has more predictors, even if it doesn't fit the data better.

Adjusted R-square can more fairly be used to compare models with different numbers of effects.

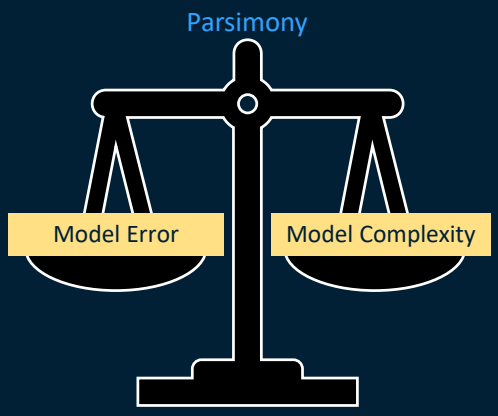


Like adjusted R-square, information criteria assess the fit of the model to your data.

Akaike's information criterion (AIC), the corrected Akaike's information criterion (AICC), and the Schwarz's Bayesian criterion (SBC) - and sometimes called the Bayesian information criterion, or BIC - are three of the most used information criteria.

Information Criteria

- assess the fit of the model
- combine information about the SSE, number of parameters in the model, and the sample size
- penalize the likelihoods in order to select the simplest model
- compare alternative models fitted to the same data set
- a model with a lower information criterion is superior to a model with a higher value



For regression models, information criteria are statistics that combine information about the error sum of squares (SSE), the number of parameters in the model, and the sample size.

Information criteria penalize the likelihoods in order to select the simplest model. These criteria are based upon concepts of information and entropy.

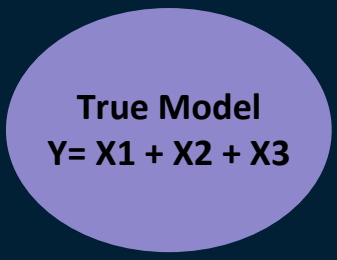
They balance minimizing the error (to explain as much of the variability in the response as possible) and minimizing the complexity of a model (to prevent overfitting). In other words, they search for the most parsimonious model. We don't want our model to become more complex unless the added complexity improves our model! For model selection, the model with the smallest value of an information criterion indicates a better fit.

These information criteria measure relative model fit. Therefore, they're used only to compare alternative models fitted to the same data set.

And all else being equal, a model with a lower information criterion is superior to a model with a higher value.

Next, we'll look at an example of one information criterion, Akaike's information criterion, or AIC.

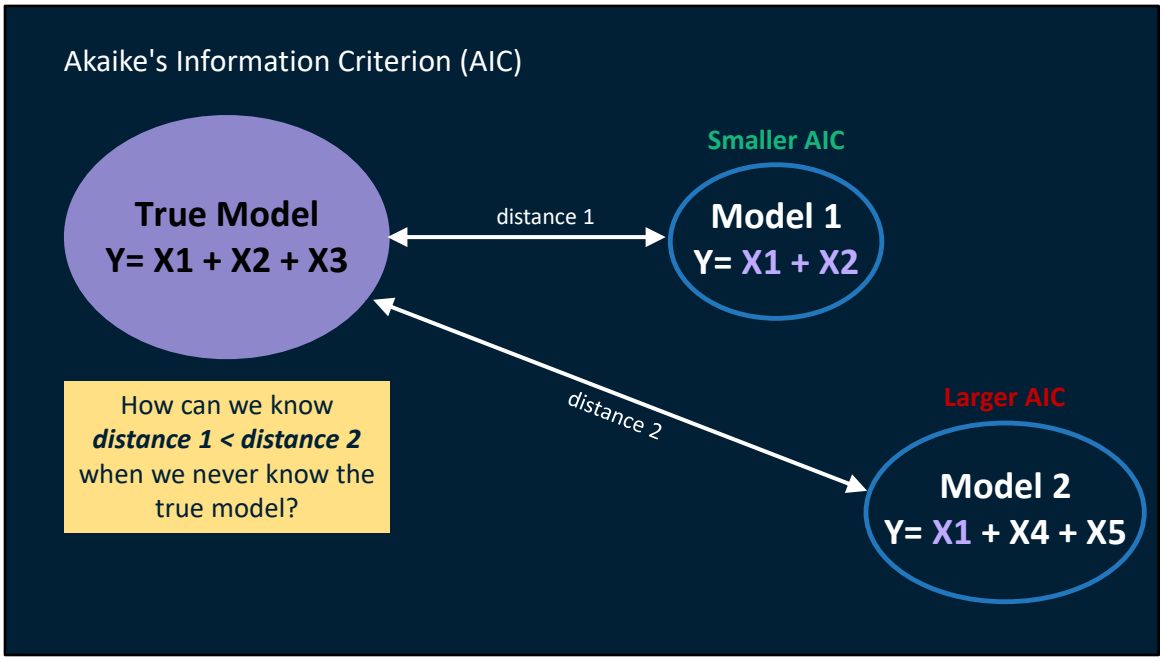
Akaike's Information Criterion (AIC)



- *AIC – A measure of relative similarity between a candidate model and the true model.*
- *Quantifying similarity – distance between a candidate model and the true model.*
- *Distance* – how much information is lost when using a particular model to approximate the true model.

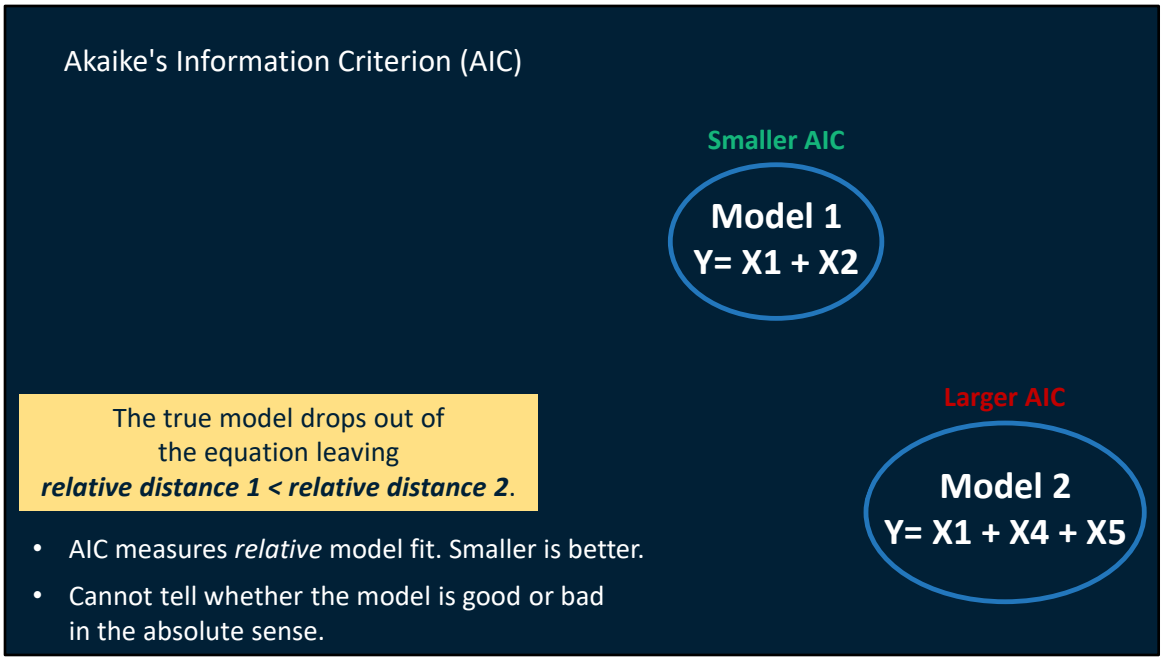
Understanding Akaike's information criterion requires reference to a true model. "True model" is actually an abstract concept. All models are simplifications by definition, and thus there is no such thing as a true model. So "true model" is really short for "best approximating model," but it is helpful to refer to this as the true model for simplicity.

Akaike's information criterion can be thought of as a measure of relative similarity between a candidate model and the unknowable true model. Imagine that we have perfect knowledge of the predictors that explain the variability in your response variable. With knowledge of this true model, we could quantify how close or far any model created from sample data is to this true model. What does "distance" mean in this context? It can be thought of as how much information is lost when using a particular model to approximate the true model. These distances are related to the likelihoods for each model.



Suppose the true model for Y involves the predictors X1, X2, and X3. Now suppose one research group collects a sample from the population and produces the model Y = X1 + X2. A second research group determines from another sample that the best model for Y is X1 + X4 + X5. Model 1 has two of the three correct predictors, and model 2 has only one of the correct predictors and two incorrect predictors. We could say that model 1 is closer to the unknowable true model than model 2. Thus, it will have a smaller AIC than model 2.

But how can we ever know that model 1 is closer to the true model when the truth is unknowable?



Akaike showed that it is possible to calculate the relative distances to the true model, despite not knowing the population parameters. This is possible because the true model drops out of the equation that compares the models, and what is left is the *relative distance* of each model to the true model.

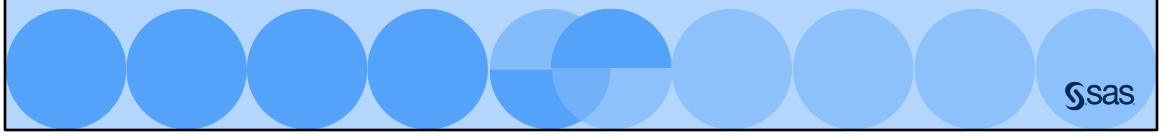
When one model has a lower AIC than another, it can be considered a better model (all else remaining equal) in a relative sense. All the models being compared might be excellent models or poor models in absolute terms, but AIC does not quantify this. It tells us "better" or "worse" than other candidate models, not "good" or "bad" in any absolutes sense.

So AIC as well as other information criteria are useful only for model comparisons, and they don't help us evaluate a single model.

For a readable and detailed explanation of AIC, see *Model Selection and Multimodel inference: A Practical Information-Theoretic Approach* by Burnham and Anderson.

Most Used Information Criteria

To explore this learning resource in greater detail, please visit the course page on learn.sas.com. There, you'll find comprehensive course materials, tutorials, and updates to support your learning journey.



Question: AIC

Fit statistics were calculated for the model

$$\widehat{Income} = 58.1 + 2.35 \cdot Age + 9.02 \cdot HrsWeek$$

including R-square and AIC. What can we conclude from AIC = 660?

a. The high value of AIC indicates good model fit.
b. The nonnegative value of AIC indicates poor model fit.
c. AIC = 660 implies a high R-square.
d. We can't evaluate this model based on AIC because it is a relative measure.

Answer: AIC

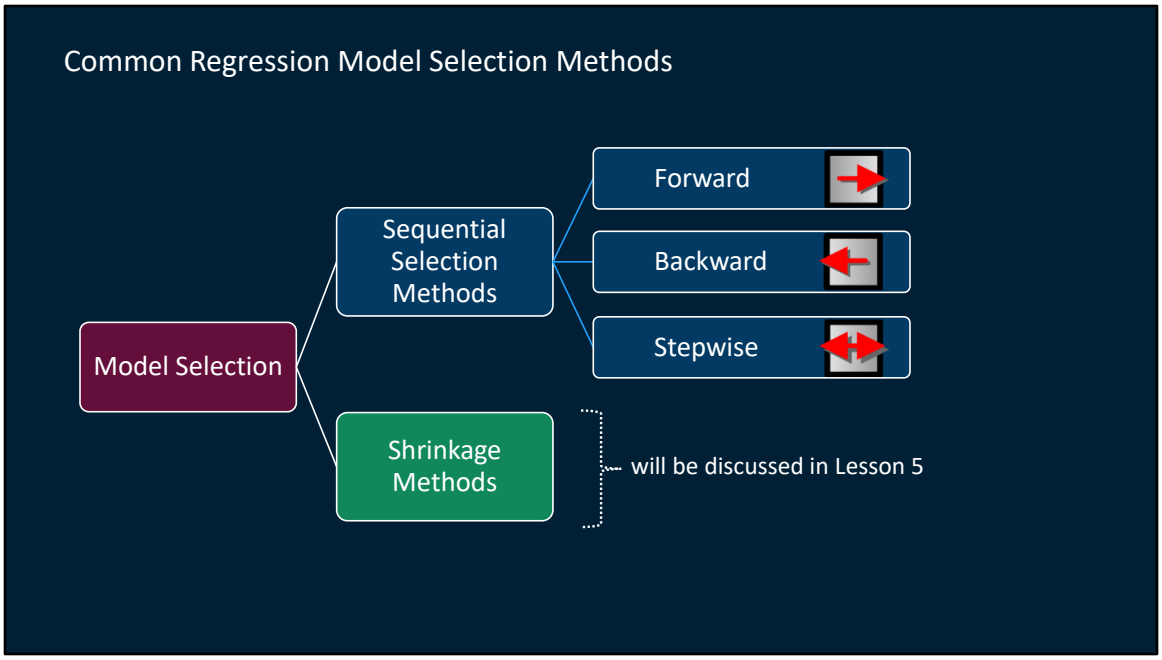
Fit statistics were calculated for the model

$$\widehat{Income} = 58.1 + 2.35 \cdot Age + 9.02 \cdot HrsWeek$$

including R-square and AIC. What can we conclude from AIC = 660?

a. The high value of AIC indicates good model fit.
b. The nonnegative value of AIC indicates poor model fit.
c. AIC = 660 implies a high R-square.
d. We can't evaluate this model based on AIC because it is a relative measure.

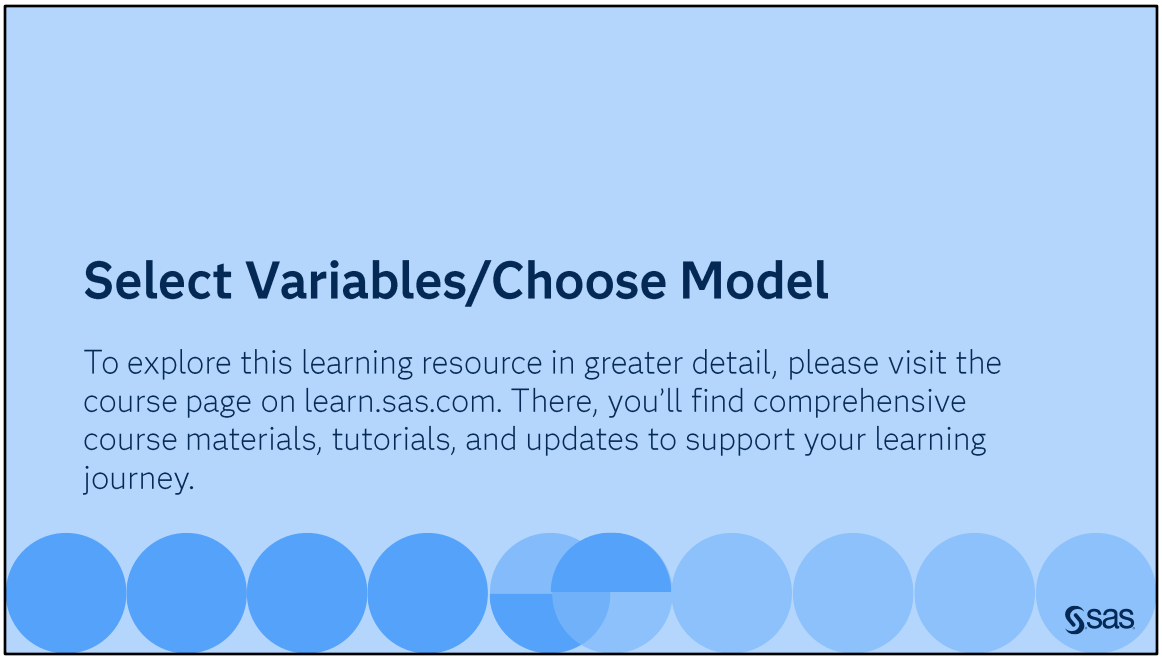
The correct answer is d. AIC and all other information criteria can be used to evaluate models as better or worse in a relative sense but not on an absolute scale. Large or small values of any information criterion do not indicate the quality of a single model.

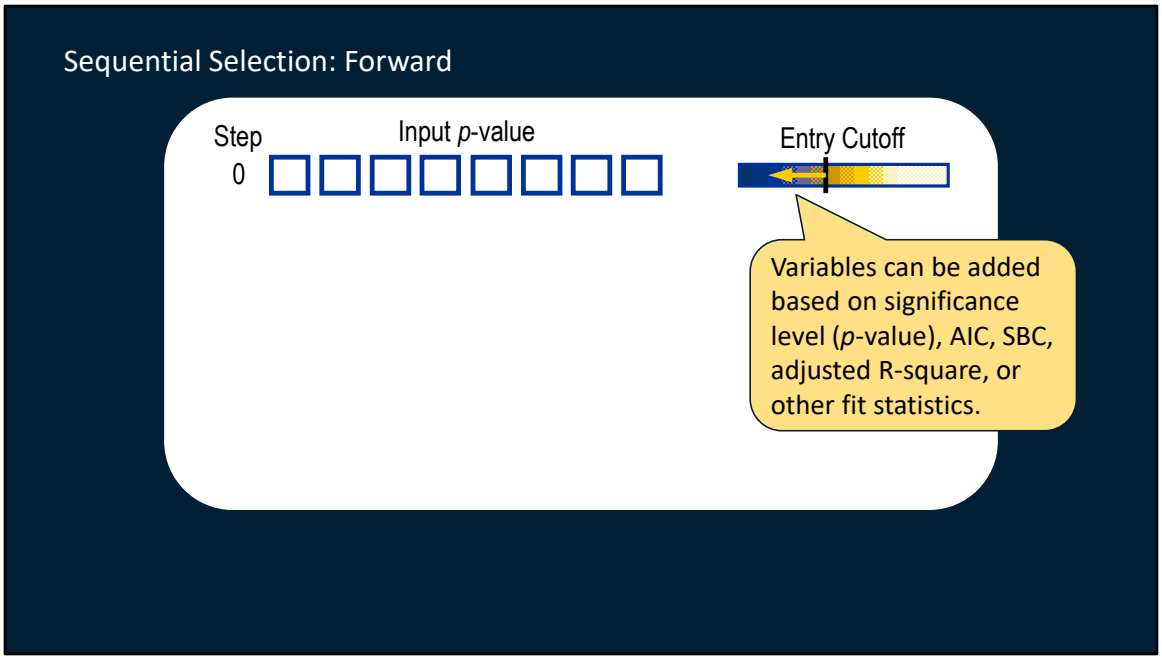


Now that we have fit statistics like adjusted R-square and information criteria such as AIC and SBC to compare models, how do we come up with candidate models in the first place?

With many predictors that could potentially be included in a model, trying all possible models is often not feasible. We need a computationally efficient way of searching through possible models to find a subset that will fit the data well. Many of these model selection methods exist for regression models. Some have existed for decades, such as the sequential selection methods. Other methods, such as the shrinkage methods listed here, are newer additions to the statistical toolbox and will be discussed in lesson 5.

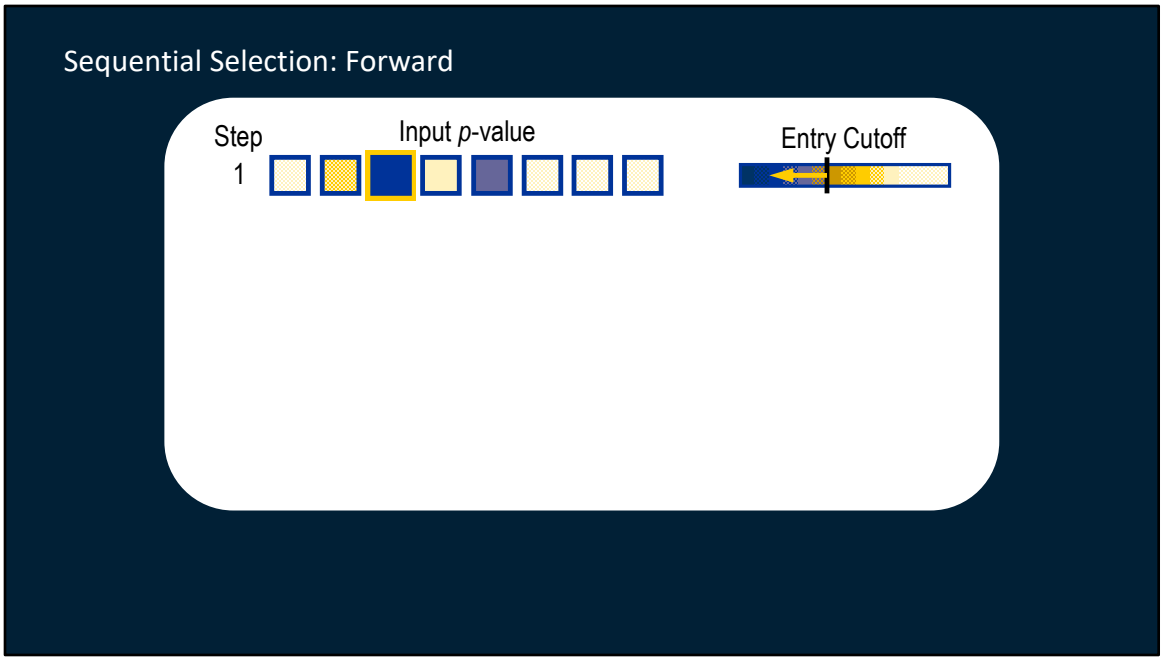
We're going to focus on the most common sequential methods: forward, backward, and stepwise model selection.





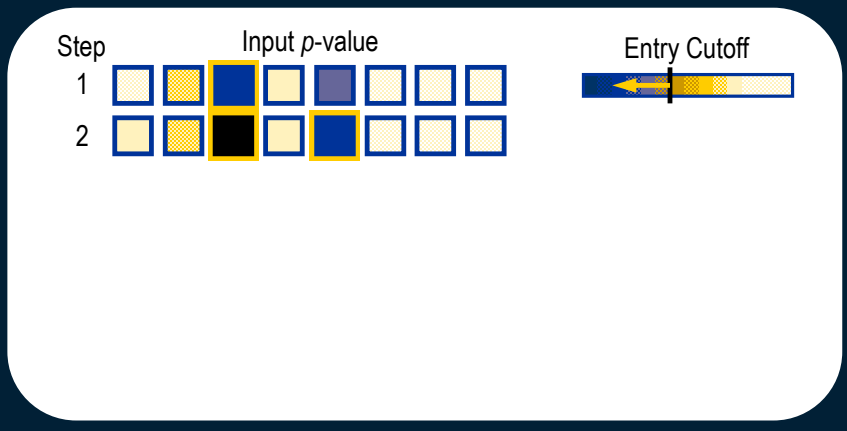
Let's look more closely at some of the traditional model selection methods. Forward selection creates a sequence of models of increasing complexity. The sequence starts with the baseline model, an intercept-only model predicting the overall average target value for all cases. We'll describe this process based on significance tests, but variables could also be added based on information criteria, adjusted R-square, or other fit statistics.

For example, here out of eight potential predictors, the most significant variable will be represented by the darkest blue color, whereas the most insignificant will be represented by the lightest gold color.



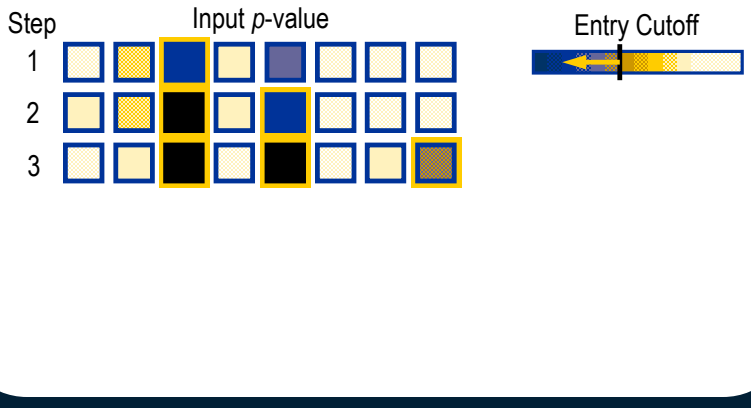
Forward selection adds predictors one at a time based on having the lowest *p*-value, as long as it is under a predetermined threshold value, called the *entry cutoff*. The algorithm searches the set of one-input models and selects the model that most improves on the baseline model.

Sequential Selection: Forward

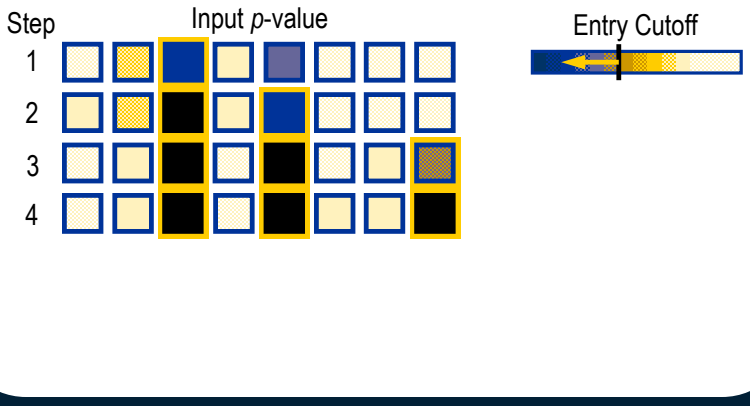


In forward selection, once variables are added, they are never removed. The next variable to be added is the one with the lowest *p*-value under the significance level to enter the model. By adding a new input to those selected in the previous step, a nested sequence of increasingly complex models is generated.

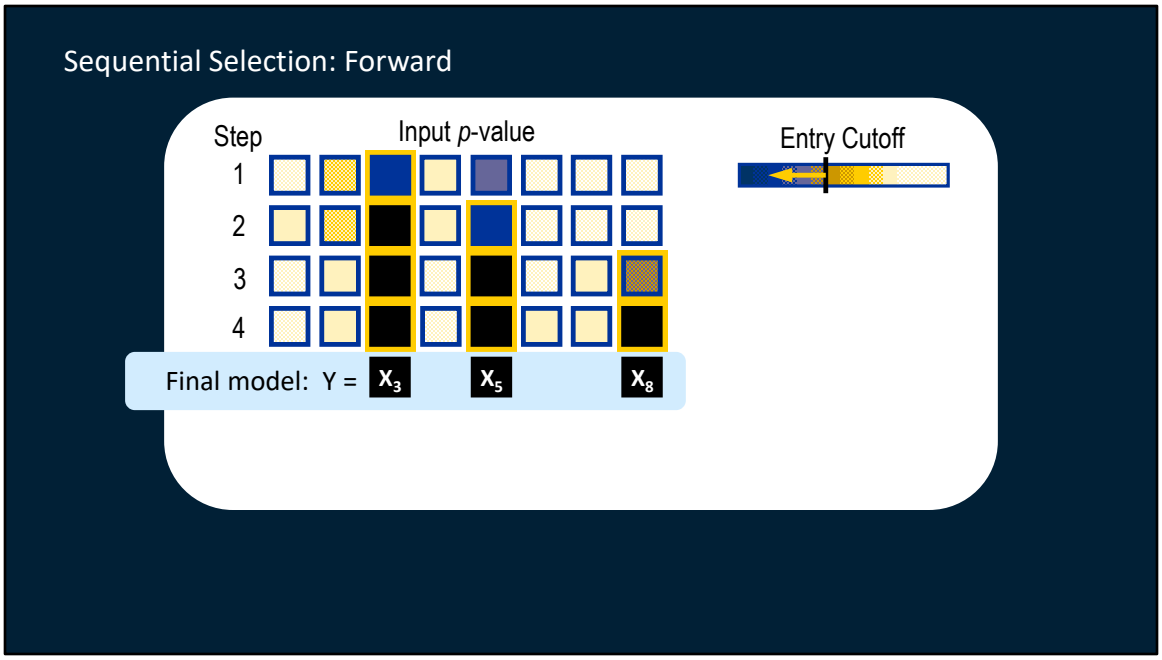
Sequential Selection: Forward



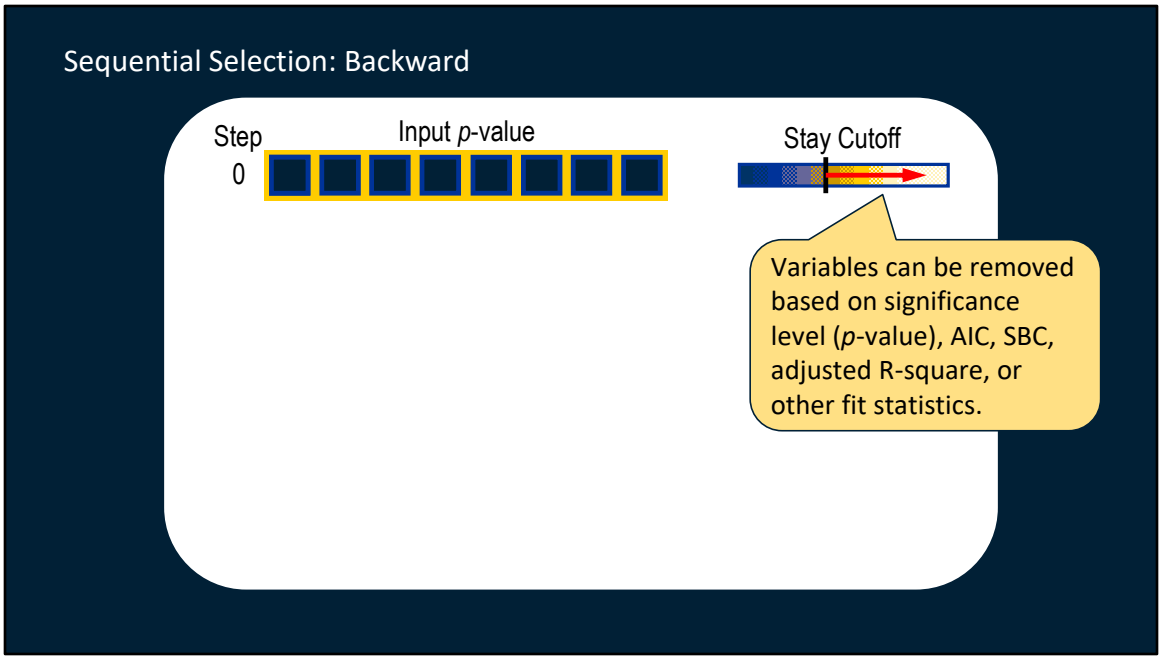
Sequential Selection: Forward



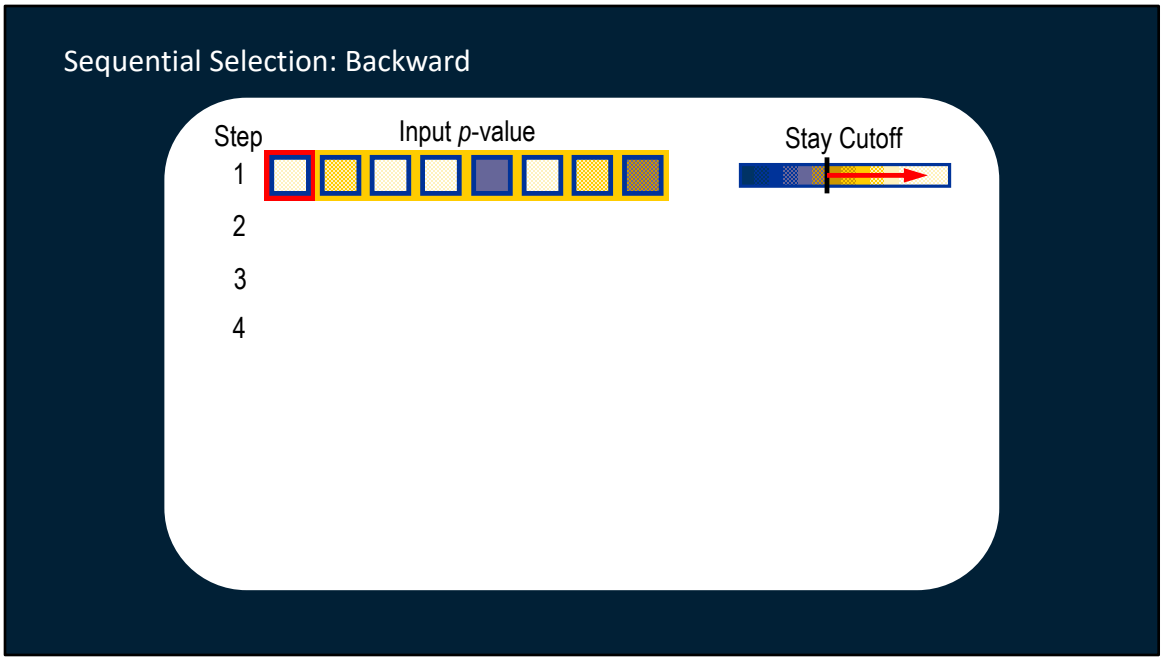
Adding terms in this nested fashion always increases a model's overall fit statistic.



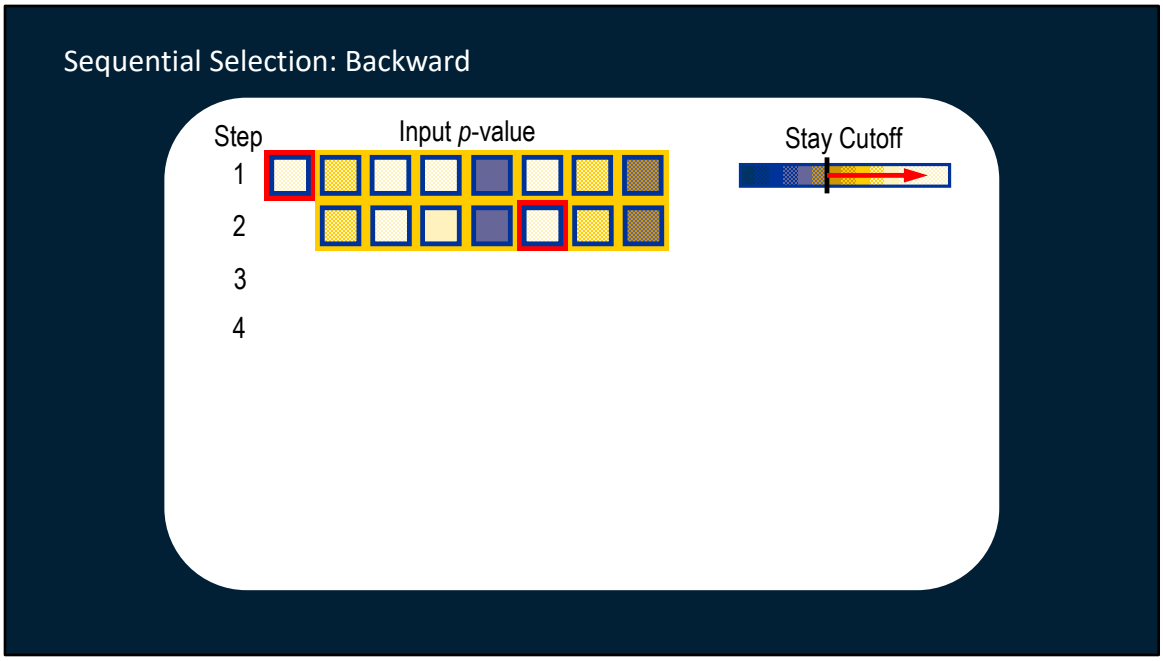
Eventually, all variables outside of the model have p-values too high to enter, and the forward selection process stops.



In contrast to forward selection, backward selection creates a sequence of models of decreasing complexity. The sequence starts with a saturated model, which is a model that contains all available inputs, and therefore, has the highest possible fit statistic.

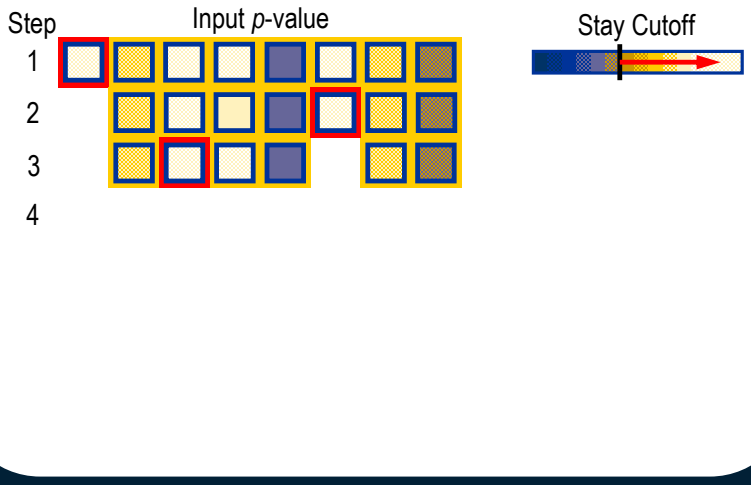


When significance level is the selection criterion, the variable with the highest *p*-value is removed first, as long as the *p*-value is above a predetermined threshold, called the *stay cutoff*.

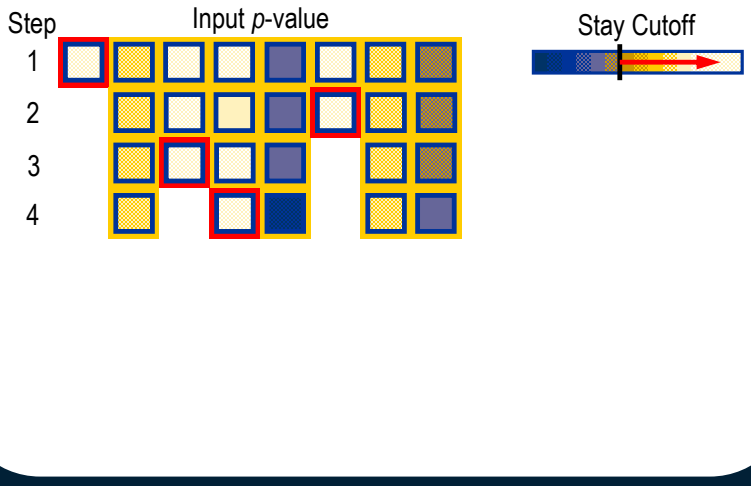


Variables continue to be removed one at a time based on *p*-values. Variables are not allowed to reenter the model.

Sequential Selection: Backward

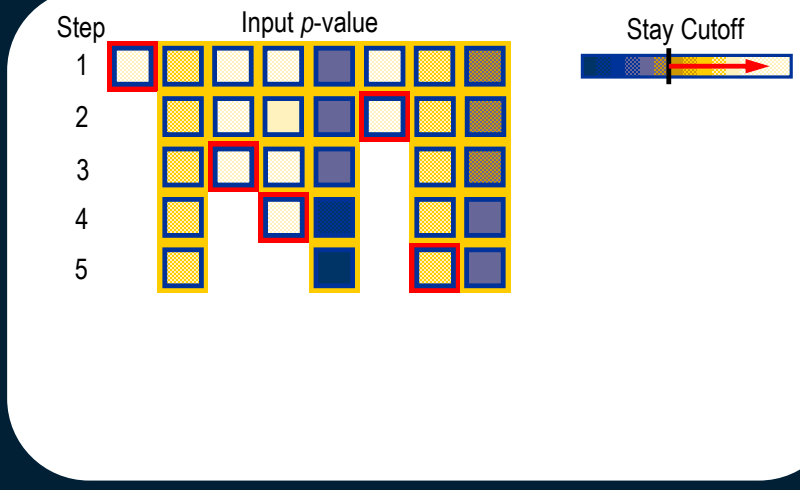


Sequential Selection: Backward

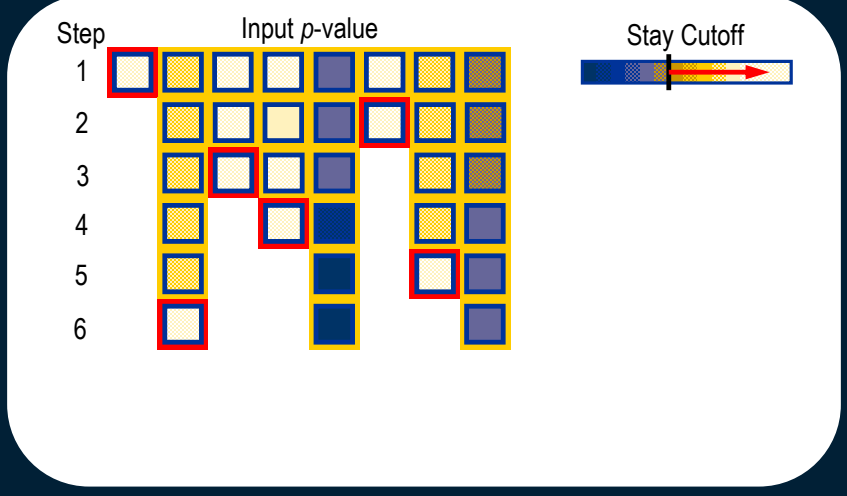


Inputs are sequentially removed from the model. At each step, the input chosen for removal least reduces the overall model fit statistic.

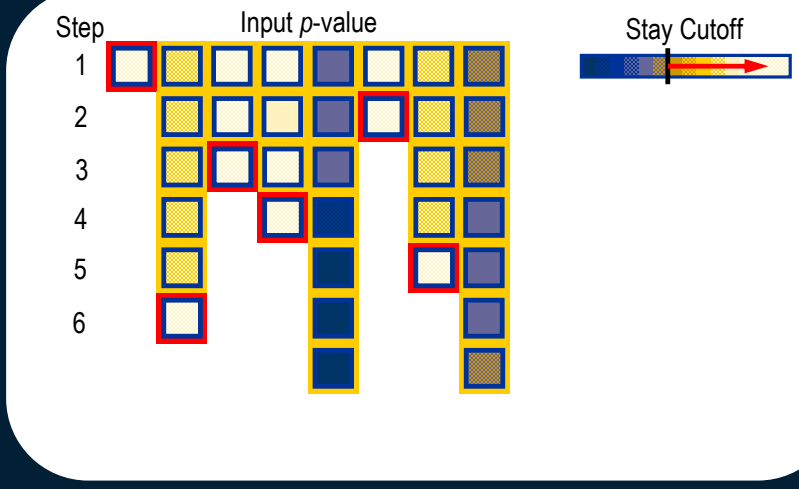
Sequential Selection: Backward

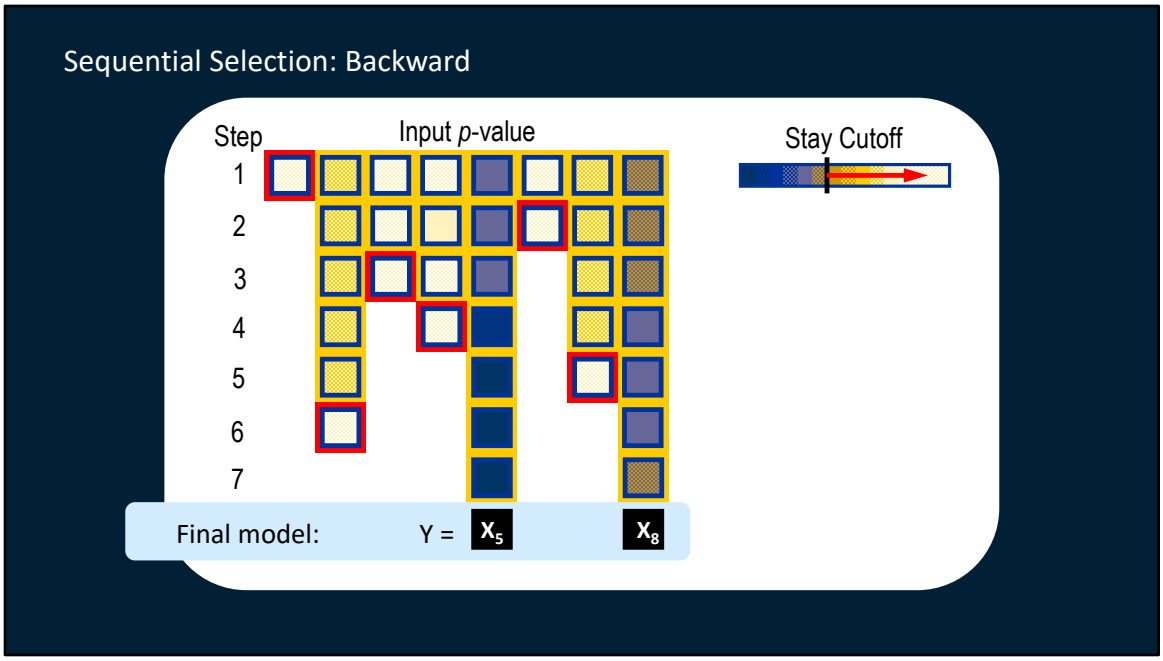


Sequential Selection: Backward

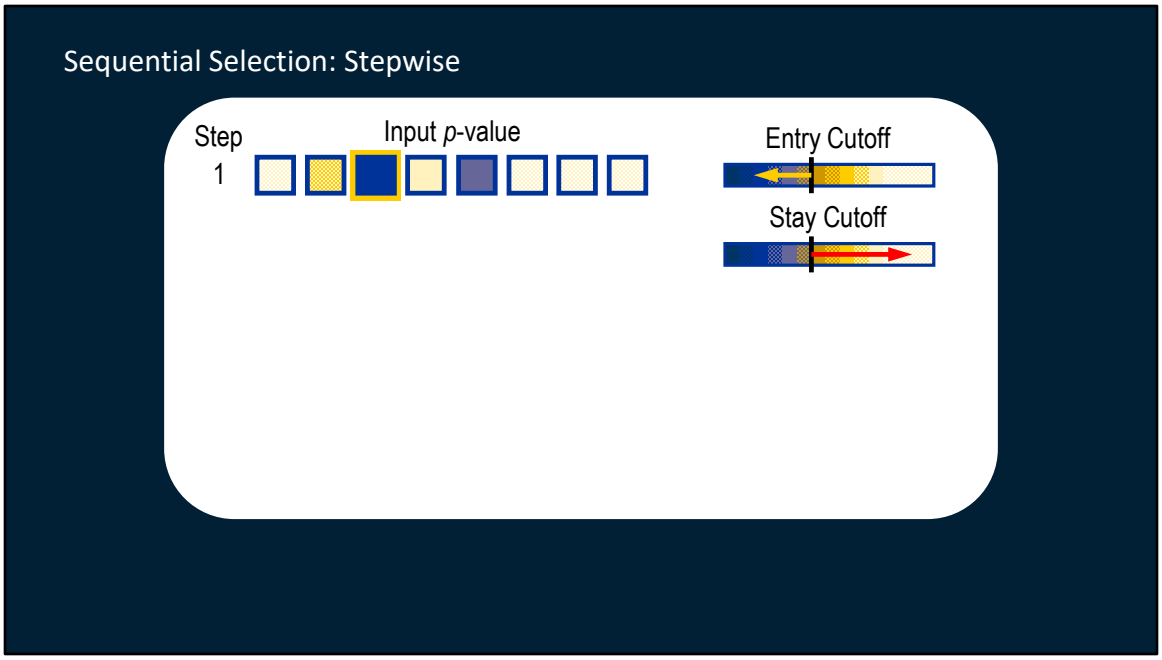


Sequential Selection: Backward



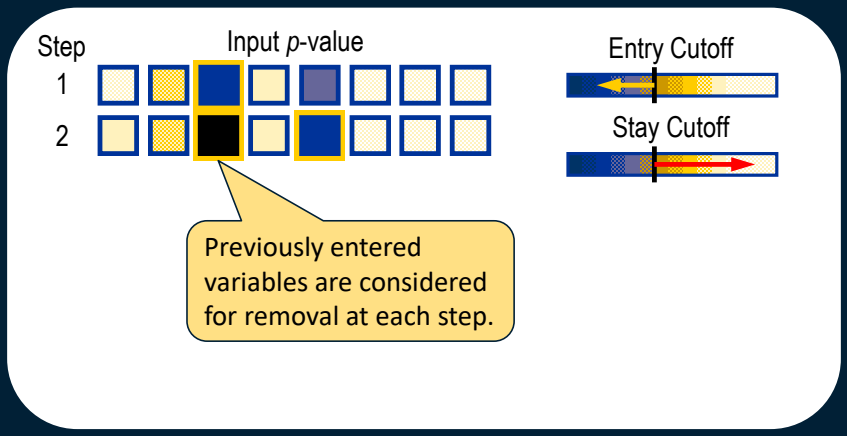


When all variables in the model have *p*-values low enough to stay, the backward elimination process stops.



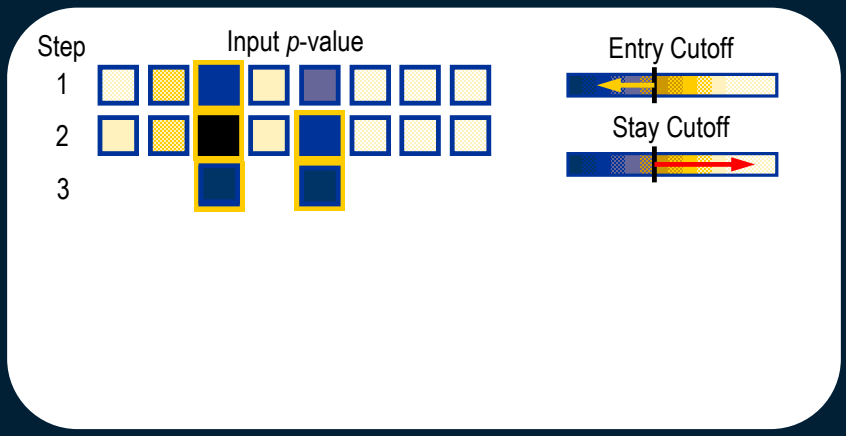
Stepwise combines elements of forward selection and backward elimination. It begins with an intercept-only model. Variables are added one at a time as in forward selection.

Sequential Selection: Stepwise



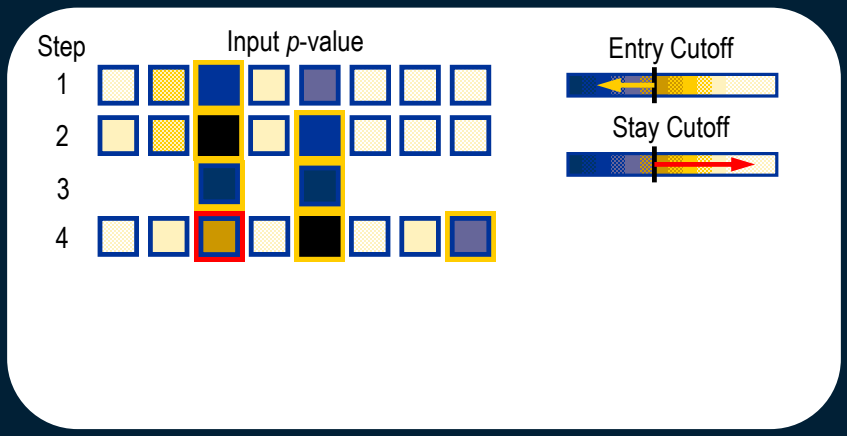
Once a second variable is added, previously added variables are removed in a backward step if they no longer meet the criterion for remaining in the model.

Sequential Selection: Stepwise



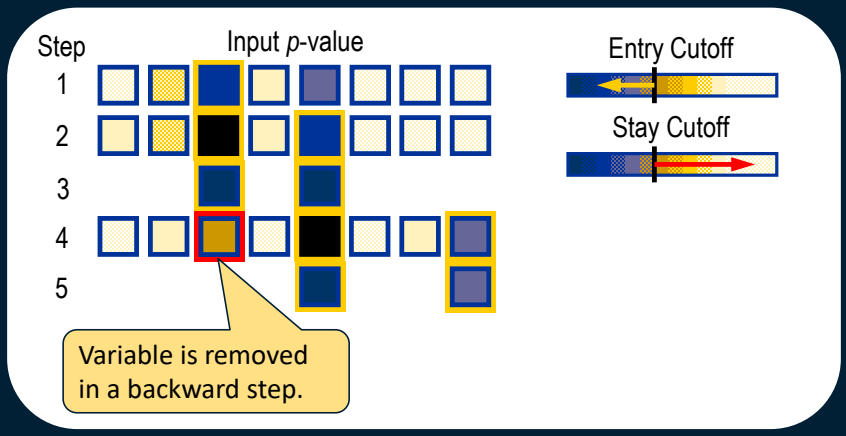
In the backward elimination stage of this example, neither of the two variables were removed because they both met the stay cutoff.

Sequential Selection: Stepwise



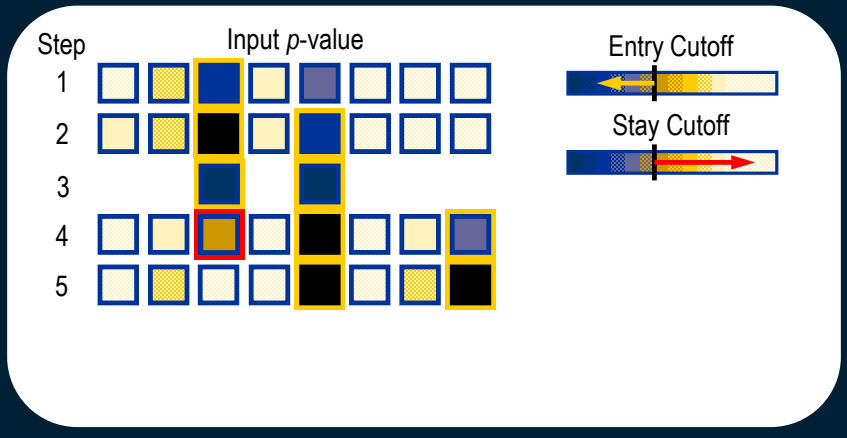
In the next step, when another variable entered in the model, one of the previous variables that was present in the model became insignificant.

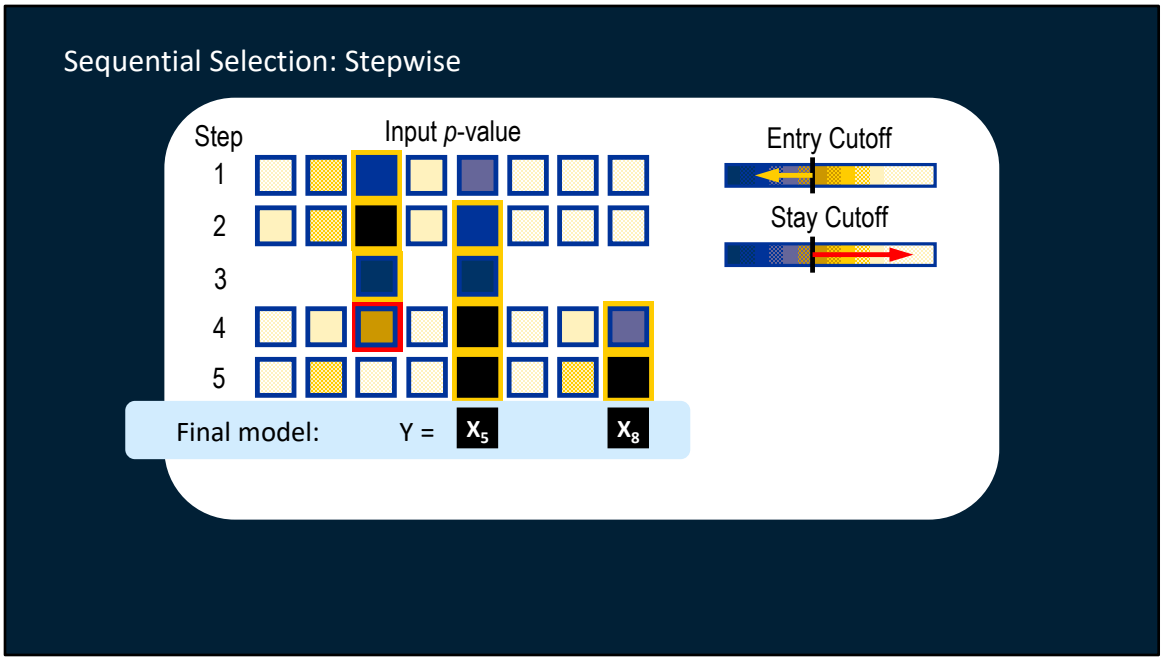
Sequential Selection: Stepwise



In step 4, a previously added variable is removed in a backward step.

Sequential Selection: Stepwise





The process terminates when all inputs available for inclusion in the model have *p*-values in excess of the entry cutoff, and all inputs already included in the model have *p*-values below the stay cutoff.

Question: Sequential Selection Methods

The STEPWISE, BACKWARD, and FORWARD strategies result in the same final model if the same significance levels are used in all three.

a. True
b. False

Answer: Sequential Selection Methods

The STEPWISE, BACKWARD, and FORWARD strategies result in the same final model if the same significance levels are used in all three.

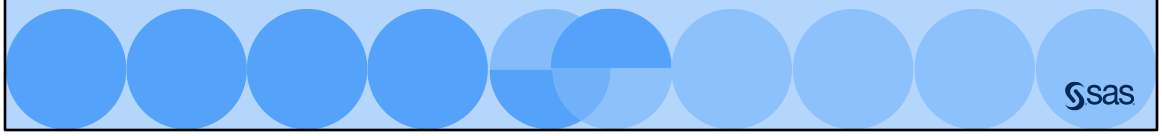
a. True
b. False

Answer: False

Because the techniques for selecting or eliminating effects differ among the three sequential methods, they don't always produce the same final model.

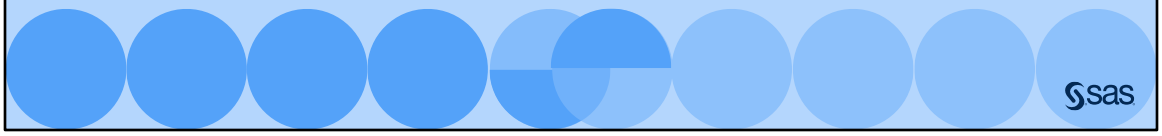
Problems with Stepwise Selection Methods

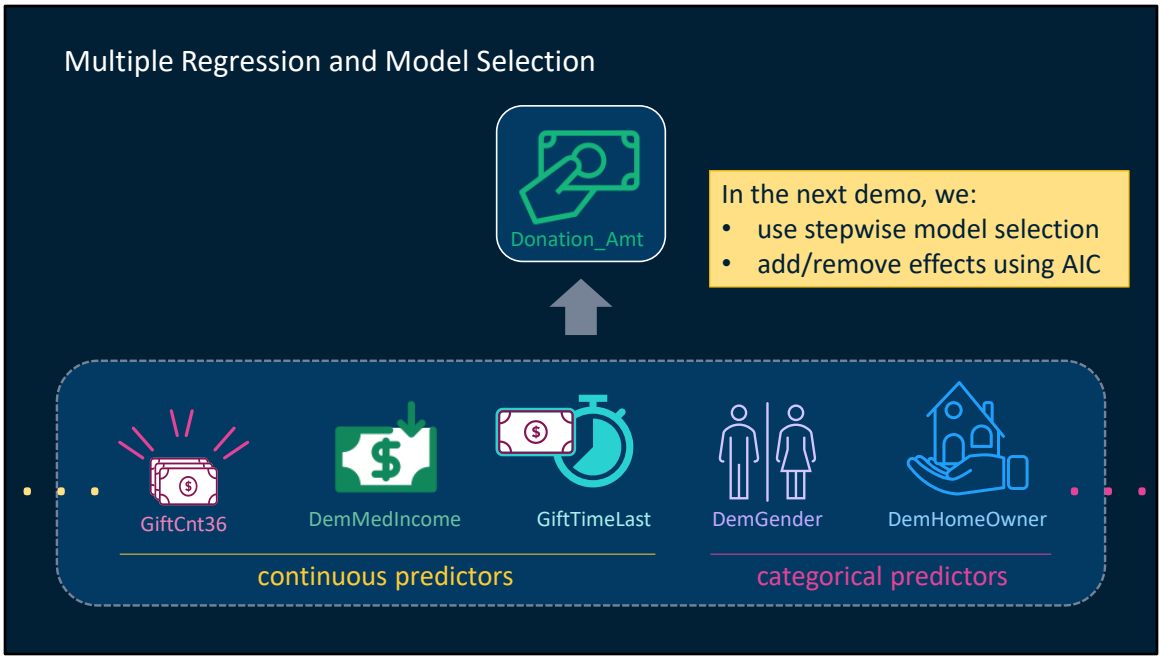
To explore this learning resource in greater detail, please visit the course page on learn.sas.com. There, you'll find comprehensive course materials, tutorials, and updates to support your learning journey.



Shrinkage Methods

To explore this learning resource in greater detail, please visit the course page on learn.sas.com. There, you'll find comprehensive course materials, tutorials, and updates to support your learning journey.





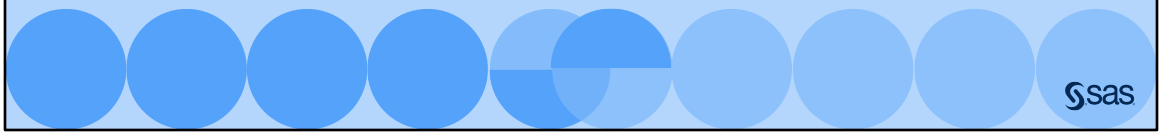
In the next demonstration, we fit a multiple linear regression model. We will include continuous predictors such as **Gift Count 36 Months**, **DemMedIncome**, and **GiftTimeLast** and categorical predictors such as **DemGender** and **DemHomeOwner**. We will use stepwise model selection and use Akaike's information criterion to add and remove variables from the model.

Demo: Multiple Regression and Model Selection

To explore this learning resource in greater detail, please visit the course page on learn.sas.com. There, you'll find comprehensive course materials, tutorials, and updates to support your learning journey.

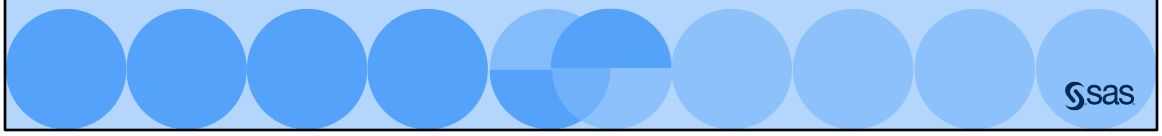
Candidate Models

To explore this learning resource in greater detail, please visit the course page on learn.sas.com. There, you'll find comprehensive course materials, tutorials, and updates to support your learning journey.

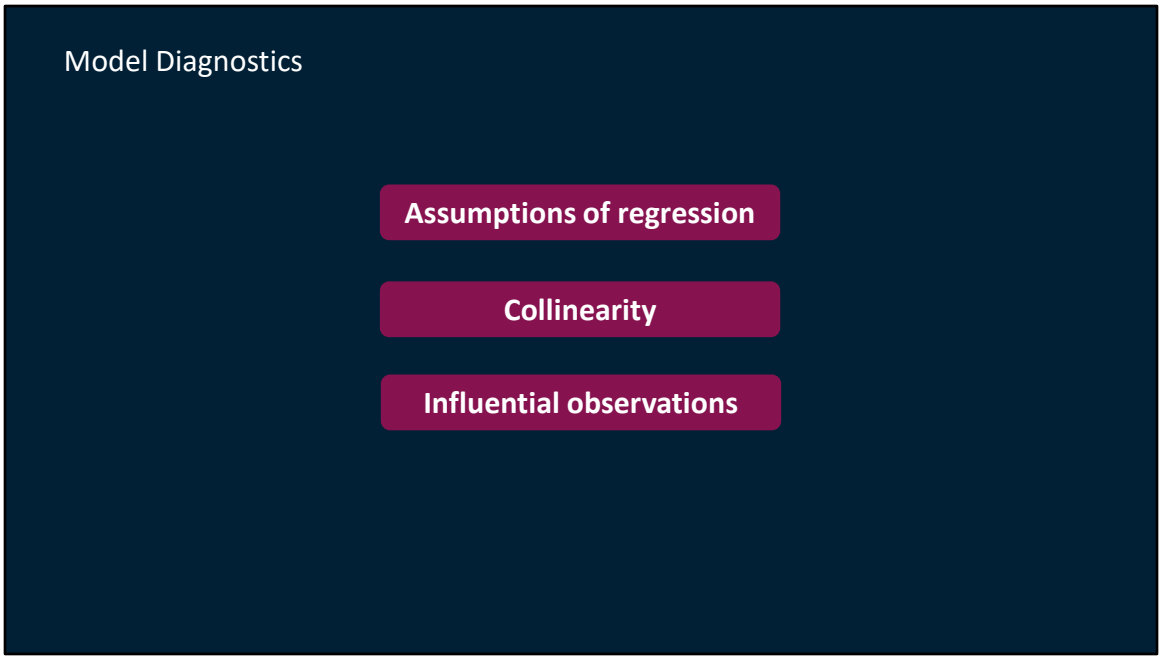


Which Model to Use?

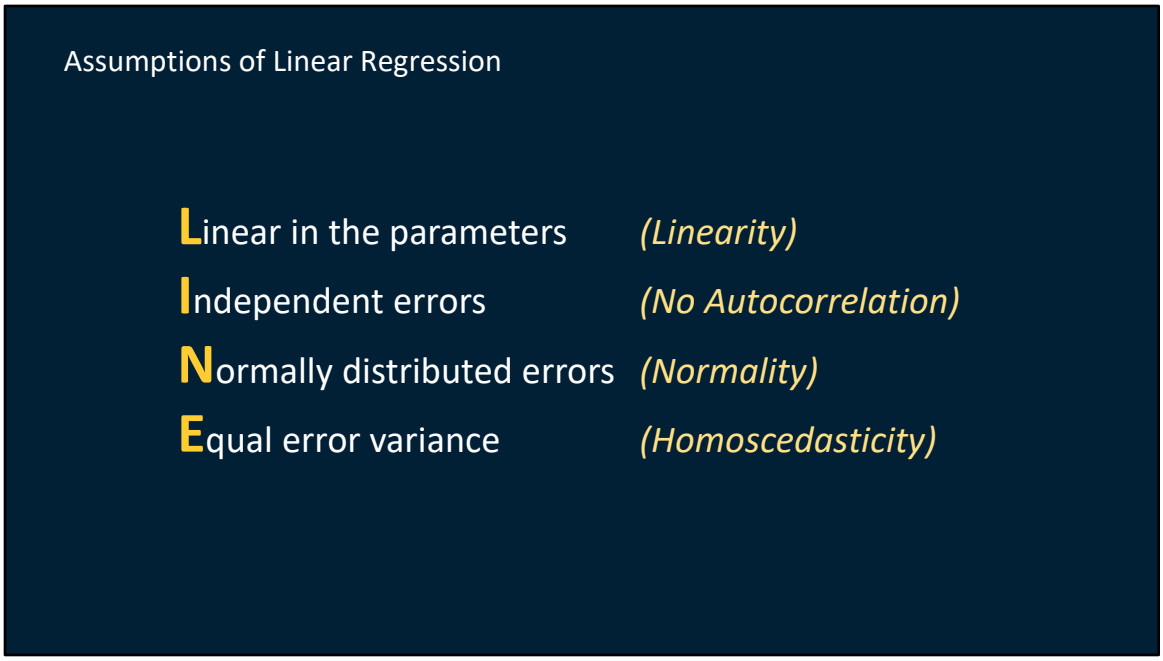
To explore this learning resource in greater detail, please visit the course page on learn.sas.com. There, you'll find comprehensive course materials, tutorials, and updates to support your learning journey.



3.3 Model Diagnostics



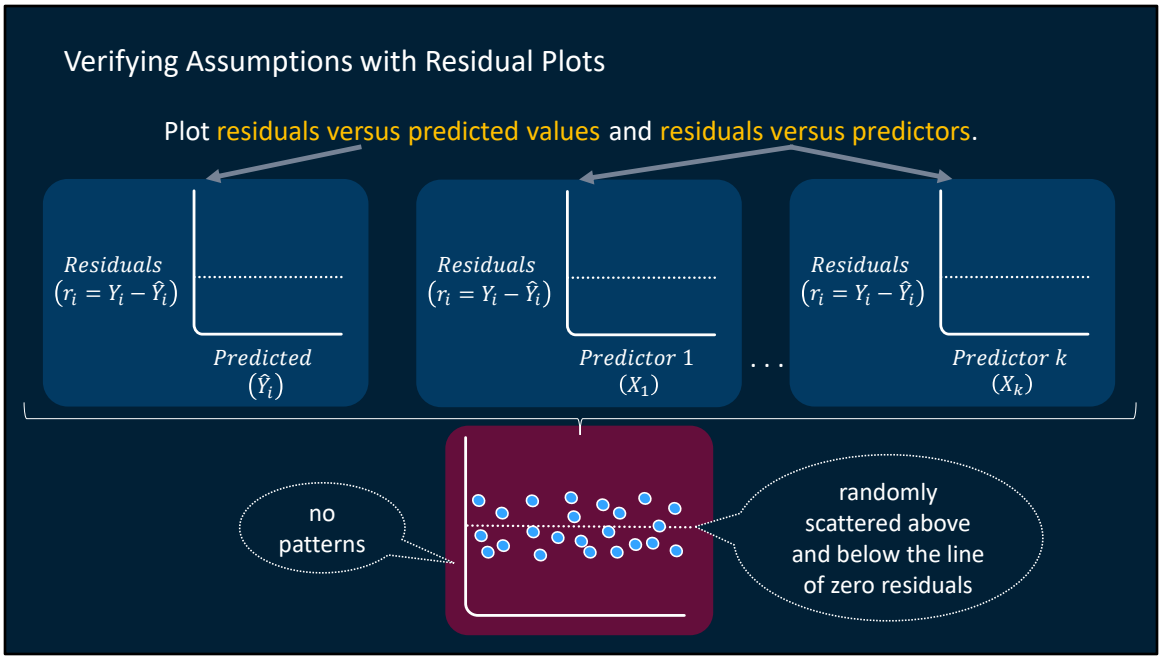
Model diagnostics for regression can involve checking the assumptions of the model, testing for collinearity (strong correlations among sets of predictors), and looking for outliers and influential observations.



Ordinary least squares regression models require making several assumptions for statistical inference and explanatory modeling. If one just wants the best linear unbiased estimates of the regression parameters, these assumptions are not necessary. But the majority of the time, analysts want to test hypotheses about the parameter estimates. This includes calculating standard errors, test statistics such as *T* and *F*, confidence intervals, and *p*-values. For these statistics to be trusted, the assumptions of the linear model must be met.

The first assumption is linear in the parameters. If the relationship between the predictors and the response is not linear, you are using the wrong model! When the model is heavily misspecified, the results will not be meaningful.

Independent normally distributed errors with constant or equal variance errors are all required for statistical inference. Conveniently, these assumptions form the acronym LINE, making them easy to remember.

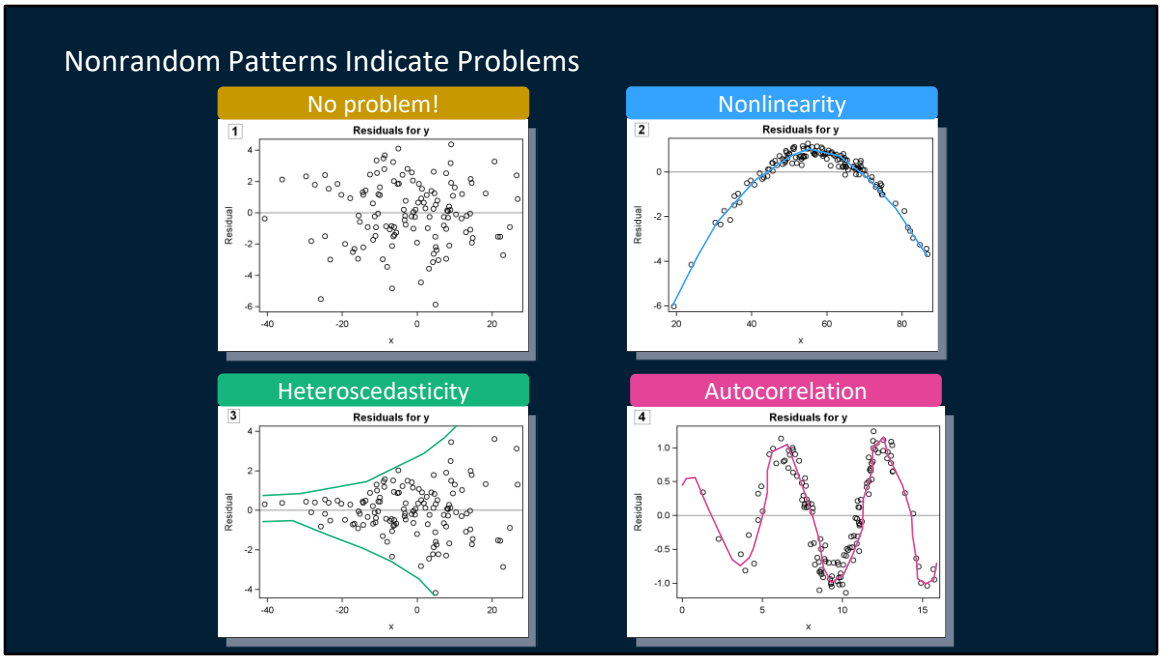


With most of these assumptions being about the errors, how do we check these assumptions?

Most of the OLS assumptions can be checked by graphing and analyzing the residuals. Common plots for checking the assumptions include plots of residuals versus predictions from the model and residuals versus each of the predictor variables, the Xs.

A random scatter above and below the line of zero residuals with no obvious patterns shown in these graphs indicate that the model assumptions were satisfied.

Nonrandom Patterns Indicate Problems



Residuals can be plotted against both predicted values from the model and against individual predictors. A random scatter such as in the top-left picture indicates no problems with the model. When patterns exist, they can indicate violations of model assumptions.

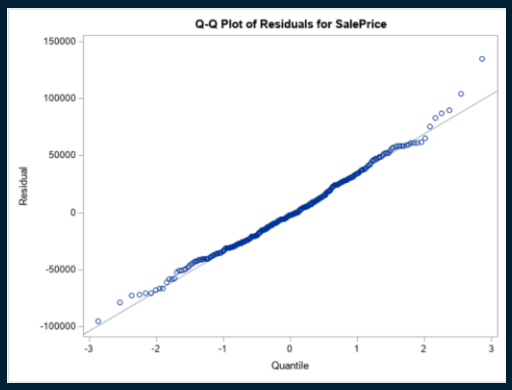
A curvilinear residual pattern (top right) can indicate a misspecified model.

A funnel shape (bottom left) indicates a violation of the constant variance assumption. Cyclical patterns (bottom right) can indicate non-independence.

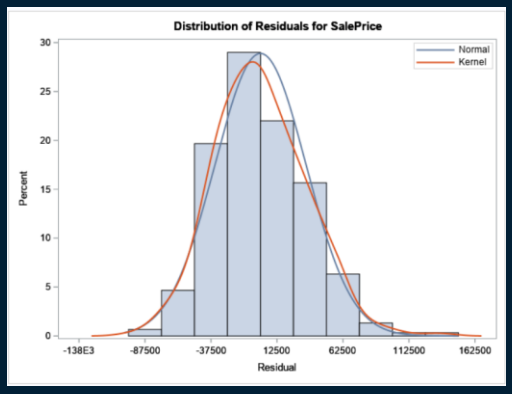
Violations of these assumptions can limit the inferences that an explanatory modeler can make from a regression, but are less problematic when the regression is used for pure prediction.

Graphical assessment of normality requires different plots.

Check Normality with Other Plots



normal quantile-quantile plot



histogram of residuals

The assumption of normality can be checked using a normal quantile-quantile plot and a histogram of residuals.

Normal quantile-quantile (QQ) plots show the quantiles of a variable plotted against the quantiles of a standard normal distribution. The closer the data points fall along the equality line, the closer the variable is to being normally distributed. Histograms of the residuals that roughly show the bell curve shape are consistent with normality.

The normality assumption can also be assessed through descriptive statistics (for example, kurtosis and skewness statistics near zero) and formal hypothesis tests.

Question: Regression Assumptions

Predictor variables are assumed to be normally distributed in linear regression models.

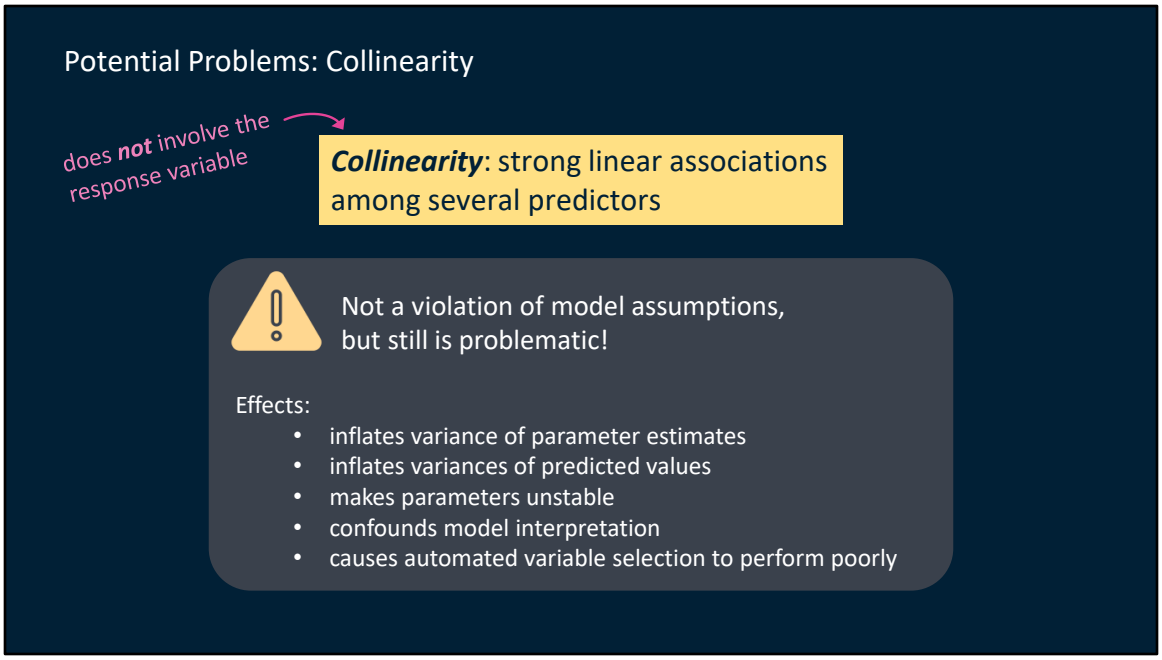
a. True
b. False

Answer: Regression Assumptions

Predictor variables are assumed to be normally distributed in linear regression models.

a. True
b. False

Answer: False
In linear regression, errors are assumed to be normally distributed.

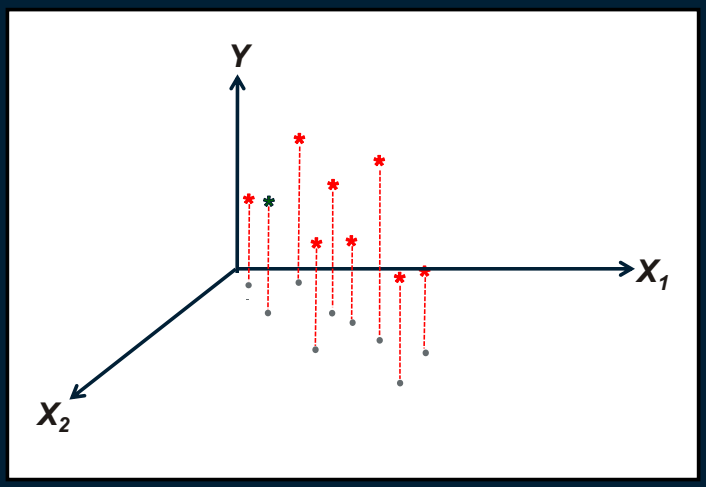


Besides checking regression model assumptions, model assessment can include checking for collinearity. Collinearity (also called multicollinearity) means strong associations among sets of predictors. Collinearity is an unsupervised concept. That is, it involves only predictors, not the response variable. Collinearity can exist between a pair of variables but can also involve three or more predictors.

Collinearity is not a violation of the assumptions of regression, but it can cause several problems for modeling. When collinearity exists among predictors, the parameter estimates become unstable, and their standard errors become inflated. Changes in the standard errors will alter p-values and make model interpretation more difficult. The variance of predicted values can increase too, especially if the values of the predictors are not in the range of the sample that generated the model. Collinearity can also cause automated variable selection methods to perform poorly, resulting in a random set of predictors being selected for your model.

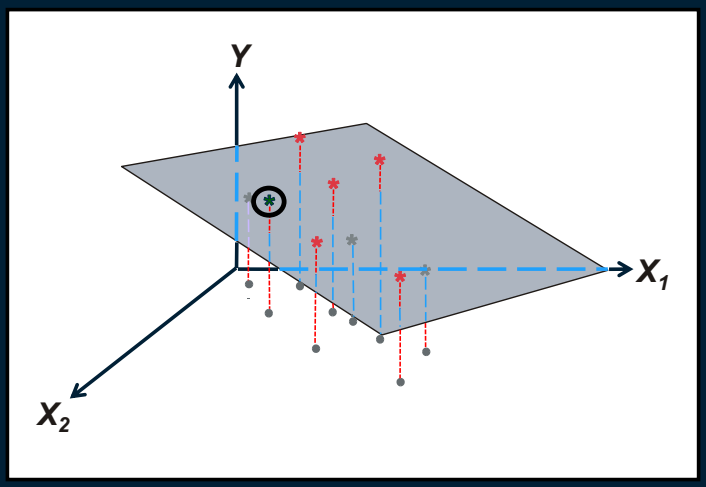
Let's see a graphical depiction of collinearity.

Illustration of Collinearity



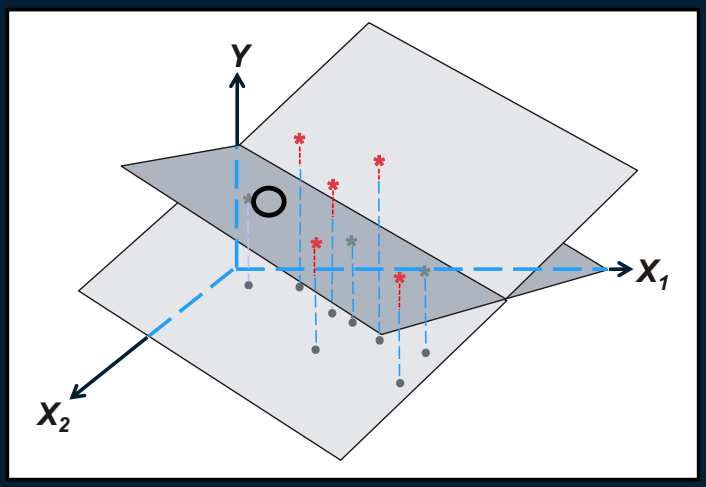
In this illustration of collinearity, the drop shadows for the data points show that X1 and X2 are correlated. This correlation makes the regression coefficients unstable. What does instability really mean? It means small changes in the data can lead to large changes in the parameter estimates.

Illustration of Collinearity

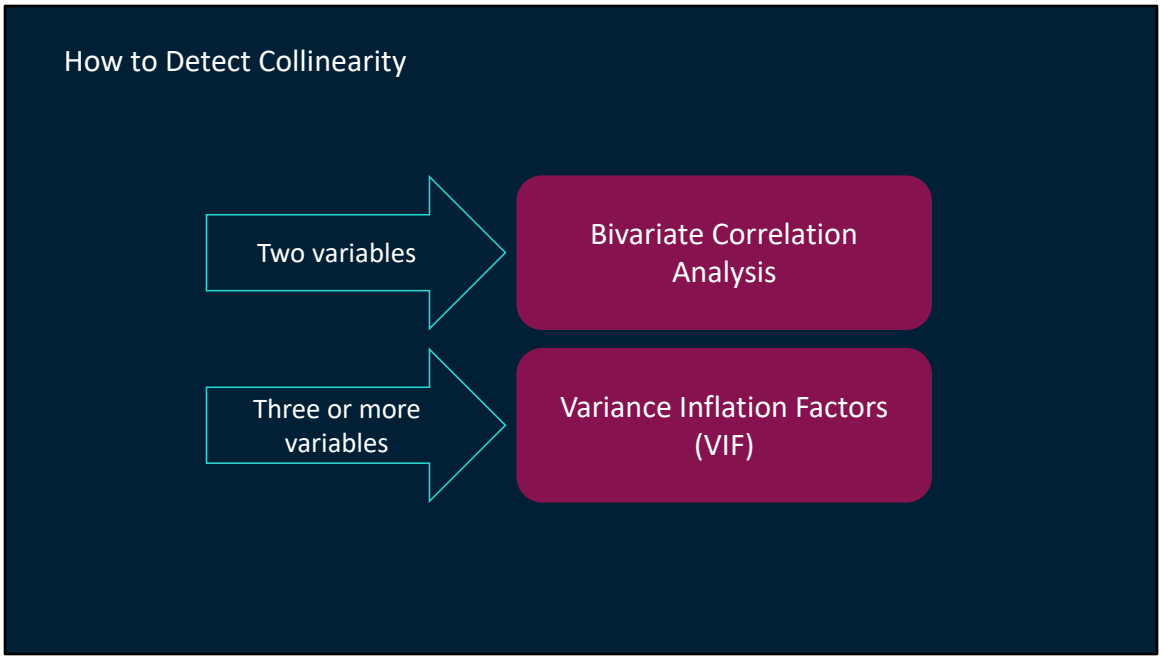


If we fit a multiple regression model to these data, the estimated coefficients can be visualized as the tilt of the best-fit regression plane. With collinearity, these regression coefficients are unstable. What would happen if we remove a single data point and refit the model?

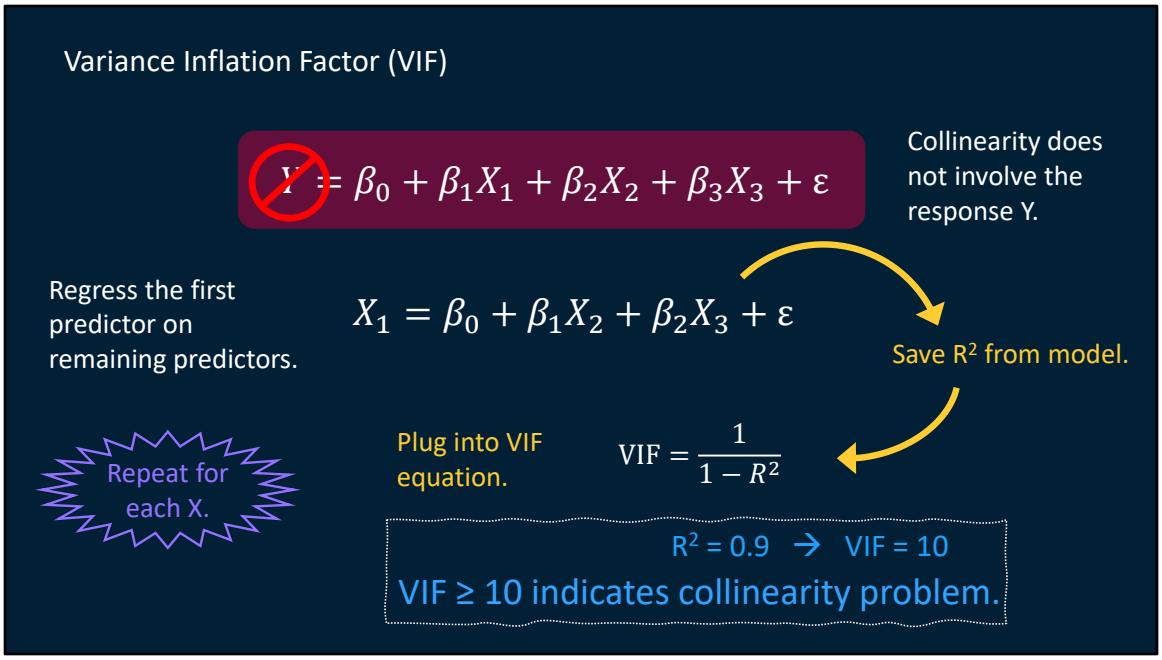
Illustration of Collinearity



Removal of a single data point causes the slopes to change sign and magnitude! Small changes in the data lead to changes in the parameter estimates that would drastically alter interpretation of this model.



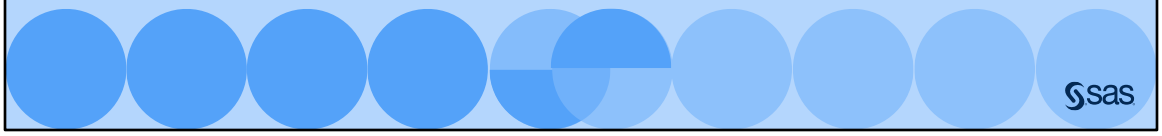
How can we detect collinearity? Collinearity between two variables can be assessed through bivariate correlation analysis, which we discussed earlier. But for collinearity involving three or more predictors, other approaches are needed. One approach is to calculate the variance inflation Factor (VIF) for each predictor. Let's see how VIF is calculated.



Variance inflation factors can be calculated to assess collinearity within a model and determine which predictors, if any, are involved in the problem. Collinearity does not involve the response variable Y, only the predictors or X's. VIFs are calculated for each predictor used in the model. To start, regress the first predictor on all the other predictors in the original model. The R-square for this regression describes the proportion of the variability in X1 that can be explained by the remaining predictors. Take this R-square and plug it into the VIF equation, the reciprocal of 1 minus R-square. If the VIF is greater than or equal to 10, a collinearity problem is present involving X1 that should be corrected. This would correspond to an R-square of 0.9. Repeat this process for each X in the original model.

Removing Collinearity

To explore this learning resource in greater detail, please visit the course page on learn.sas.com. There, you'll find comprehensive course materials, tutorials, and updates to support your learning journey.



Question: Collinearity

Which of the following assumptions does collinearity violate?

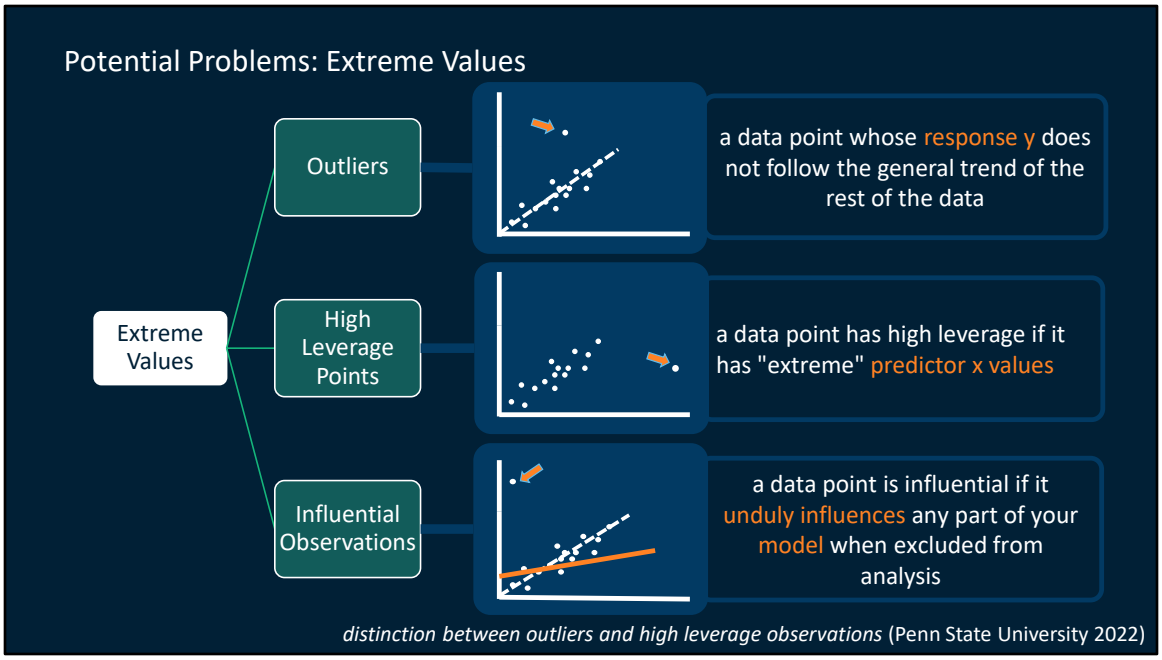
a. Independent errors
b. Constant variance
c. Normally distributed errors
d. None of the above

Answer: Collinearity

Which of the following assumptions does collinearity violate?

a. Independent errors
b. Constant variance
c. Normally distributed errors
d. None of the above

The correct answer is d.
Collinearity is not the violation of regression assumptions. It can occur in multiple regression only when two or more predictors are correlated.



Extreme values can also be problematic for regression models. Extreme values can be categorized as outliers, high leverage points, and influential observations.

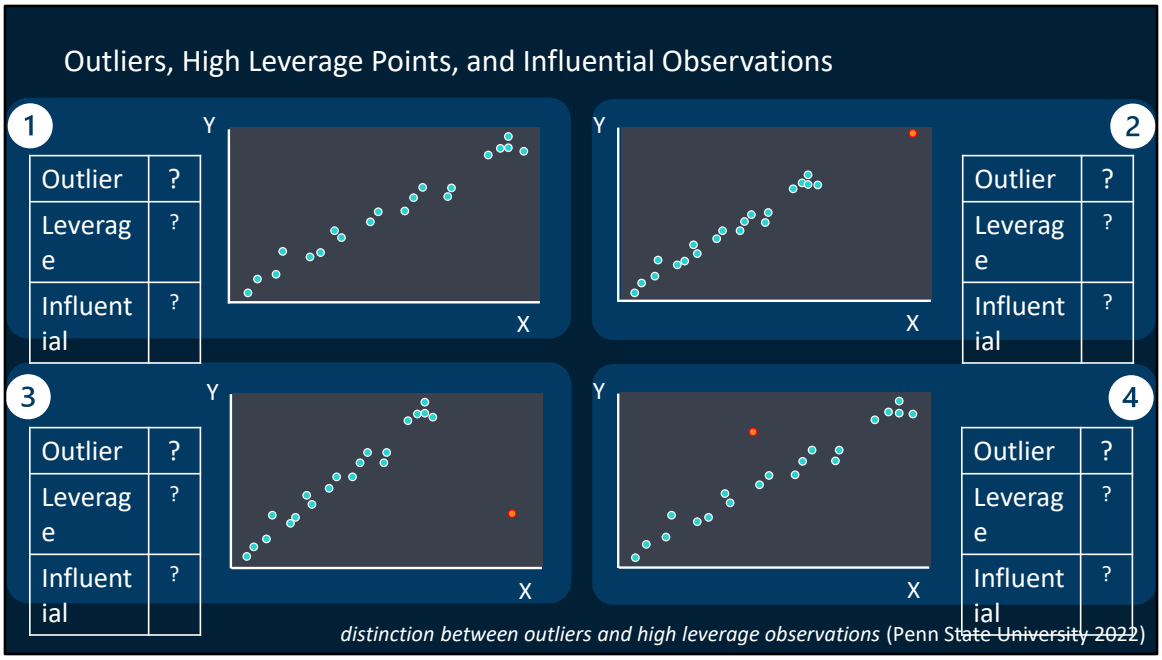
An *outlier* is a data point whose response y does not follow the general trend of the rest of the data.

A data point has *high leverage* if it has "extreme" predictor x values. With a single predictor, an extreme x value is simply one that is particularly high or low.

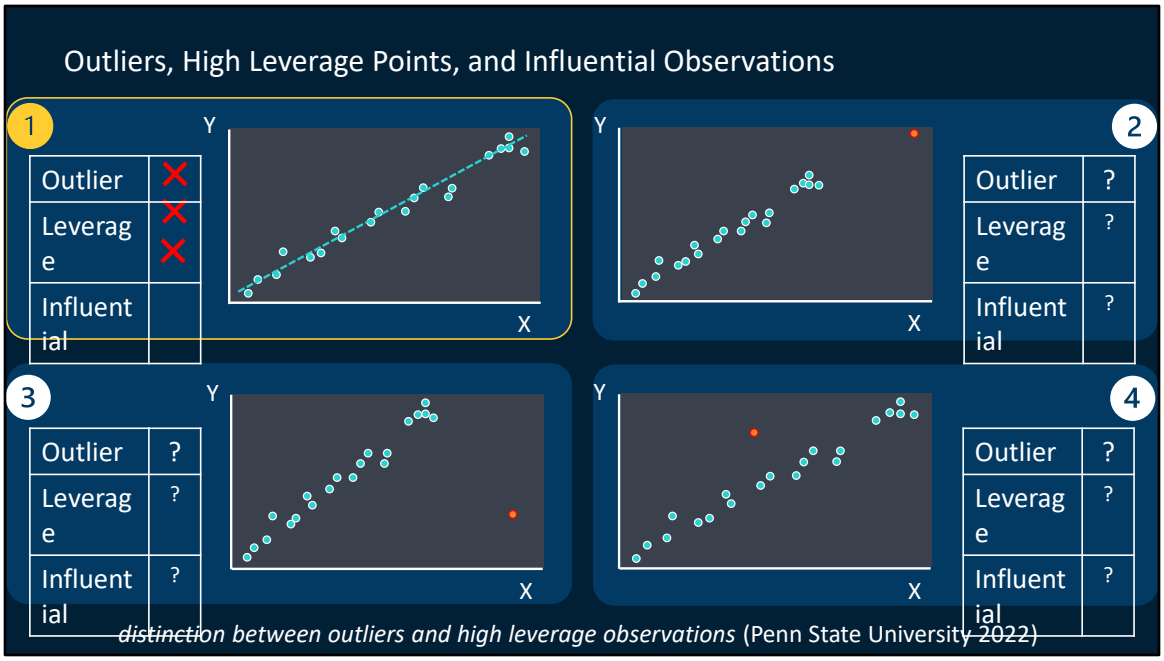
Note that — for our purposes — we consider a data point to be an outlier only if it is extreme with respect to the other y values, not the x values.

A data point is *influential* if it unduly influences any part of a regression analysis, such as the predicted responses, the estimated slope coefficients, or the hypothesis test results. Outliers and high leverage data points have the potential to be influential, but we generally must investigate further to determine whether they are influential.

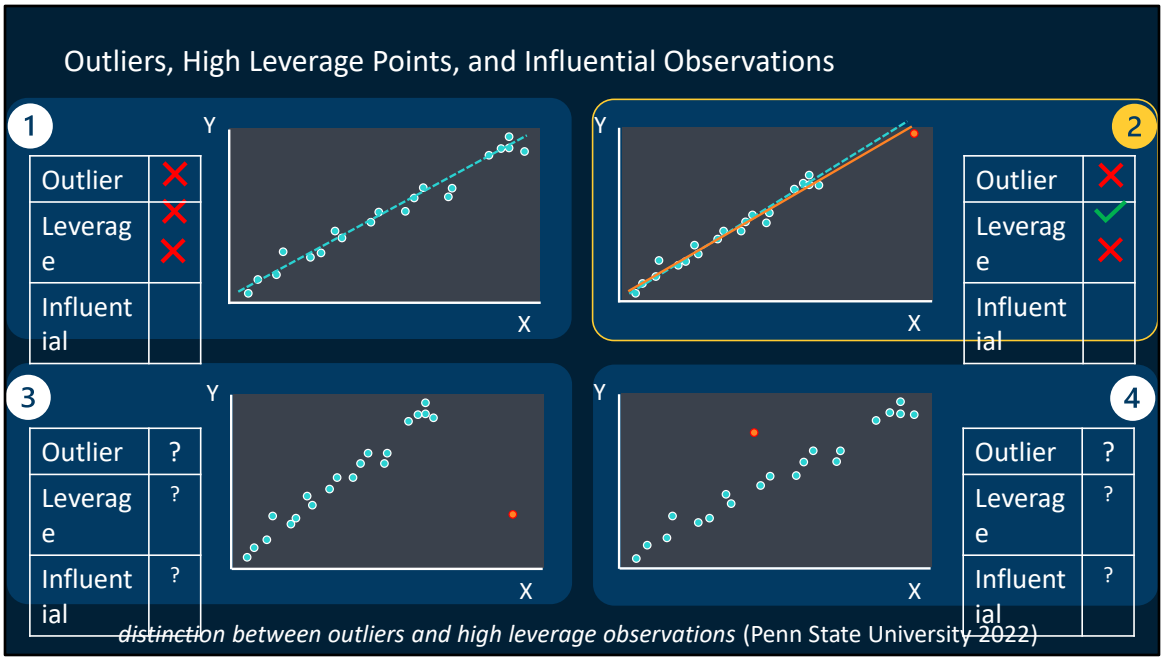
The presence of outliers or influential observations does not violate the assumptions of linear regression directly. But the unusual data points can affect the linearity assumption or distributional assumptions of the model.



Based on the definitions just now discussed, do you think the four illustrations contain any outliers? Or any high leverage points or influential observations?

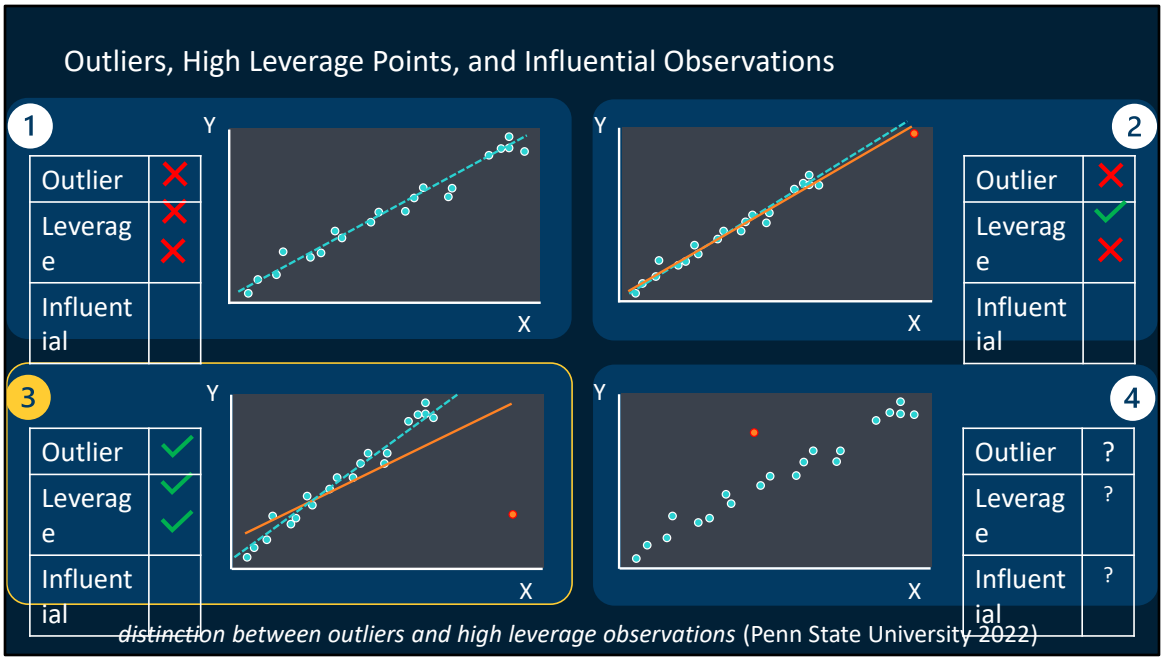


Example (1): All the data points follow the general trend of the rest of the data, so there are no outliers (in the y direction). And none of the data points are extreme with respect to x, so there are no high leverage points. Overall, none of the data points would appear to be influential with respect to the location of the best fitting line.



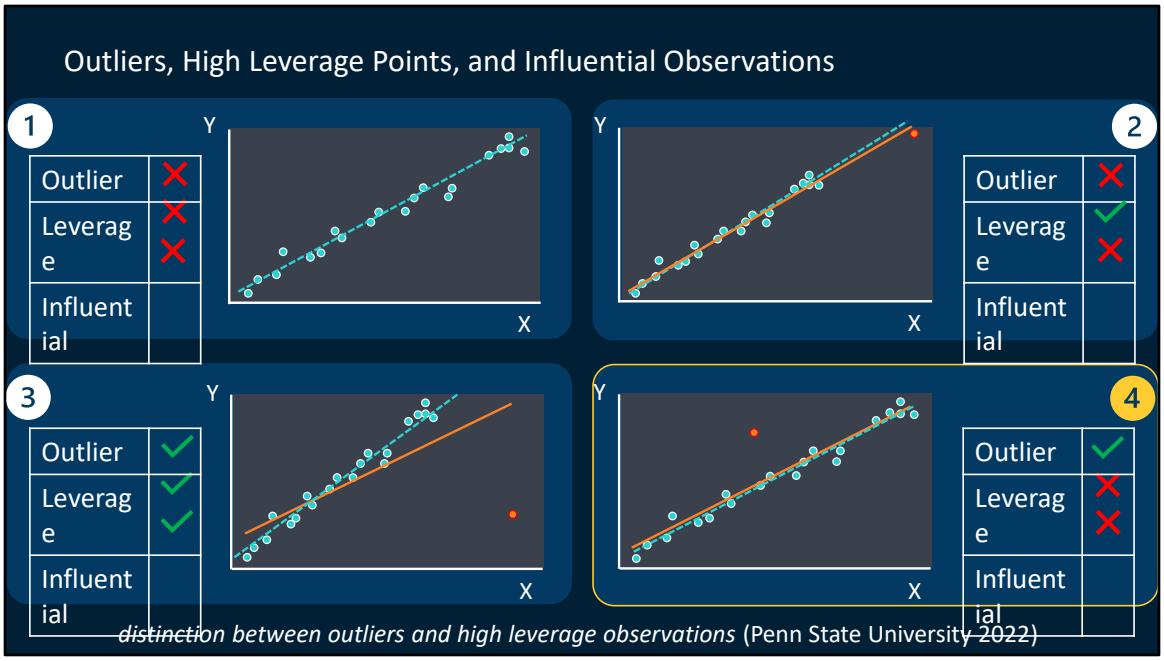
distinction between outliers and high leverage observations (Penn State University 2022)

Example (2): In this case, the orange data point does follow the general trend of the rest of the data. Therefore, it is not deemed an outlier here. However, this point does have an extreme x value, so it does have high leverage. Is the orange data point influential? An easy way to determine whether the data point is influential is to find the best fitting line twice — once with the orange data point included and once with the orange data point excluded. It's hard to even tell the two estimated regression lines apart! The solid line represents the estimated regression equation with the orange dot included, and the dashed line represents the estimated regression equation with the orange dot taken out. The slopes of the two lines are very similar, so it's not an influential observation.



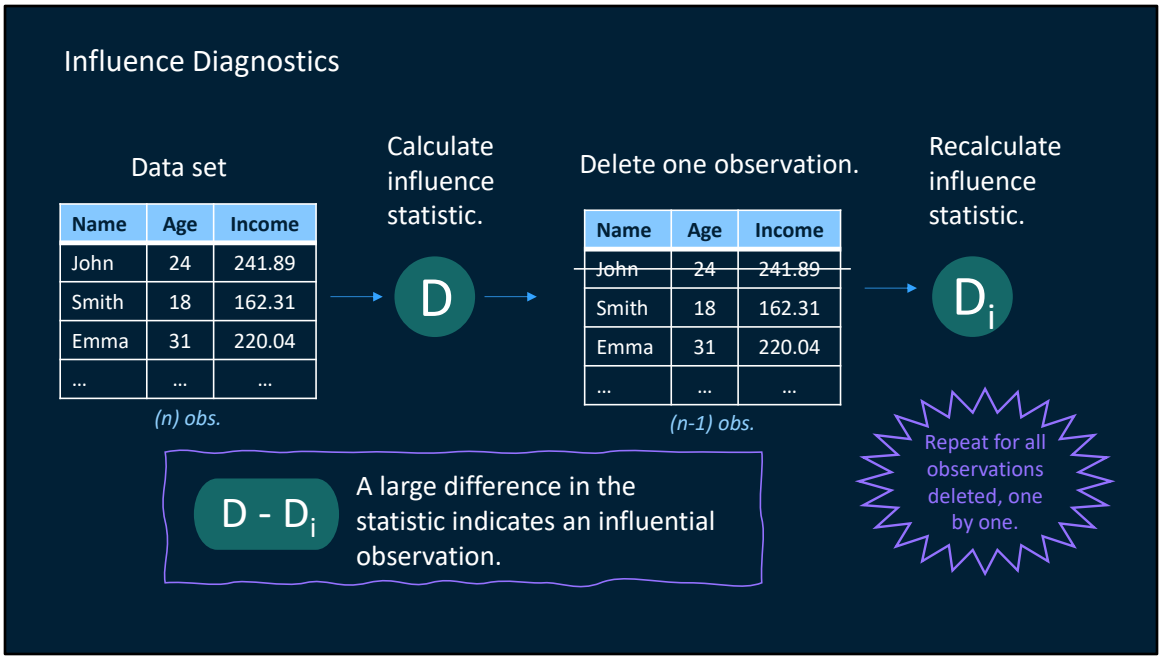
distinction between outliers and high leverage observations (Penn State University 2022)

Example (3): In this case, the orange data point is most certainly an outlier and has high leverage! The orange data point does not follow the general trend of the rest of the data, and it also has an extreme x value. And, in this case the orange data point is influential. The two best fitting lines, one obtained when the orange data point is included and one obtained when the orange data point is excluded, are substantially different. The existence of the orange data point significantly reduces the slope of the regression line.

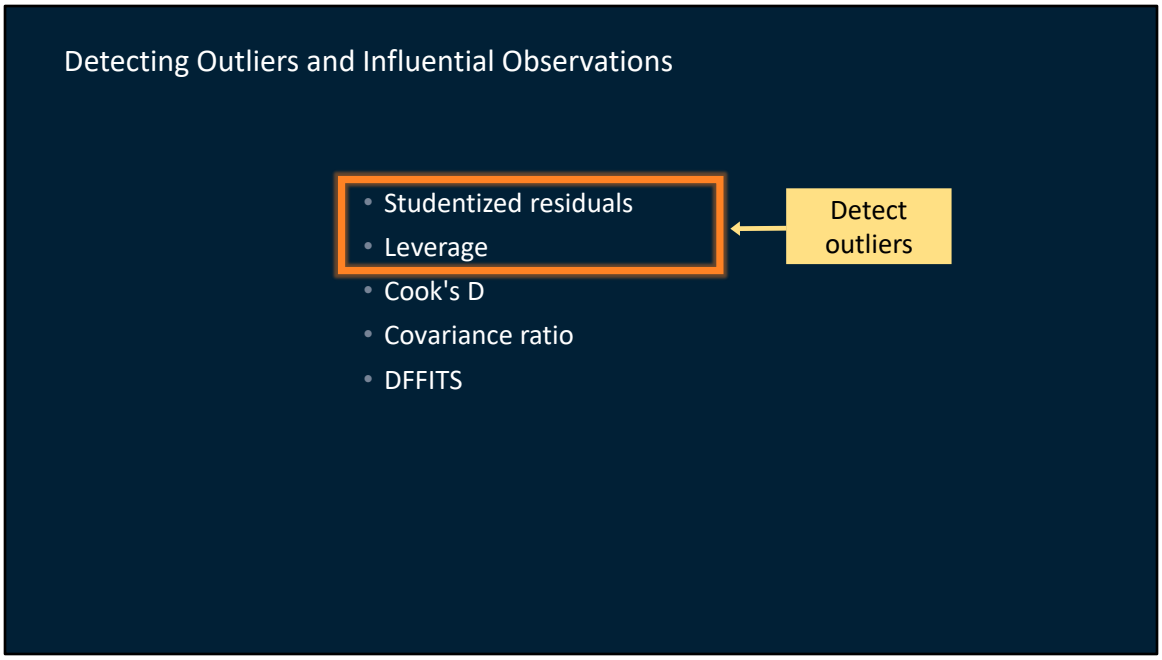


distinction between outliers and high leverage observations (Penn State University 2022)

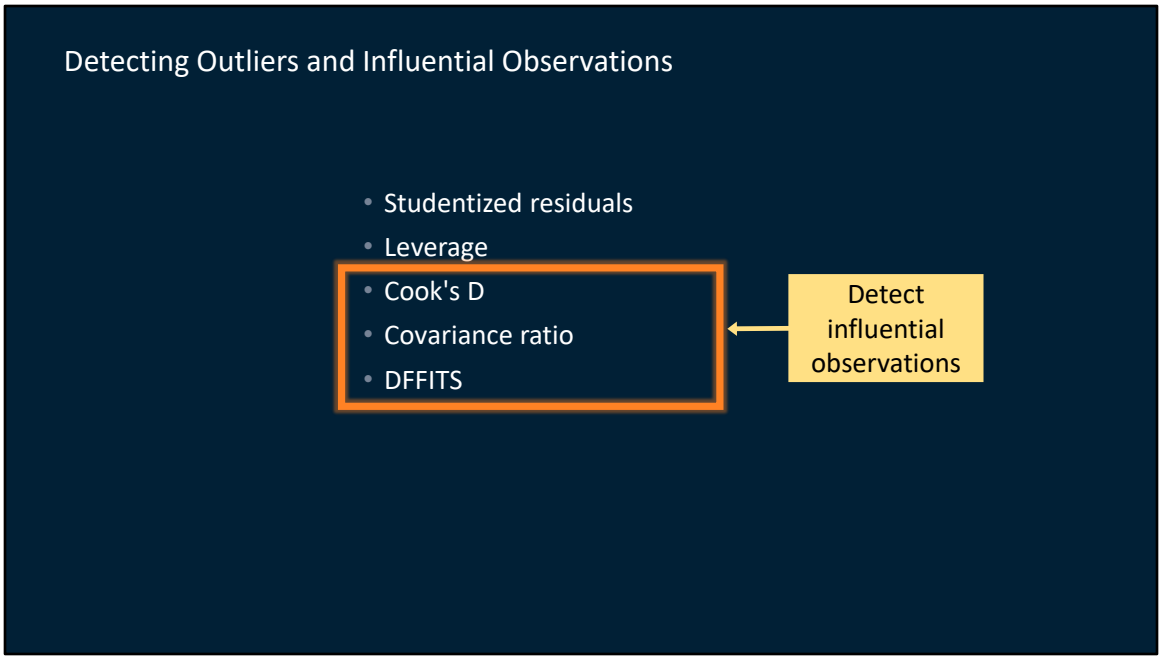
Example (4): Because the orange data point does not follow the general trend of the rest of the data, it would be considered an outlier. However, this point does not have an extreme x value, so it does not have high leverage. It's hard to even tell the two estimated regression lines apart! So, it's not an influential observation.



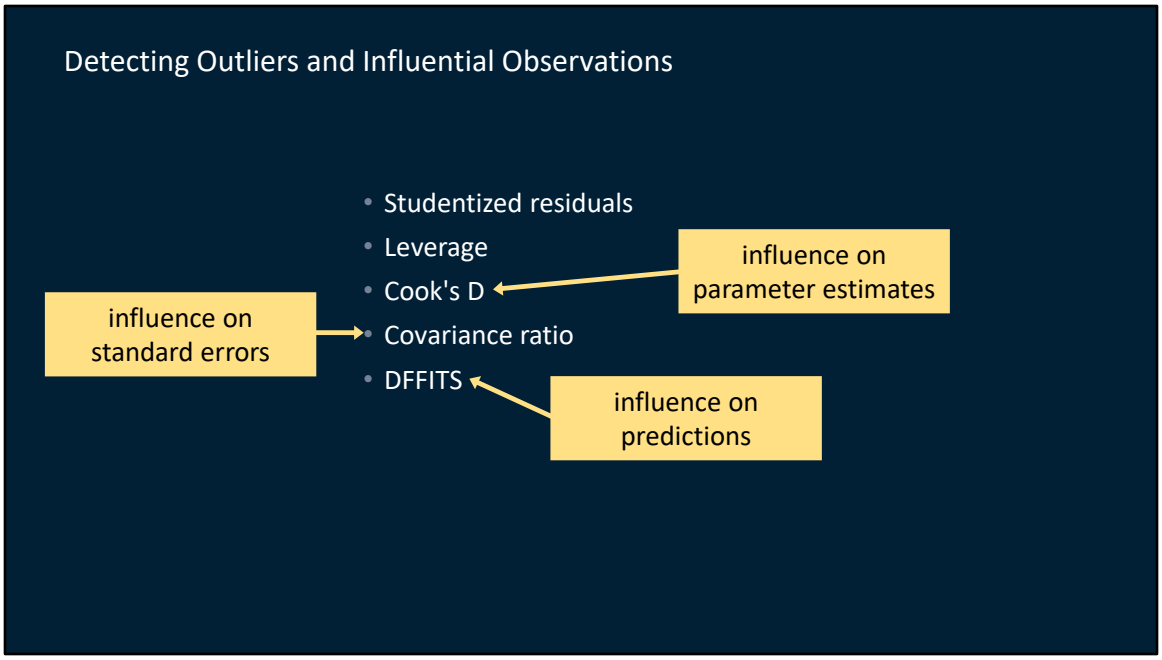
How do we decide whether an observation is influential? Most of the influence statistics take the same approach. They take your data set and calculate an influence statistic of interest. This statistic could measure any of several different parts of the model. Then an observation is deleted, and the statistic is recalculated. A difference between the statistic before and after the deletion is calculated. In some cases, the result is scaled in units of standard deviations. If the change is substantial, then that observation is considered influential. Repeat this process for each observation deleted, one by one, to determine whether or not that observation is influential.



Many influence diagnostics exist to help determine which observations are influential. Studentized residuals and leverage statistics help detect outliers.



Cook's D, covariance ratios, and DFFITS help find observations that influence various aspects of the model.



Cook's D, Covariance Ratios, and DFFITS help find observations that influence parameter estimates, their standard errors, and predicted values, respectively.

For linear regression models, there are suggested cutoffs for how large a statistic is needed to conclude that an observation is considered "influential."

Another approach to determining how big is big enough is to look for observations with large values of the influence statistics that are well separated from other values. Several advanced types of regression models such as generalized linear models and mixed models have no suggested cutoffs for influence statistics, so the subjective approach is required.

Detecting Outliers and Influential Observations

- Studentized residuals
- Leverage
- Cook's D
- Covariance ratio
- DFFITS

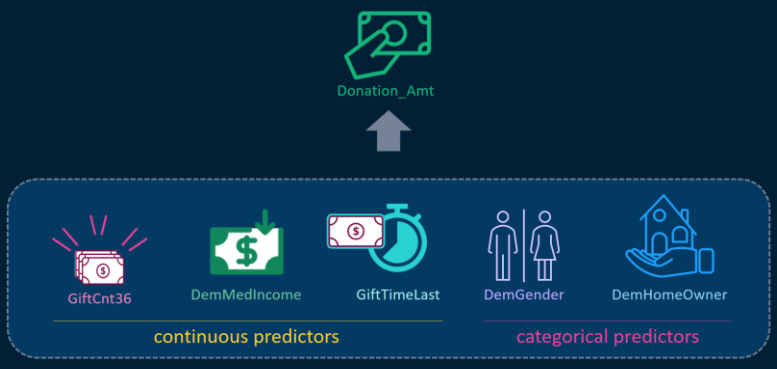
Dealing with outliers and influential observations depends on research goals.

Machine learning models are generally robust to outliers.

What to do with outliers depends on the research goals. Options for dealing with outliers are discussed in Lesson 5. These influence diagnostics are not necessarily used for big data analysis and machine learning modeling. That's because many machine learning models such as neural networks can easily handle outliers without compromising model predictions. This is due to both the flexibility of these models and having access to larger data sets.

Machine learning methods for outlier detection are discussed in a later lesson. For a more in-depth treatment of outlier detection for machine learning models, see the Advanced Machine Learning Using SAS Viya course.

Assessing the Model

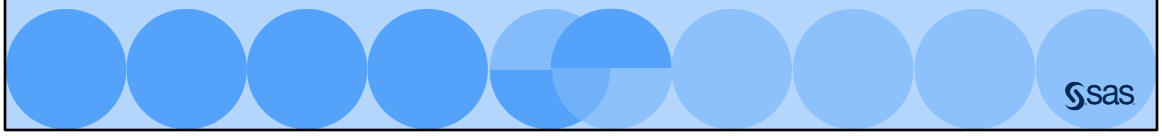


- Were the regression assumption met?
- Is there collinearity among predictors?
- Are there influential observations or outliers?

In the next demonstration, we assess the multiple regression model fit previously. Assessment includes addressing whether the model met the assumptions of linear regression, determining whether collinearity exists among the predictors, and finding any influential observations or outliers.

Demo: Model Diagnostics

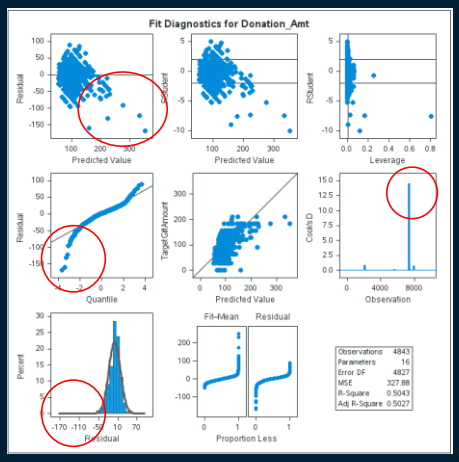
To explore this learning resource in greater detail, please visit the course page on learn.sas.com. There, you'll find comprehensive course materials, tutorials, and updates to support your learning journey.



What Did We Discover about the Model?

negative residuals, which could mean a misspecified model

skewed distribution of residuals



At least one influential observation

R-square = 0.50

Next steps?

- Include additional predictors or interactions.
- Transform predictors.
- Try other models.

Here is the diagnostics plots panel from the previous demonstration. What did we discover about this model? The residuals by predicted plot shows that higher predicted values over about $170 all have negative residuals. In other words, the model ***overestimates*** the donation amount for higher donations. Consistently overestimating the response variable suggests a misspecified model, and we might be missing important predictors or interactions from this model.

The quantile-quantile plot and the histogram of the residuals shows that the residuals are left-skewed. They are not normally distributed. In addition, based on the Cook's D plot, there is at least one observation that has high influence on the parameter estimates from this model.

Despite these issues, the model is able to explain half of the variability in donations given to the PVA.

What if we want to improve this model? What are the next steps that could be taken? Maybe including additional predictors or interactions could improve the model fit.

Transforming predictors is another option for improving the model.

Instead of trying to improve this model, different models can be explored. This model was based on stepwise selection using AIC to add and remove predictors. Typically, several models are fit and compared.

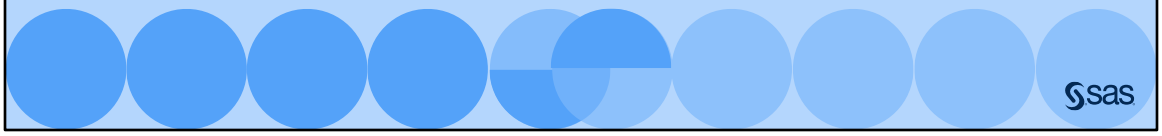
Lesson Summary

To explore this learning resource in greater detail, please visit the course page on learn.sas.com. There, you'll find comprehensive course materials, tutorials, and updates to support your learning journey.

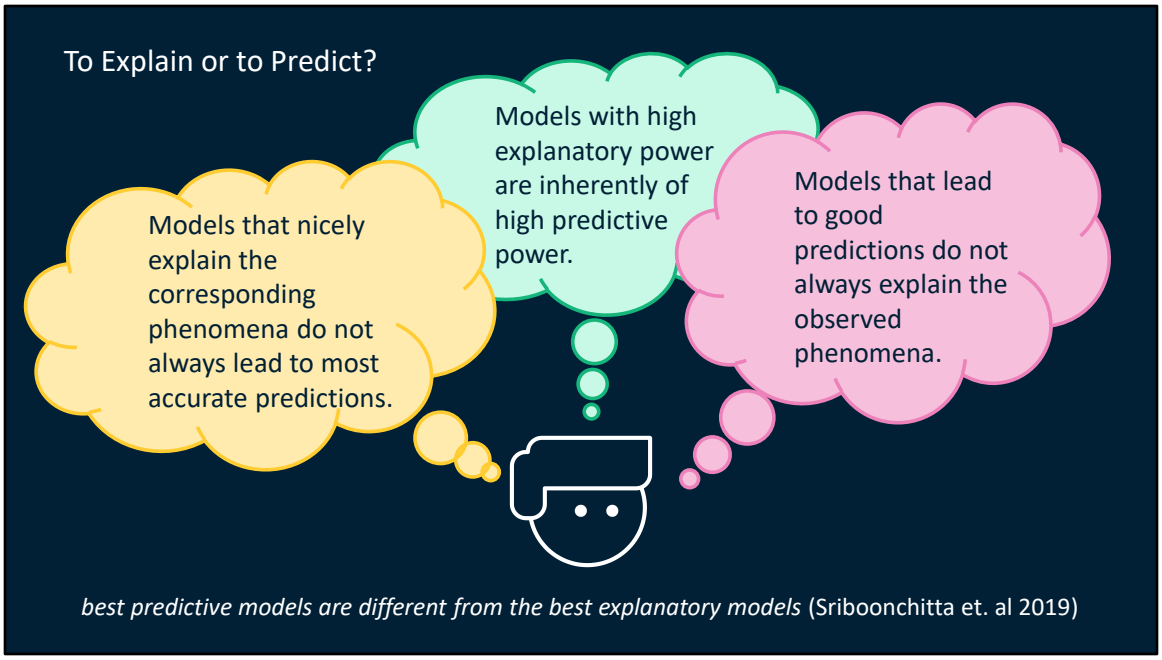
Lesson 4: Predictive Modeling Using Logistic Regression

Lesson Overview

To explore this learning resource in greater detail, please visit the course page on learn.sas.com. There, you'll find comprehensive course materials, tutorials, and updates to support your learning journey.

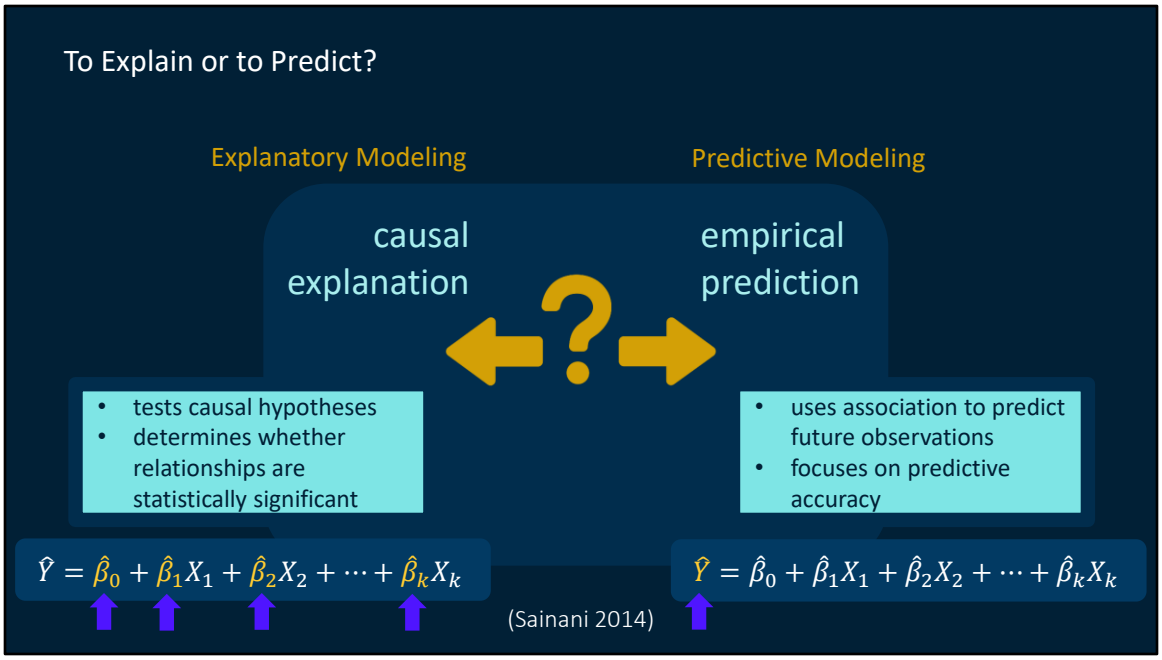


4.1 Introduction to Predictive Modeling



In statistics, it is implicitly assumed that models that are the best predictors also have the best explanatory power. Of course, an explanatory model can be used for prediction - and likewise, a predictive model can be used for explanation.

However, the best predictive models are often different from the best explanatory models. In practice, models that lead to good predictions do not always explain the observed phenomena. Vice versa, models that nicely explain the corresponding phenomena do not always lead to most accurate predictions.

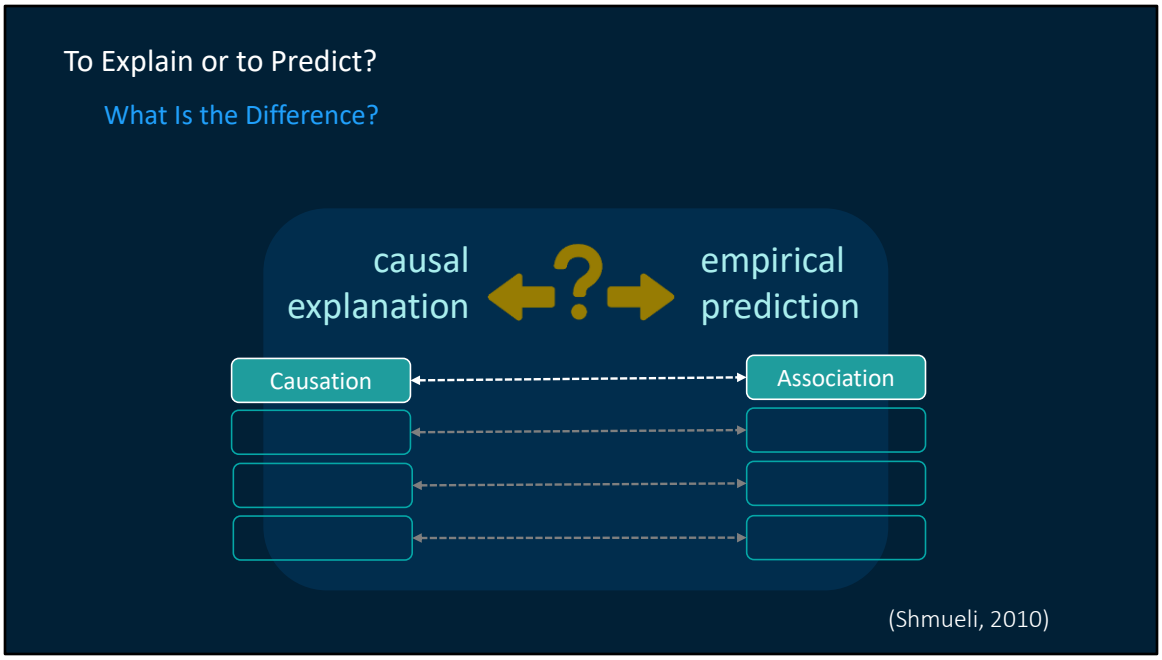


Earlier in this course, you learned to build models for the purpose of understanding and explaining the relationships between a response and a set of predictors. But model building works differently for purely predictive models. This distinction affects every aspect of model building and evaluation.

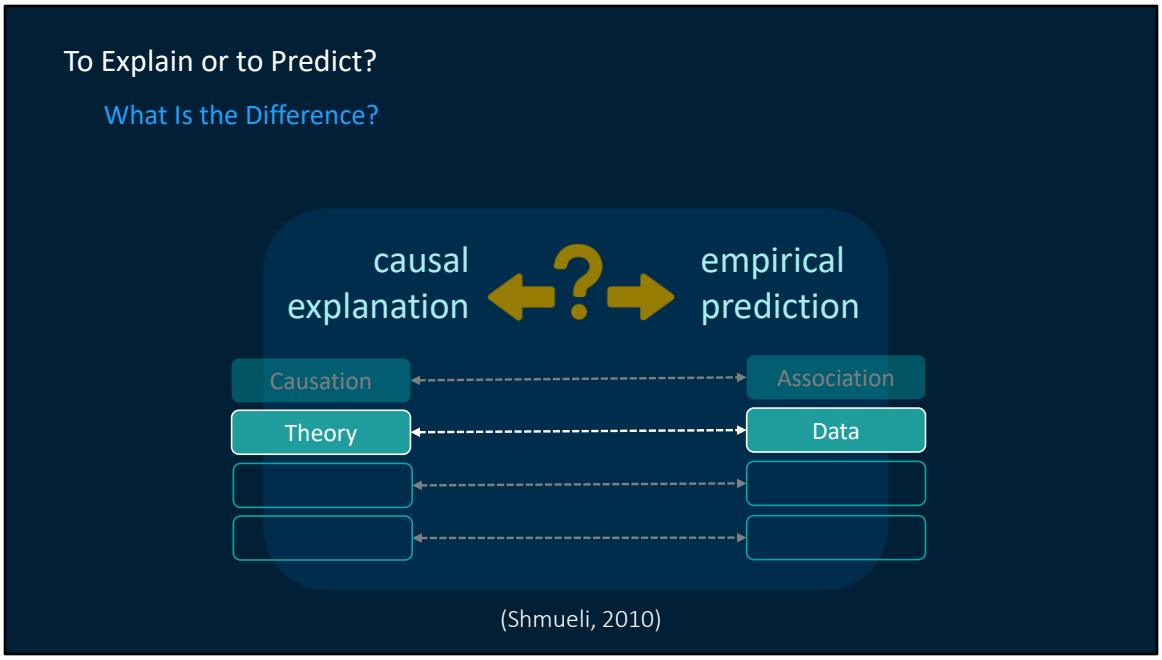
Explanatory modeling and predictive modeling have distinct goals that they are aimed at causal explanation and empirical prediction, respectively.

In explanatory modeling, statistical models are applied to data in order to test causal hypotheses. Thus, looking at the model equations, explanatory modeling focuses on the estimates of the beta coefficients.

In predictive modeling our interest is different. Predictive modeling is a process of applying a statistical model or data mining and machine learning algorithm to data for the purpose of predicting new or future observations. Here, the goal is to use the associations between predictors and the response to generate good predictions for future outcomes. As a result, predictive models are created very differently from explanatory models. The primary goal is predictive accuracy. Thus, the focus is not so much on the parameters of the model as it is in the predictions of observations. In the model equation, these are represented on the left side of the equation.



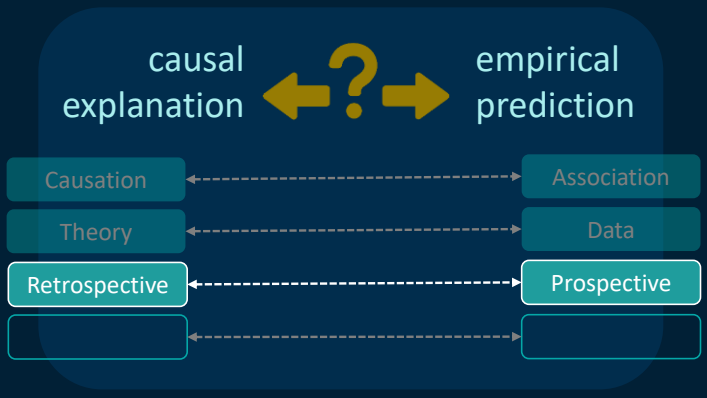
In explanatory modeling, the model represents an underlying causal function, and input is assumed to cause the target. In predictive modeling, the model captures the association between input and target.



In explanatory modeling, the model is carefully constructed based on a function in a fashion that supports interpreting the estimated relationship between input and target and testing the causal hypotheses. In predictive modeling, the model is often constructed from the data. Direct interpretability in terms of the relationship between input and target is not required, although sometimes transparency of the model is desirable.

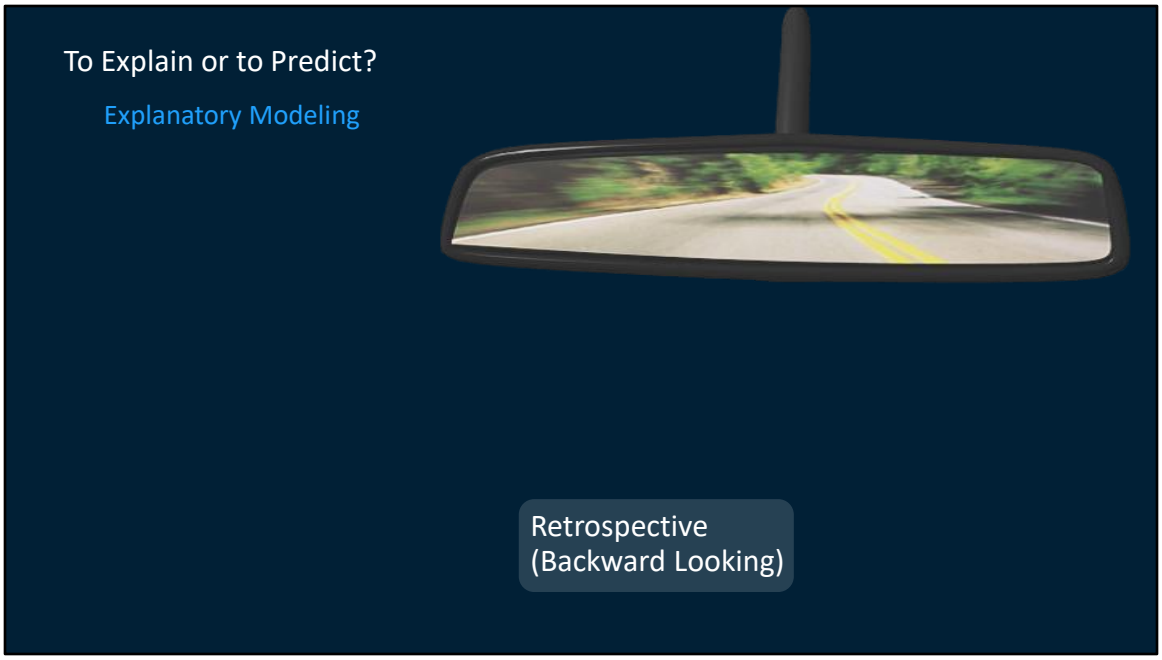
To Explain or to Predict?

What Is the Difference?



(Shmueli, 2010)

Predictive modeling is forward looking, in that the model is constructed for predicting new observations. In contrast, explanatory modeling is retrospective, in that the model is used to test an already existing set of hypotheses.

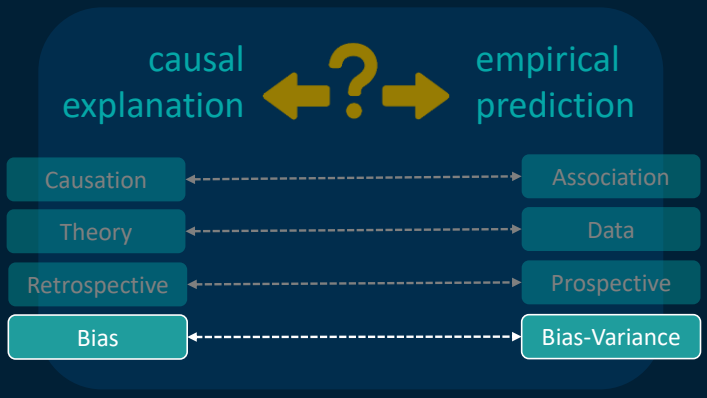


Explanatory modeling provides tools to analyze the *past*. It's retrospective, or looking backward to the past. Using explanatory modeling, you can act only on information that describes the *history*, and looking *forward* becomes a guess. So, solving your business problem based on historical data is like driving your car looking only at the rearview mirror.

In contrast, *predictive modeling* holds the promise of being able to make actionable predictions of customer and market behavior based on historical data. Predictive modeling is prospective or looking forward into the *future* - like seeing the truck coming toward you and being able to develop strategies to avoid the crash while also keeping eye on the rearview mirror.

To Explain or to Predict?

What Is the Difference?



(Shmueli 2010)

In explanatory modeling, the focus is on minimizing bias to obtain the most accurate representation of the underlying theory. In contrast, predictive modeling seeks to minimize the combination of bias and estimation variance, occasionally sacrificing theoretical accuracy for improved empirical precision.

We will discuss this bias-variance trade-off in detail in the context of predictive modeling.

Question: Explanation versus Prediction

Using prior usage of services and demographic inputs, a telecom company wants to determine which customers are likely to churn. They built a regression model, which of the following statements is correct?

(Select all that apply)

a. This is essentially an explanatory modeling problem.
b. This is essentially a predictive modeling problem.
c. Perhaps the same model can be used to explain why customers were churned as well as to predict which customers likely to churn.
d. None of the above.

Answer: Explanation versus Prediction

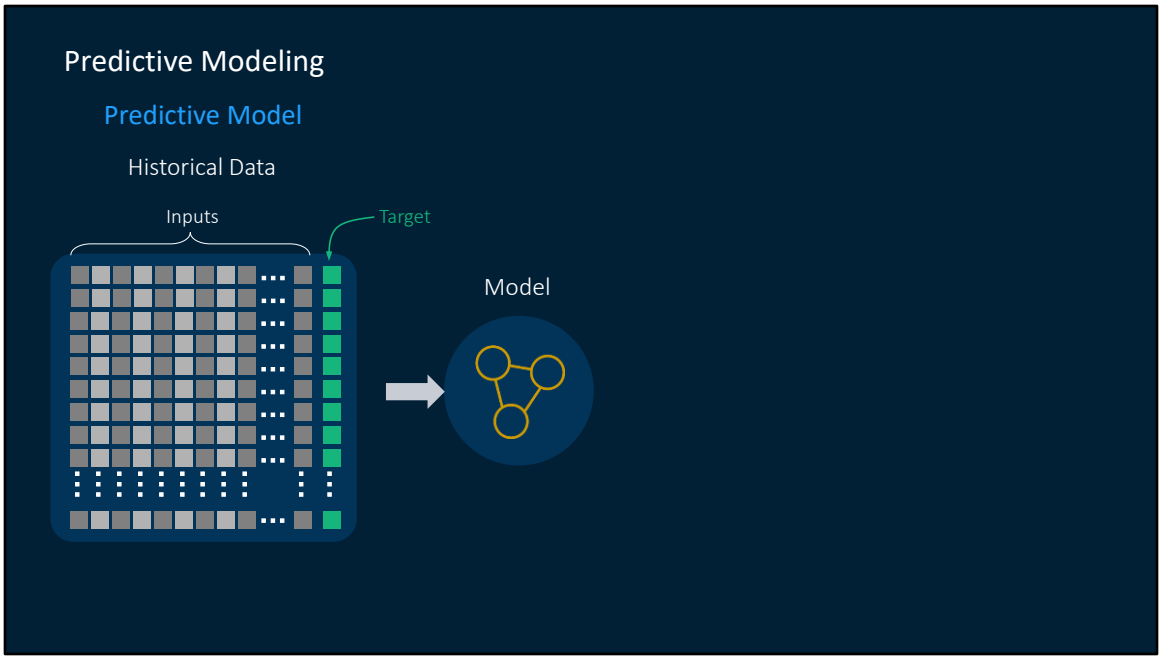
Using prior usage of services and demographic inputs, a telecom company wants to determine which customers are likely to churn. They built a regression model, which of the following statements is correct?

(Select all that apply)

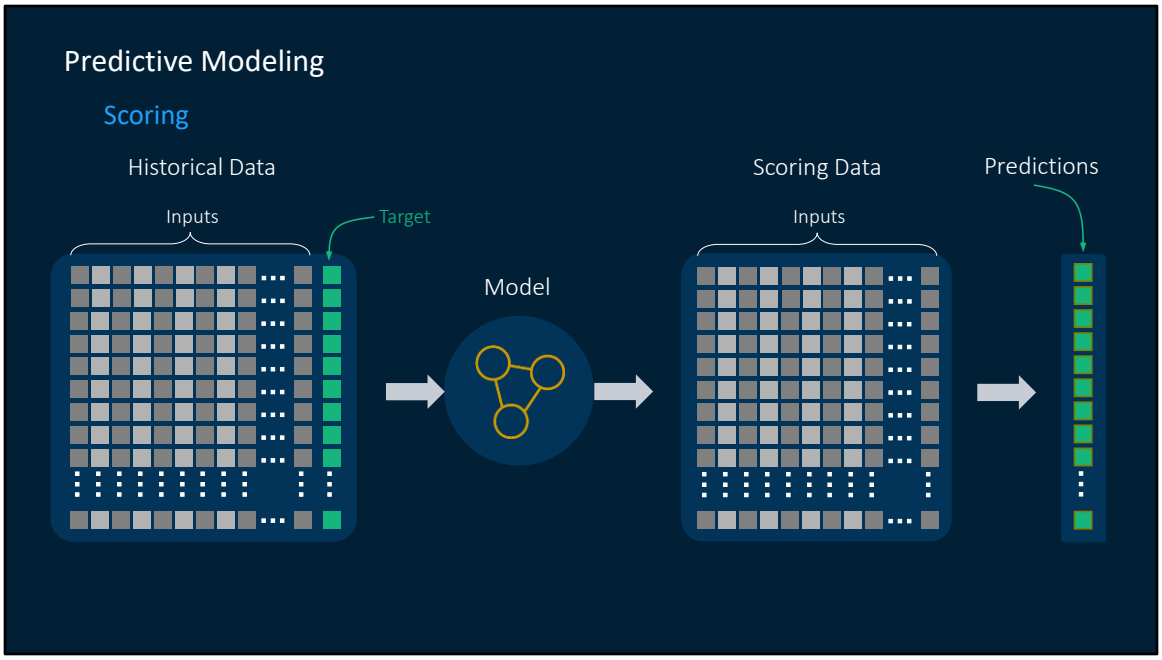
a. This is essentially an explanatory modeling problem.
b. This is essentially a predictive modeling problem.
c. Perhaps the same model can be used to explain why customers were churned as well as to predict which customers likely to churn.
d. None of the above.

ANSWER: b and c

Predicting churn is fundamentally a predictive modeling problem. The distinction between using regression for an explanation and prediction is somewhat artificial. A model developed for prediction is probably a good explanatory model. Conversely, a model developed for an explanation is probably a good prediction model.

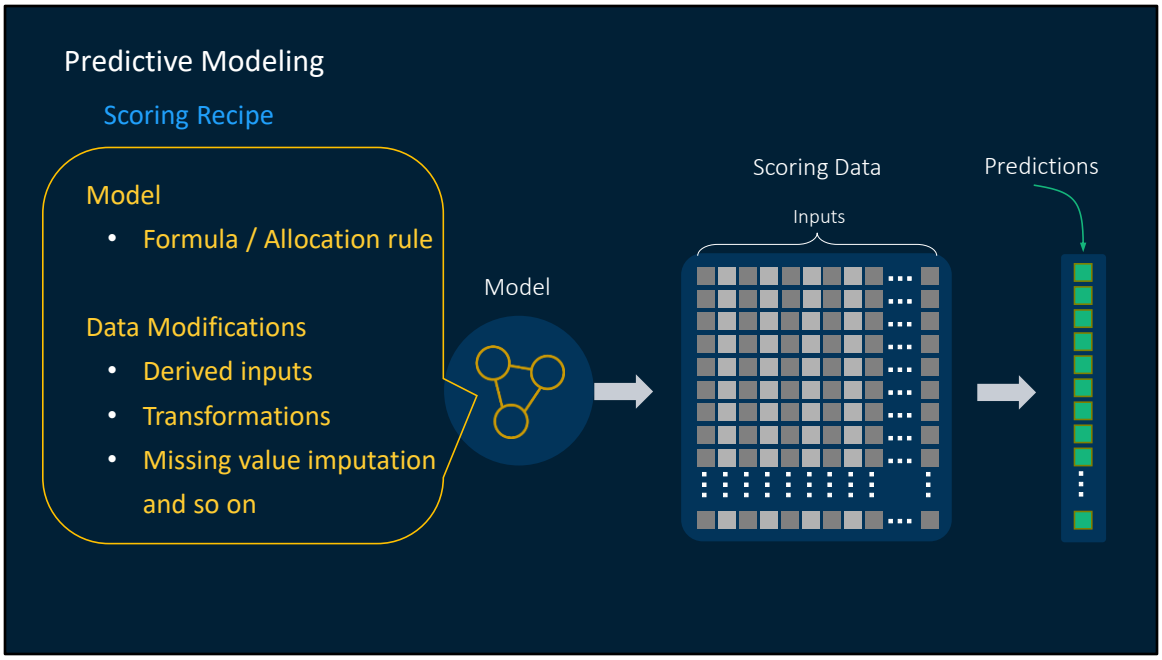


The historical data is used to construct a model (allocation rule) that can predict the values of the target from the inputs. The *predictive model* is a concise representation of the association between the inputs and the target. The task is referred to as *supervised* because the prediction model is constructed from data where the target is known.

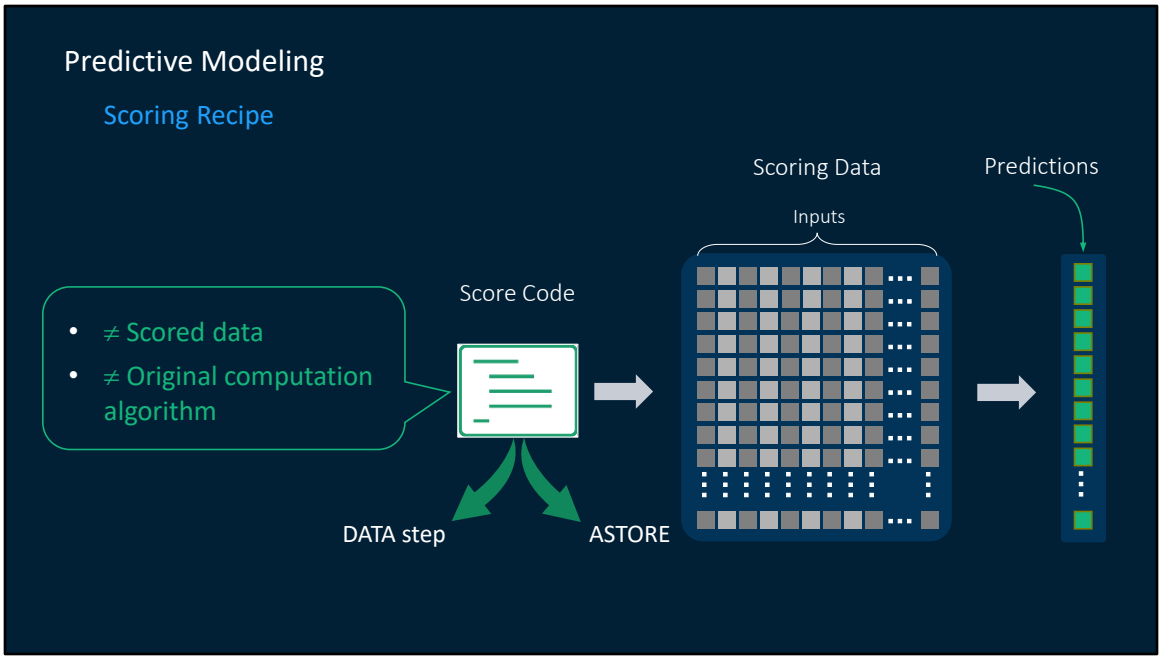


If the targets are known, why do we need a prediction model? It allows you to predict *new* cases when the target is unknown. Typically, the target is unknown because it refers to a future event. In addition, the target may be difficult, expensive, or destructive to measure.

The predictive modeling task is not completed when a model and allocation rule are determined. The model must be practically applied to new cases. This process is called *scoring*. Scoring is the generation of predicted values for a data set that might not contain a target variable. The model is applied to the scoring data set to obtain predicted outcomes. Based on the predictions from the model, the enterprise makes business decisions and acts.



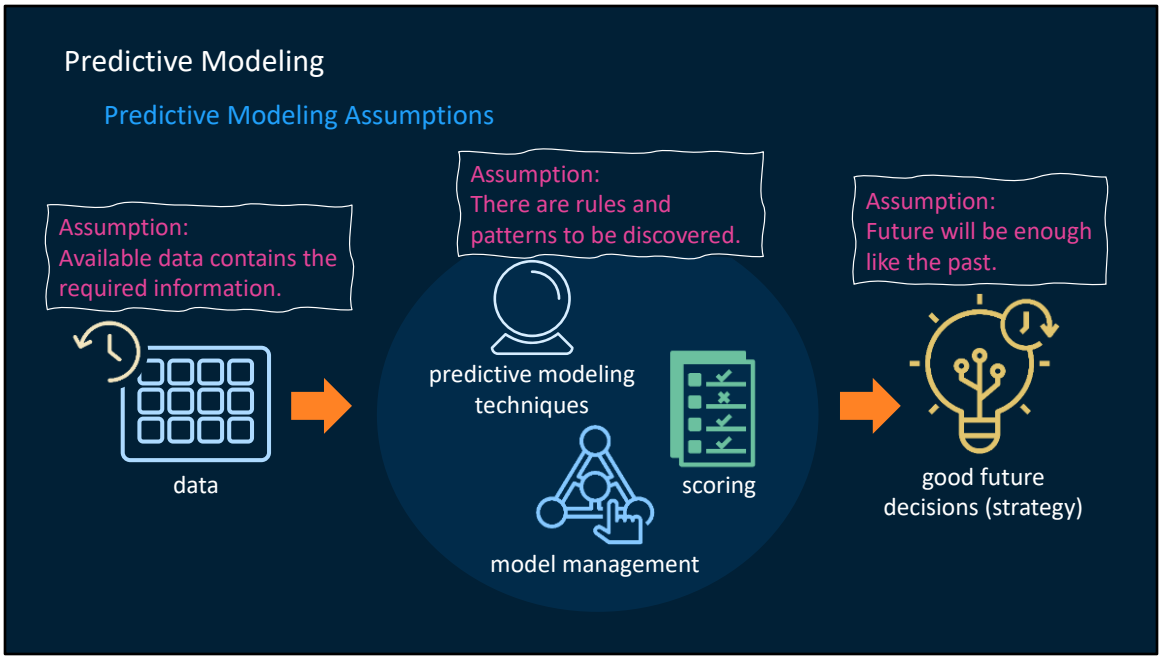
The scoring process is sometimes referred to as *model deployment*, *model production*, or *model implementation*. All tasks performed earlier in the analytics life cycle led to this task: generating predictions through scoring. Therefore, scoring must be more than just applying the model equation or allocation rules. It must incorporate all data manipulation tasks done before generating the model like feature engineering, transformations, missing value imputation, and so on.



For scoring, the model is first translated into another format typically called *score code*. In the SAS Viya environment, score code is a SAS program that you can easily run on a scoring data set. Then the model is applied to the scoring data set to obtain predicted outcomes generally known as *scored data*. Based on the predictions from the model, the enterprise makes business decisions and acts.

Thus, the score code be neither confused with the scored data nor it is equivalent to the original computation algorithm.

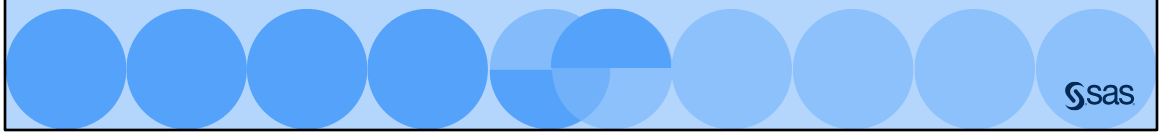
SAS Viya creates two types of SAS language score code: either DATA step code or analytic store (ASTORE) code. An ASTORE file is a useful, transportable, universal, and compact binary file that represents the state of an analytic procedure after training.



All data comes from the past. Predictive modeling techniques, paired with scoring and good model management, enable you to use your data about the past and the present to make good decisions for the future. This process assumes several things. The most important assumption is that the future will be enough like the past that lessons learned from past data will remain applicable in the future. Another important assumption is that there are rules and patterns to be discovered and that the available data contains the required information.

Analytical Methods for Predictive Modeling

To explore this learning resource in greater detail, please visit the course page on learn.sas.com. There, you'll find comprehensive course materials, tutorials, and updates to support your learning journey.



Question: Score Code in SAS Viya

Which of the following statements is true regarding score code in SAS Viya? (Select all that apply)

a. Score code is SAS code that implements all modeling tasks including data preprocessing.
b. The score code is strictly equivalent to the original modeling algorithm.
c. Score code creates prediction estimates.
d. You can use the score code to score a data set.

Answer: Score Code in SAS Viya

Which of the following statements is true regarding score code in SAS Viya? (Select all that apply)

a. Score code is SAS code that implements all modeling tasks including data preprocessing.

b. The score code is strictly equivalent to the original modeling algorithm.

c. Score code creates prediction estimates.

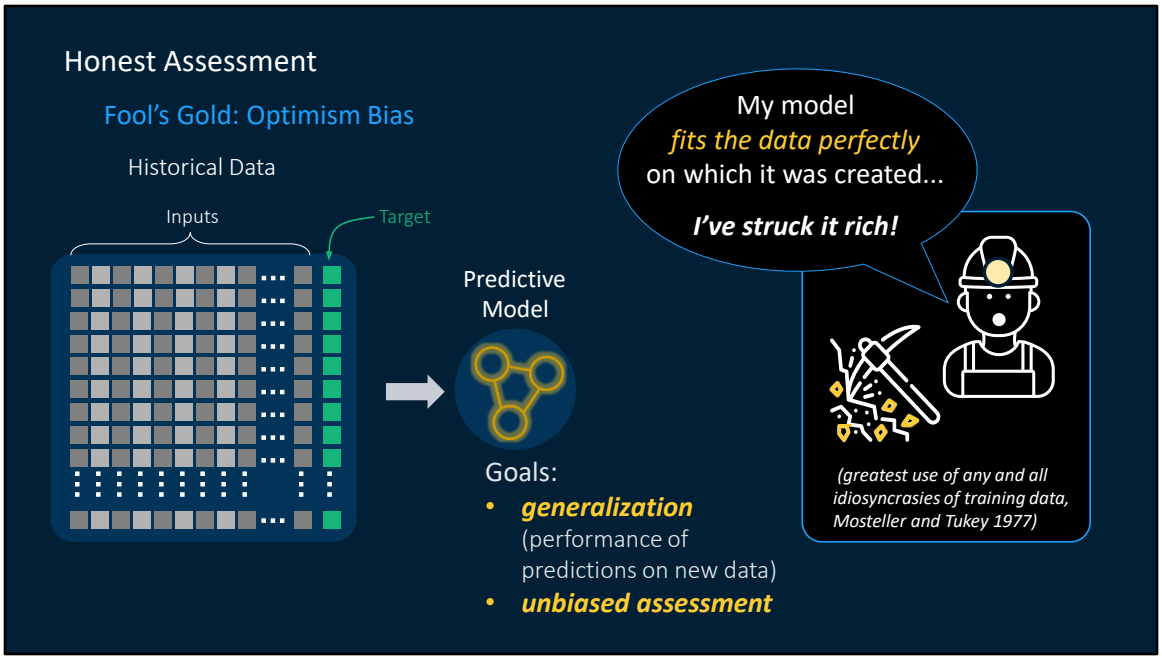
d. You can use the score code to score a data set.

ANSWER: a, c and d.

Score code incorporates all data manipulation tasks done before generating the model like feature engineering, transformations, and missing value imputation. SAS Viya creates two types of SAS language score code: either DATA step code or analytic store (ASTORE) code. Score code can be used for scoring a new data set.
Note that SAS Studio does not create any path analytic flow, hence data preprocessing code is not included in the score code coming out of a modeling task.

Timeframes for Modeling

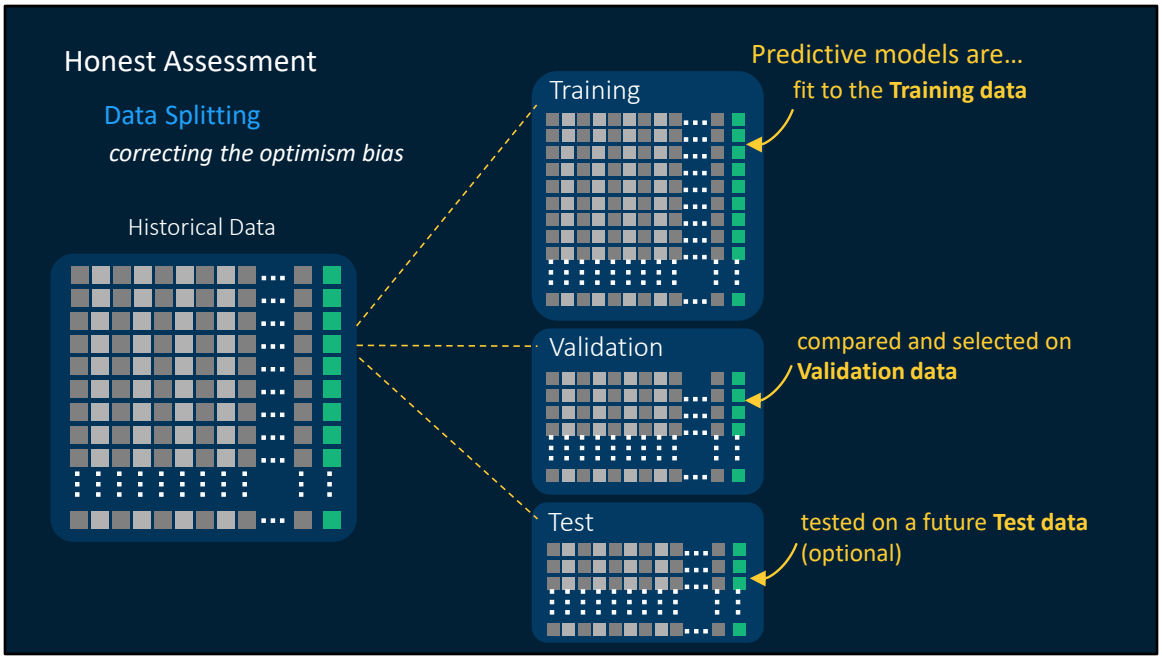
To explore this learning resource in greater detail, please visit the course page on learn.sas.com. There, you'll find comprehensive course materials, tutorials, and updates to support your learning journey.



When you fit a predictive model on a set of historical data, you might think it's a good idea to assess the model using the same data that you used to fit the model. However, when you assess the accuracy of a predictive model on the same data that was used to fit the model, you tend to get better assessment statistics than when you assess the model on other data. This bias is known as the *optimism bias*.

The purpose of predictive modeling is *generalization*, which is the performance of the predictions on new data. Evaluating the model on the same data the model was fit on usually leads to an optimistically biased assessment.

Testing the procedure on the data that gave it birth is almost certain to overestimate performance, for the optimizing process that chose it from among many possible procedures will have made the greatest use of any and all idiosyncrasies of those particular data.

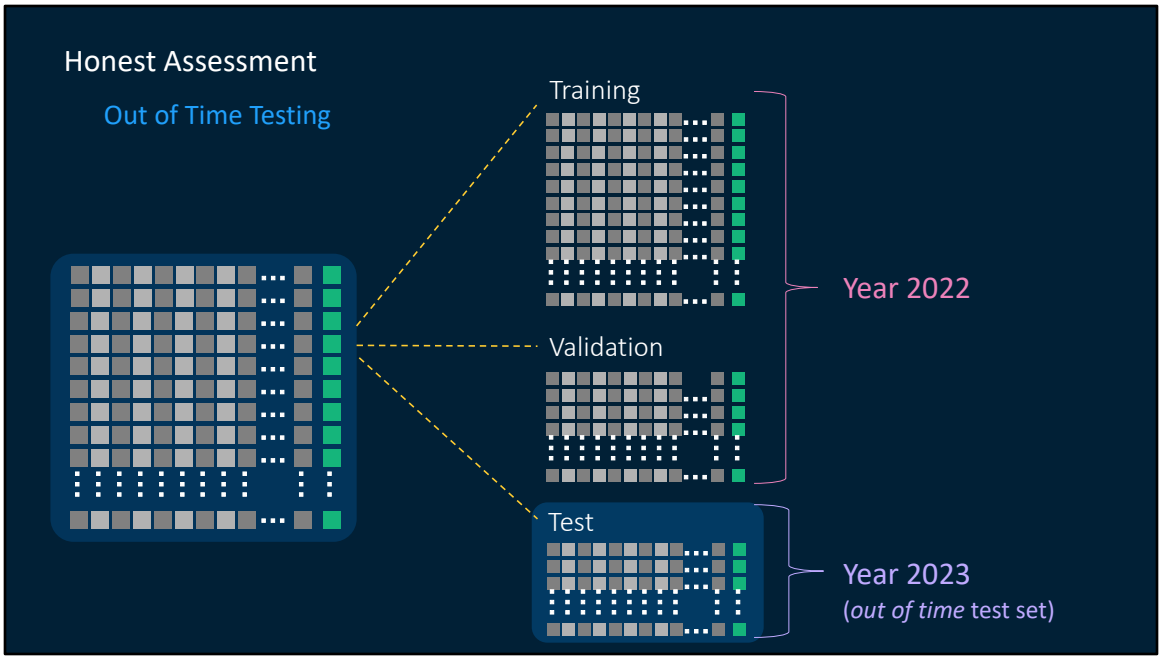


To avoid an optimistically biased assessment and create a predictive model that generalizes well, you need to assess the performance of the model on new data that was not used to fit the original model. This approach is called *honest assessment*. There are various ways to perform an honest assessment. In this lesson, you learn the simple method of splitting the data. Later in the course, you also learn about a few other methods of performing an honest assessment: k-fold cross validation and bootstrapping.

In data splitting, a portion of the data is used to fit the model and the rest is held out for empirical validation. A major portion is used for fitting the model - the training data set.

The *validation data set* is used for monitoring and tuning the model to improve its generalization. The tuning process usually involves selecting among models of different types and complexities. The tuning process optimizes the selected model on the validation data. Consequently, a further holdout sample is needed for a final, unbiased assessment.

The *test data set* has only one use: to give a final honest estimate of generalization. Consequently, cases in the test set must be treated just as new data would be treated. They cannot be involved whatsoever in the determination of the fitted prediction model. In some applications, there might be no need for a final honest assessment of generalization. A model can be optimized for performance on the test set by tuning it on the validation set. It might be enough to know that the prediction model will likely give the best generalization possible without being able to say what it is. In this situation, no test set is needed.



For predictive models, the test set should come from a different time period than the training and validation sets. The proof of a model's stability is in its ability to perform well month after month. A test set from a different time period, often called an *out of time* test set, is a good way to verify model stability, although such a test set is not always available.

Question: Honest Assessment

Which of the following statements are true regarding honest assessment? (Select all that apply)

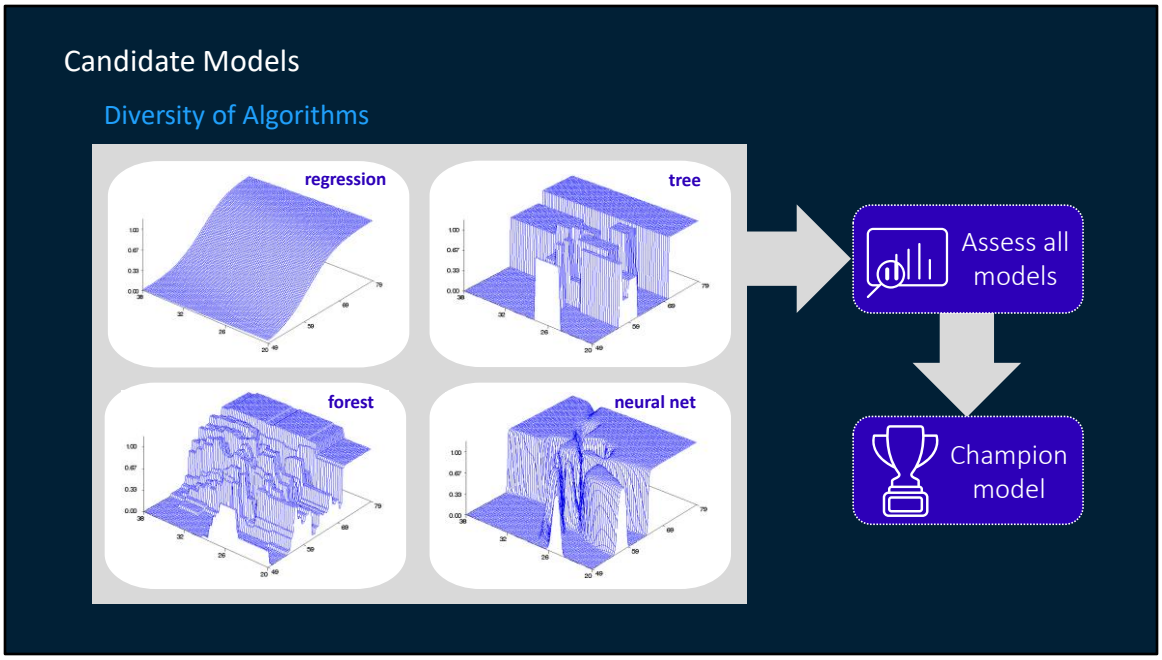
a. Training, validation, and test data are derived from historical data.
b. Test data are strictly larger than the training data.
c. Validation and test data do not necessarily contain the target variable.
d. The model is built on training data and assessed on validation or test data.

Answer: Honest Assessment

Which of the following statements are true regarding honest assessment? (Select all that apply)

a. Training, validation, and test data are derived from historical data.
b. Test data are strictly larger than the training data.
c. Validation and test data do not necessarily contain the target variable.
d. The model is built on training data and assessed on validation or test data.

Correct answer: a and d
To perform honest assessment by way of data splitting, you need to extract three subsets from the historical data: the training, validation, and test sets, all contain the target variable.
A major portion, called the *training data set*, is used for fitting the model. The *validation data set* is used for monitoring and tuning the model to improve its generalization. The *test data set* has only one use: to give a final honest estimate of generalization.

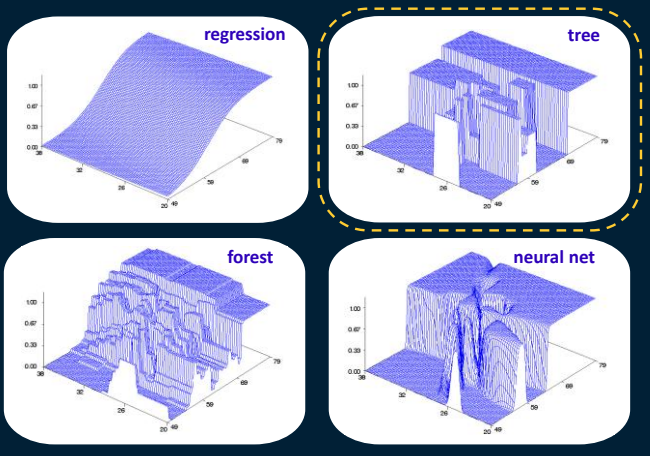


Because predictive modeling is mostly data driven, you cannot anticipate which model or algorithm works well for a given problem. Therefore, a common methodology includes building several candidate models. The models that you build can be based on diverse algorithms such as regression, decision trees, random forest, neural networks, and so on. You can see *prediction surface* plots showing how each of these four machine learning algorithms, for example, predicts a coarse grid across the input feature space.

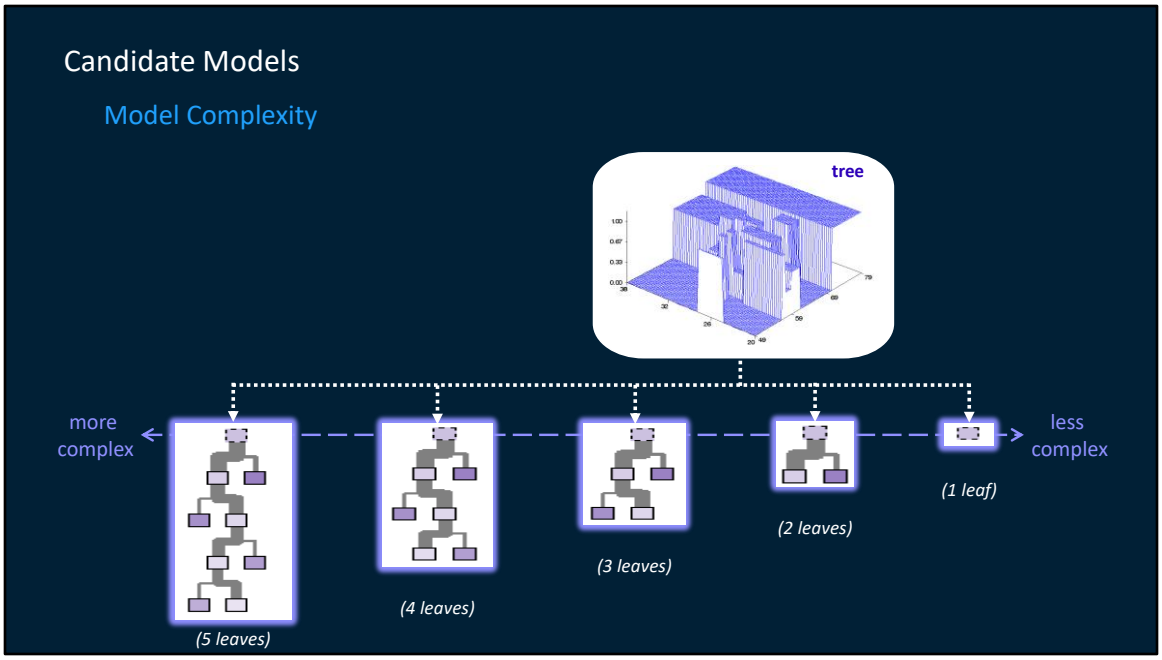
Each type of model has its own metrics by which it can be assessed, but there are also assessment tools that are independent of the type of model. By honest assessment, several statistics and statistical plots can be used to find out the best model called the champion model. The champion model is the best predictive model that is chosen from a pool of candidate models. Before you identify the champion model, you can evaluate the structure, performance, and resilience of candidate models.

Candidate Models

Diversity within an Algorithm



Predictive modeling typically involves choices from among a set of models. These might be different *types* of models.



Or they might be different *complexities* of models of the *same* type, such as multiple decision trees that use different splitting and pruning criteria from the same training data set. Fitting a model to data in fact requires searching through the space of possible models. Constructing a model with good generalization requires choosing the right complexity.

Question: Model Diversity

Because data mining and machine learning are data driven, it's always a good idea to create several candidate models and then choose the best from them.

a. True
b. False

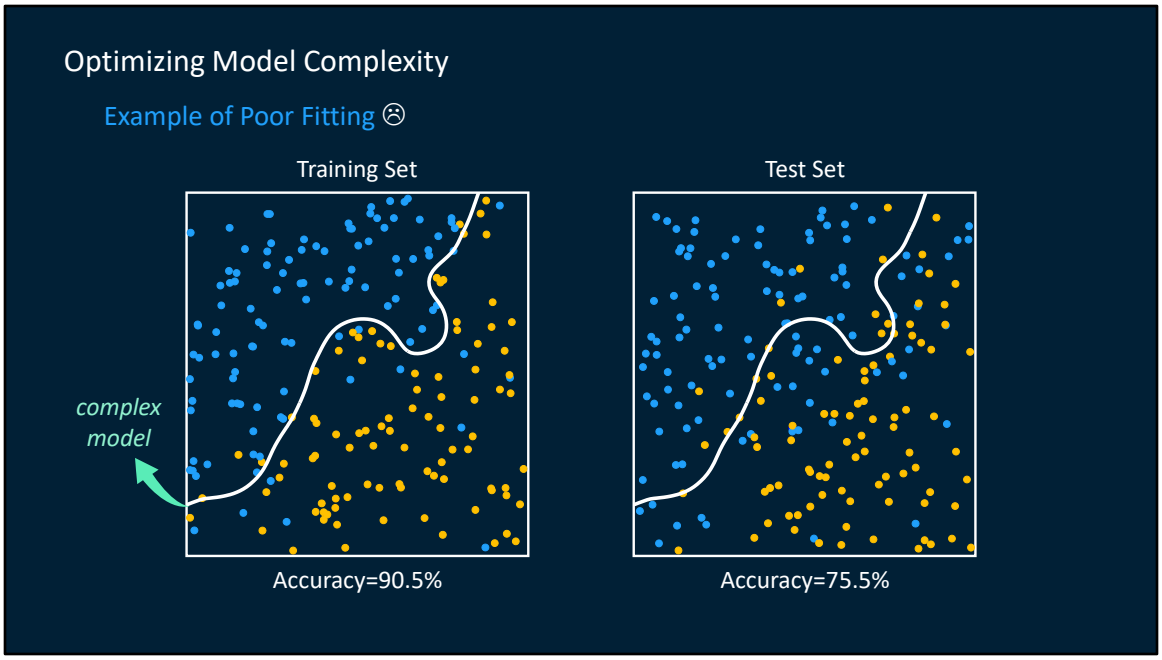
Answer: Model Diversity

Because data mining and machine learning are data driven, it's always a good idea to create several candidate models and then choose the best from them.

a. True
b. False

Correct answer: True

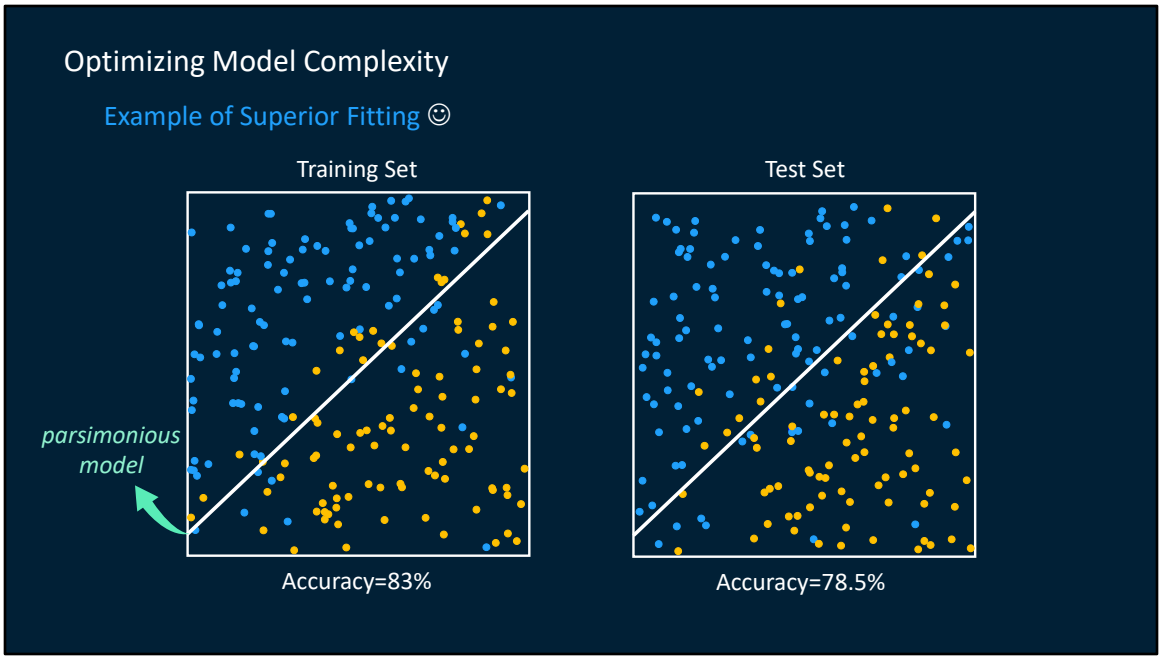
It's always a good idea to create several candidate models and then choose the best from them. These might be different types of models, or they might be different complexities of models of the same type.



A very complex model was used on the classification problem shown here, where the goal was to discriminate between the blue and yellow classes. The classifier fit the training data well, making only 19 errors among the 200 cases (90.5% accuracy).

However, on a test set of data, the classifier did not do as well, making 49 errors among 200 cases (75.5% accuracy). The flexible model snaked through the training data accommodating the signal (meaningful information) as well as the noise (random variation or fluctuation that interferes with the signal).

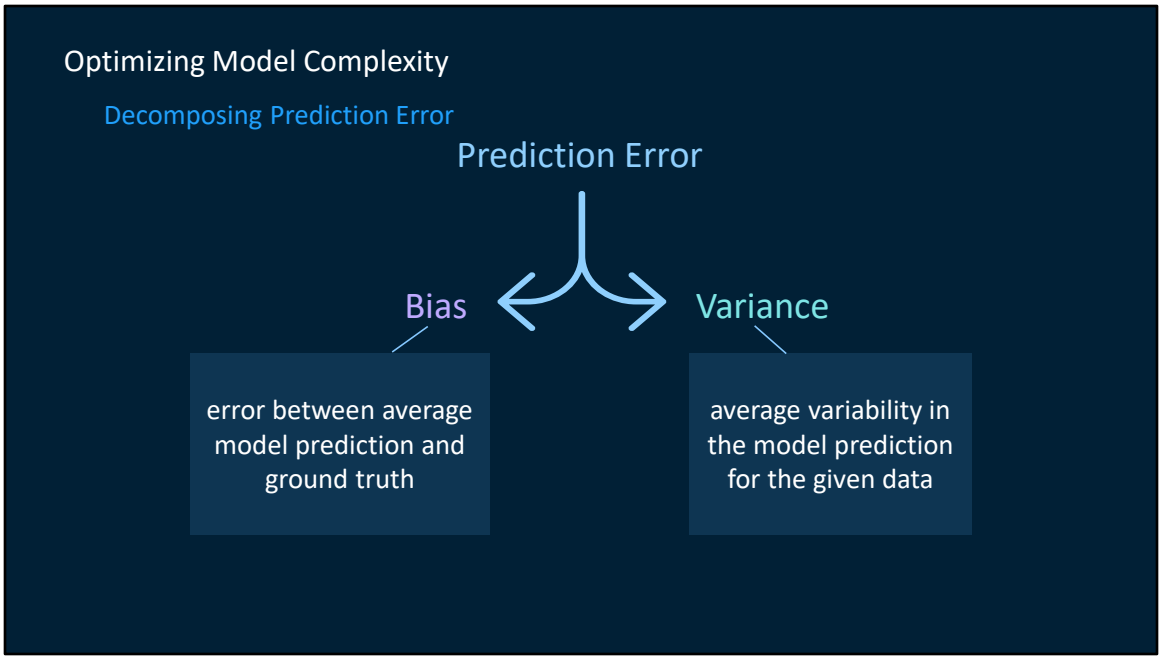
Signal and noise are covered in detail later in the course.



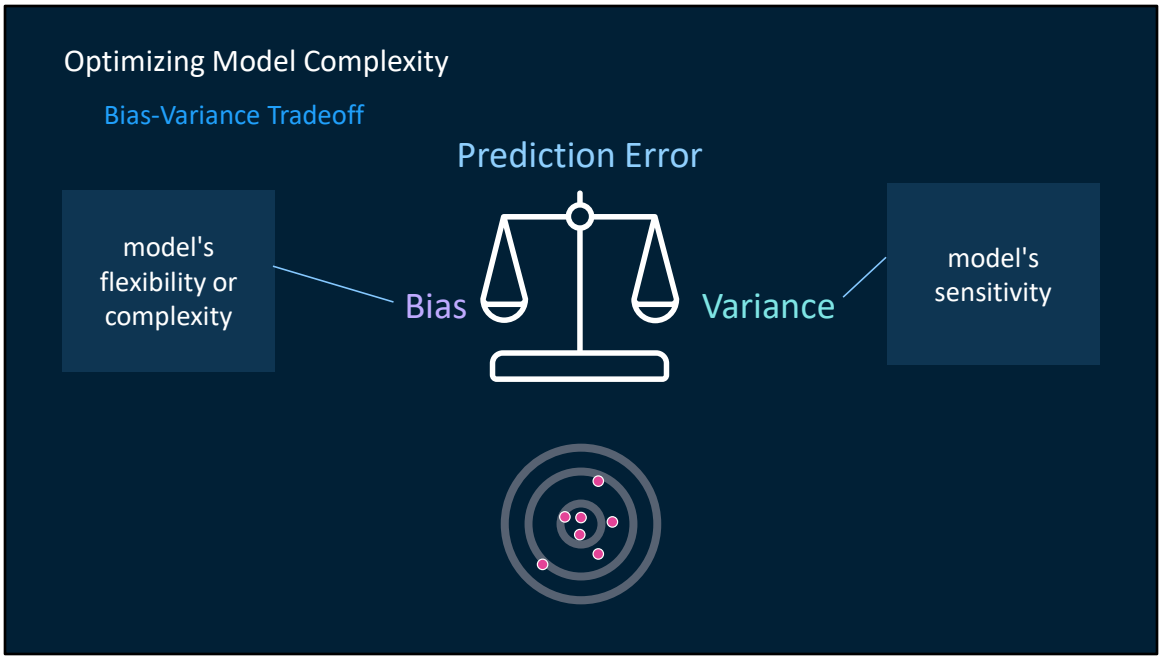
A more *parsimonious* model was fit to the training data. The apparent accuracy was not quite as impressive as the complex model (34 errors, 83% accuracy. However, it gave better performance on the test set (43 errors, 78.5% accuracy).

How would you decide which model to choose as best?

You can know that you picked the best model by assessing the prediction error. The prediction problem ignores questions of model correctness and focuses on the predicted values. Prediction error refers to the difference between the predicted values made by some model and the actual values.



This expected prediction error can largely be decomposed into bias and variance. Bias is the difference between the expected prediction values and the correct value of the response. Variance is the variability of the predictions due to the model variability.

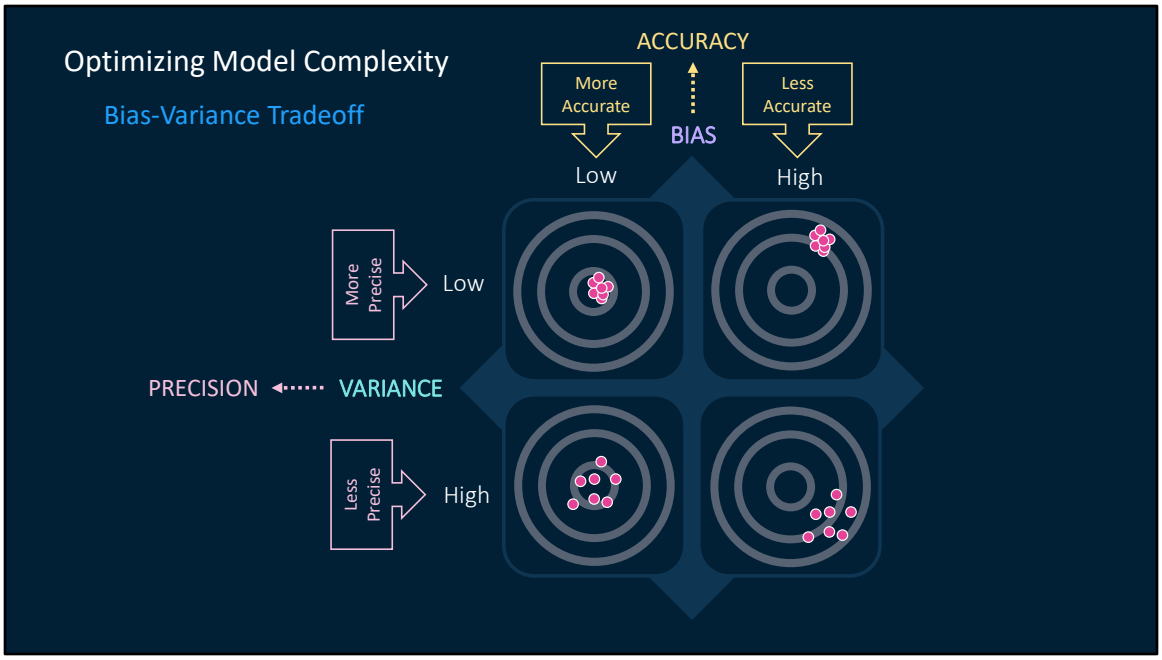


A bias occurs when an algorithm has limited flexibility to learn from data. Such models pay very little attention to the training data and oversimplify the model. The bias of the estimated function tells us the capacity of the underlying model to predict the values.

Variance defines the algorithm's sensitivity to specific sets of data. A model with a high variance pays a lot of attention to training data and overcomplicates the model that does not generalize. The variance of the estimated function tells you how much the function can adjust to the change in the data set.

Selecting model complexity involves a trade-off between bias and variance. Both bias and variance are forms of prediction error in machine learning. Thus, you select a model that simultaneously achieves *low variance* and *low bias*. But bias and variance are inversely related to each other. Trying to reduce one component, the other component of the model will increase. The true art lies in creating a good fit by balancing both.

Let's try to understand this trade-off using a simple example. Imagine that the center of the target, or the bull's-eye, perfectly predicts the correct value of your model. The dots marked on the target then represent an individual realization of your model based on your training data. In certain cases, the dots will be densely positioned close to the bull's-eye, ensuring that predictions made by the model are close to the actual data. In other cases, the training data will be scattered across the target. The more the dots deviate from the bull's-eye, the higher the bias and the less accurate the model will be in its overall predictive ability.



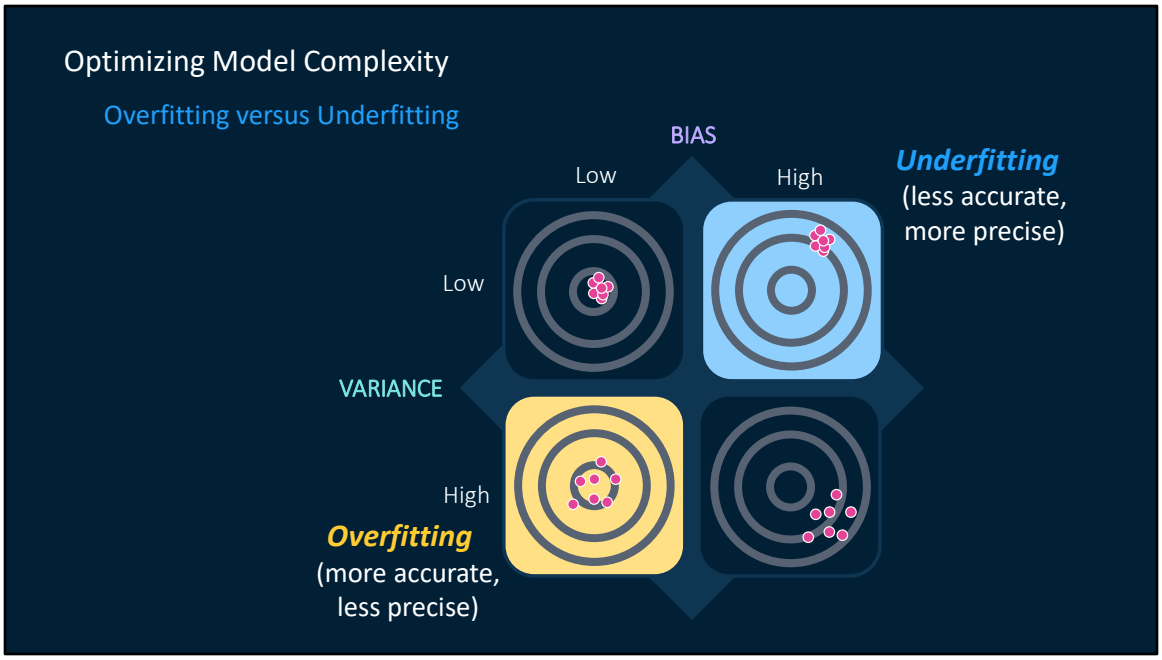
In the first target (located on the top left) , we can see an example of low bias and low variance. Bias is low because the hits are closely aligned to the center and there is low variance because the hits are densely positioned in one location.

The second target (located on the bottom left) shows a case of low bias and high variance. Although the hits are not as close to the bull's-eye as in the previous example, they are still close to the center, and bias is therefore relatively low. However, there is high variance this time because the hits are spread out from each other.

The third target (located on the top right) represents high bias and low variance, and the fourth target (located on the bottom right) shows high bias and high variance.

Hence, you can never make exact measurements in an experiment. You often use accuracy and precision for assessing predictive models. How far away you are from the "mark" is described by *accuracy,* and how well you measure is described by *precision*. An analogy can be made to the relationship between accuracy and precision. Accuracy is a description of bias. Low bias corresponds to more accurate. Conversely, high bias corresponds to less accurate. Accuracy is measuring near the target/true value.

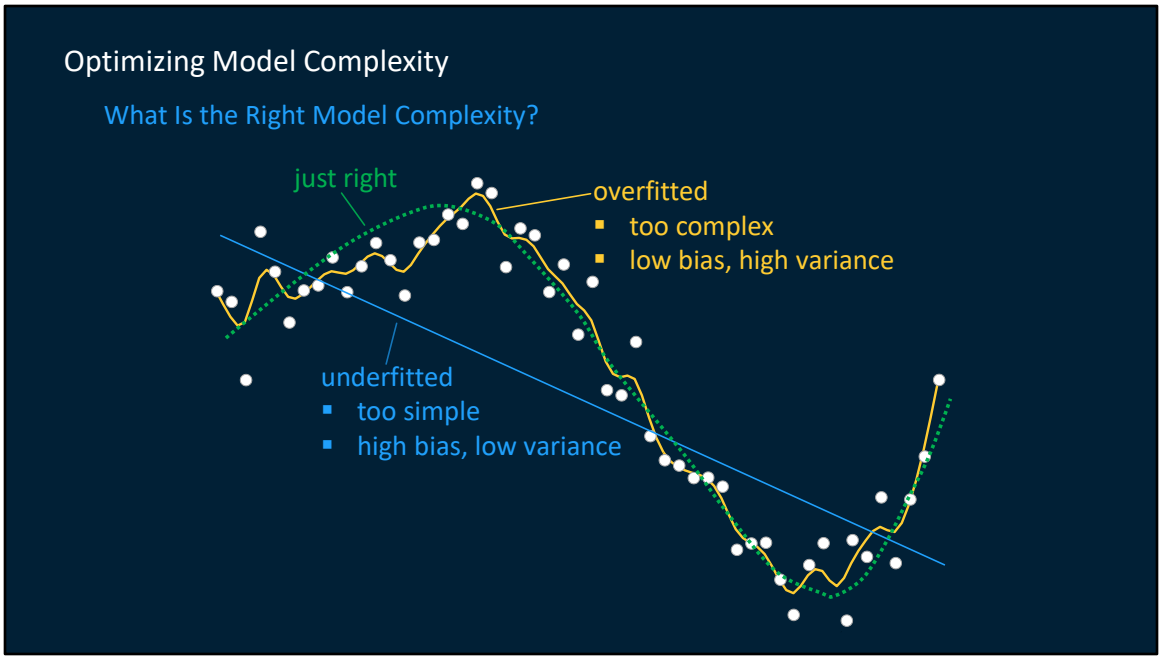
Precision is a description of variance. Low variance corresponds to more precise and whereas high variance corresponds to less precise. Precision is getting consistent results of repeated measurements.



Underfitting happens when a model is unable to capture the underlying pattern of the data. These models usually have high bias and low variance. In other words, they are less accurate and more precise.

Overfitting happens when a model captures the noise along with the underlying pattern in data. These models have low bias and high variance. In other words, they are more accurate and less precise.

A model with low bias and low variance is desirable, and a model with high bias and high variance is not.



Fitting a model to data requires searching through the space of possible models. Constructing a model with good generalization requires choosing the right complexity.

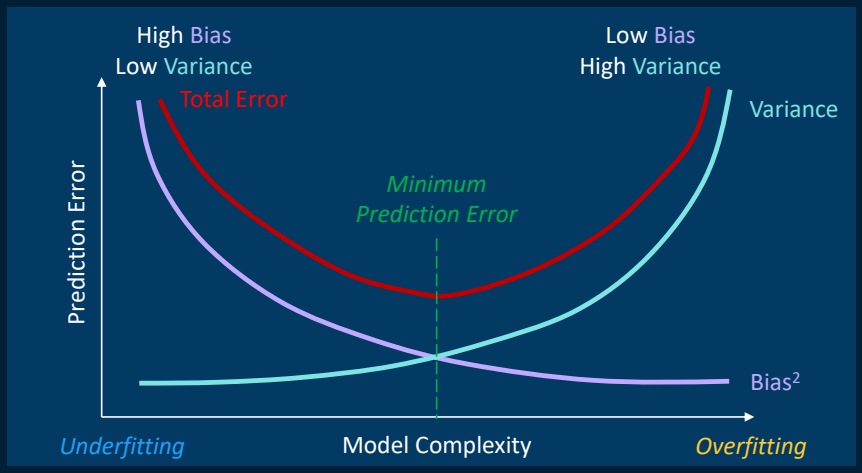
An insufficiently complex model might not be flexible enough. This leads to underfitting – systematically missing the signal. Simple, less complex models do not have the flexibility to fit every data set that well. They do not approximate the true relationship between predictors and the response variable.

An overly complex model might be too flexible. This will lead to overfitting – accommodating nuances of the random noise in the sample. These models, due to their flexibility, can fit training data too much, creating poor prediction performance in a new data set.

A model with just enough flexibility will give the best generalization. The strategy for choosing model complexity in data mining and machine learning is to select the model that performs best on the validation data set. Using the training data to evaluate performance usually leads to selecting too complex a model. (The classic example of this is selecting linear regression models based on R^2.)

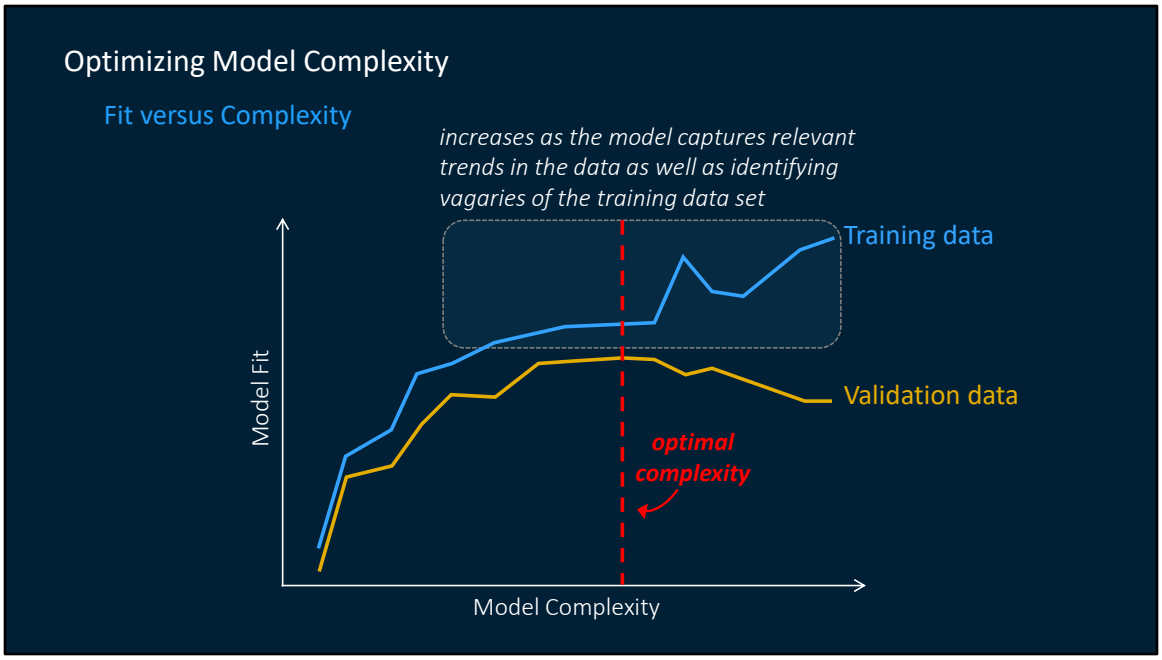
Optimizing Model Complexity

Optimum Balance of Bias and Variance



(Hastie, Tibshirani, and Friedman 2001)

In reality, there is always a trade-off between bias and variance. Bias and variance both contribute to error, but it is the prediction error that you want to minimize, not bias or variance specifically. An optimal balance of bias and variance would never overfit or underfit the model. Increasing the variance with complex models will decrease the bias, but that might overfit the model. Conversely, simple models will increase the bias at the expense of the model variance, and that might underfit the model. There is an optimal point for model complexity, a balance between overfitting and underfitting. In practice, there is no analytical way to find this optimal complexity. Instead, we must use an accurate measure of prediction error, explore different levels of model complexity, and choose the complexity level that minimizes the overall error.

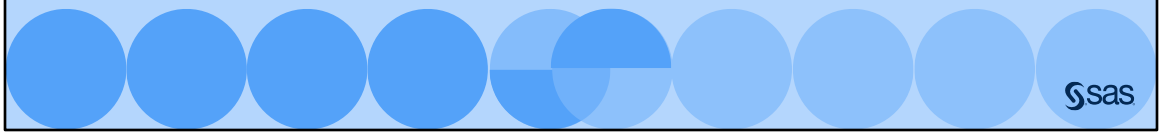


To compare many models, an appropriate fit statistic must be selected. For a series of models with increasing complexity, which could be generated by an automatic selection routine, it is conceivable to plot a fit measure against some index of complexity.

Typically, model performance follows a straightforward trend. As the complexity increases the fit on the training data gets better. After a point, the fit might plateau, but on the training data, the fit gets better as model complexity increases. Some of this increase is attributable to the model capturing relevant trends in the data. Detecting these trends is the goal of modeling. Some of the increase, however, is due to the model identifying vagaries of the training data set. This behavior has been called overfitting. Because these vagaries are not likely to be repeated, in the validation data or in future observations, it is reasonable to want to eliminate those models. Hence, the model fit on the validation data, for models of varying complexity, is also plotted. The typical behavior of the validation fit line is an increase (as more complex models detect more usable patterns) followed by a plateau, which might finally result in a decline in performance. The decline in performance is due to overfitting. The plateau just indicates more complicated models that have no fit-based arguments for their use. A reasonable rule would be to select the model associated with the complexity that has the highest validation fit statistic.

Decomposing Prediction Error

To explore this learning resource in greater detail, please visit the course page on learn.sas.com. There, you'll find comprehensive course materials, tutorials, and updates to support your learning journey.



Question: Predictive Models

Predictive models are widely used and are available in many varieties. Any model, however, must perform the following essential tasks:

- provide a rule to transform an input measurement into a prediction
- be able to adjust its complexity to balance overfitting and underfitting.

a. True
b. False

Answer: Predictive Models

Predictive models are widely used and are available in many varieties. Any model, however, must perform the following essential tasks:

- provide a rule to transform an input measurement into a prediction
- be able to adjust its complexity to balance overfitting and underfitting.

a. True
b. False

The correct answer is True.

4.2 Categorical Associations

Question: Categorical Variables

Which of the following would likely be considered categorical in the PVA_Donors data? (Select all that apply.)

a. Age (**DemAge**)
b. Gender (**DemGender**)
c. Gift Amount Last (**GiftAvgLast**)
d. Time Since Last Gift (**GiftTimeLast**)
e. Homeowner (**DemHomeOwner**)

Answer: Categorical Variables

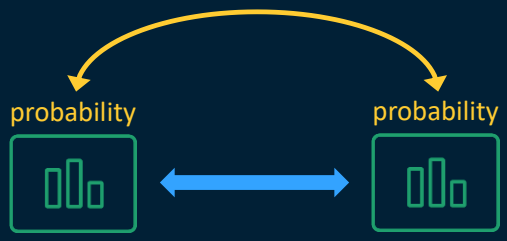
Which of the following would likely be considered categorical in the PVA_Donors data? (Select all that apply.)

a. Age (**DemAge**)
b. Gender (**DemGender**)
c. Gift Amount Last (**GiftAvgLast**)
d. Time Since Last Gift (**GiftTimeLast**)
e. Homeowner (**DemHomeOwner**)

The correct answer is b and e.

Gender is a categorical variable having three levels: male, female, and unknown. Similarly, **Homeowner** is also categorical having two levels: yes and no. **Age**, **Gift Amount Last**, and **Time Since Last Gift** are continuous variables.

Associations between Categorical Variables

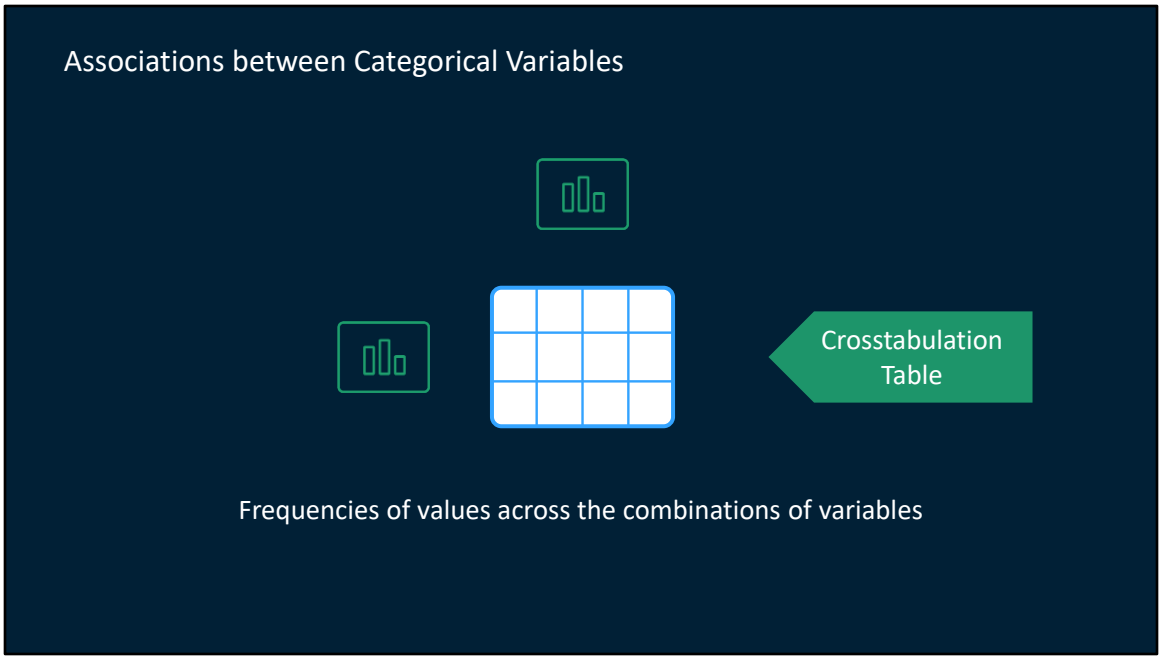


- The distribution of one variable changes when the value of the other variable changes.
- The probability of one variable depends on the probability of the other.

By examining distributions of categorical variables, you can determine the frequencies of data values and possible associations among variables. An association exists between two categorical variables if the distribution of one variable changes when the value of the other variable changes.

You can also think of an association as the probability that one variable depends on the probability of the other.

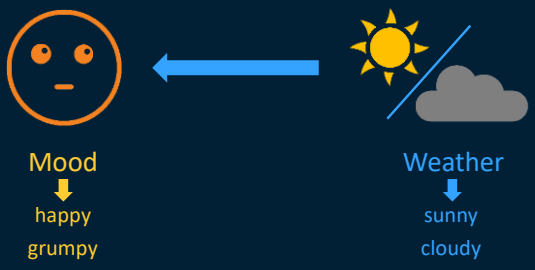
More precisely, the difference between conditional and marginal probabilities is an indication that the variables might be associated. If there's no association, the distribution of the first variable is the same, regardless of the level of the other variable.



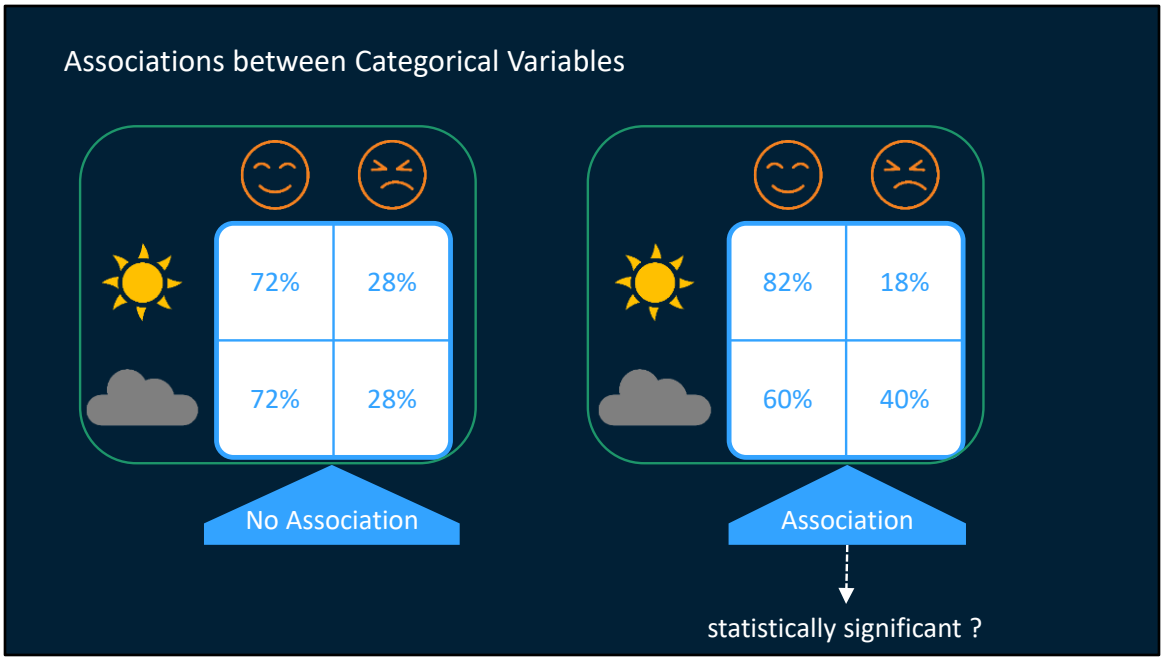
To look for associations, you examine the frequencies of values across the combinations of variables.

Crosstabulation is commonly used to quantitatively analyze the relationship between two or more categorical variables. Cross tabulations — also referred to as contingency tables or crosstabs — group categorical variables together and enable you to understand their association.

Associations between Categorical Variables



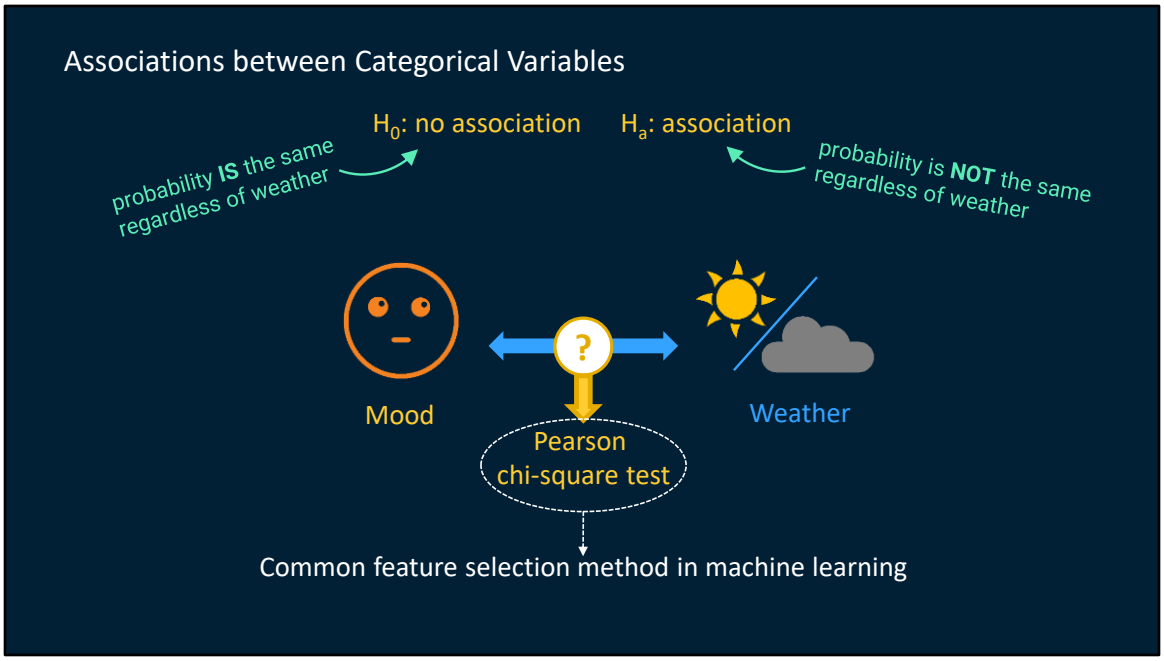
For example, suppose you want to determine whether your mood is affected by the weather. The categorical response variable is **Mood**, and its values are either *happy* or *grumpy*. The categorical predictor variable is **Weather**, and its values are either *sunny* or *cloudy*.



This table shows row frequency percentages for combinations of values of the two variables. On sunny days, you're happy 72% of the time and grumpy 28% of the time. Now look at the frequency percentages for your mood on cloudy days. The row percentages are the same in each column, indicating that there's no change in your mood based on the weather. So, there's no association between these two variables.

What if these were your results? In this table, the row percentages are different in each column. On sunny days, you're happy 82% of the time and grumpy 18% of the time. On cloudy days, you're happy 60% of the time and grumpy 40% of the time. Your mood changes based on the weather. It appears you're more likely to be happy if the weather is nicer, indicating a possible association between mood and weather.

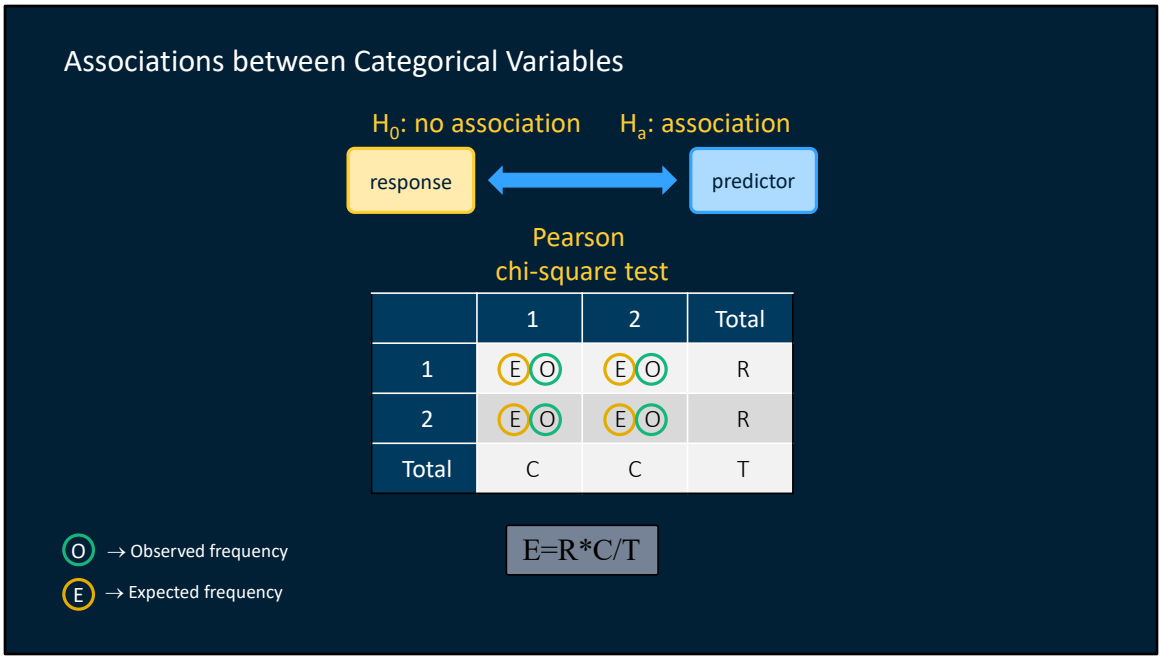
We need to assess whether the differences between the percentages of **Mood** across levels of **Weather** are greater than would be expected by chance. To be certain that the variables are associated, we need to run a formal test of association, the chi-square test.



Let's start with our null hypothesis, that there's no association between the variables **Weather** and **Mood**, meaning that the probability of a happy mood is the same regardless of weather.

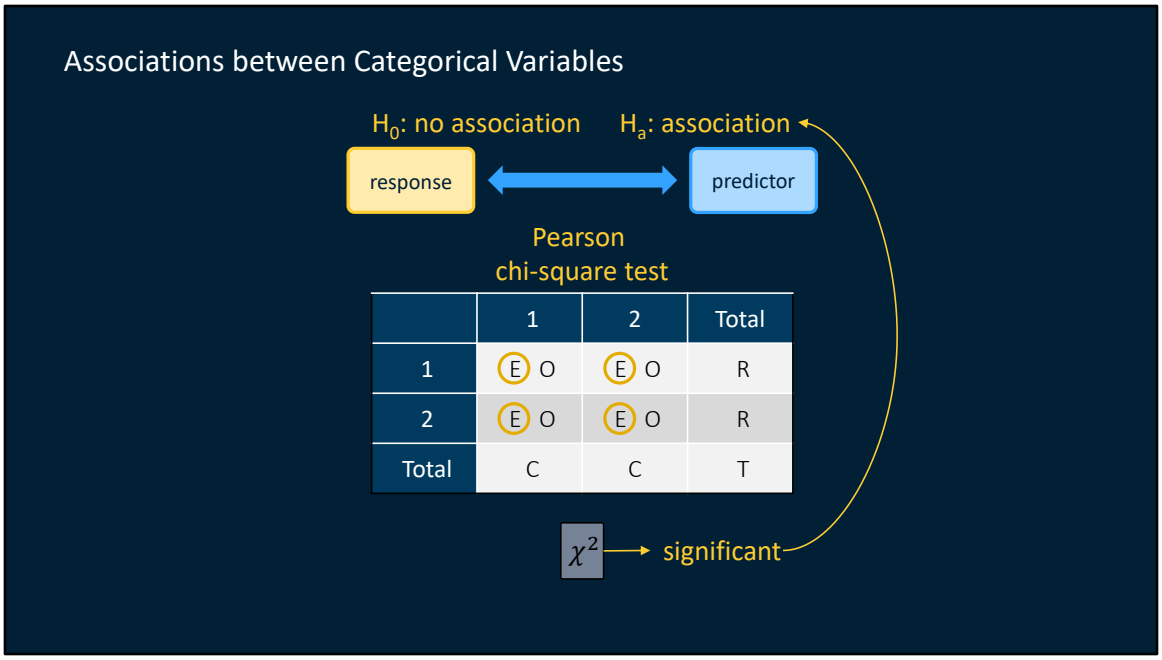
The alternative hypothesis is that there *is* an association between **Weather** and **Mood**, meaning the probability of a happy mood is *not* the same for sunny and cloudy weather.

To formally test the association for statistical significance, we'll use the Pearson chi-square test, often referred to as simply the chi-square test. Chi-square test is commonly used for feature selection in machine learning. For example, it can be used as a split search criterion in decision trees. In feature selection, we aim to select the features that are highly dependent on the response.



The chi-square measures the difference between the observed cell counts and the cell counts that are expected if there's no association between the variables and the null hypothesis is in fact true.

The expected counts for each cell is calculated by multiplying the row total (R) by the column total (C), and then dividing the result by the total sample size (T).

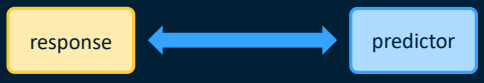


The larger the difference between the observed and expected cell counts, the more evidence of a statistically significant association between the variables.

A significant chi-square statistic provides evidence to reject the null hypothesis and conclude that an association exists. To measure the magnitude of an association, we'll use measures of association such as Odds ratio.

Associations between Categorical Variables

H_0: no association H_a: association



Pearson chi-square test

	1	2	Total
1	E O	E O	R
2	E O	E O	R
Total	C	C	T

$$\sum \frac{(\text{observed} - \text{expected})^2}{\text{expected}}$$

To calculate the chi-square test statistic, you square the difference between the observed and expected counts for each cell, and divide by the expected cell count to get the cell chi-square values.

Associations between Categorical Variables

H_0: no association H_a: association

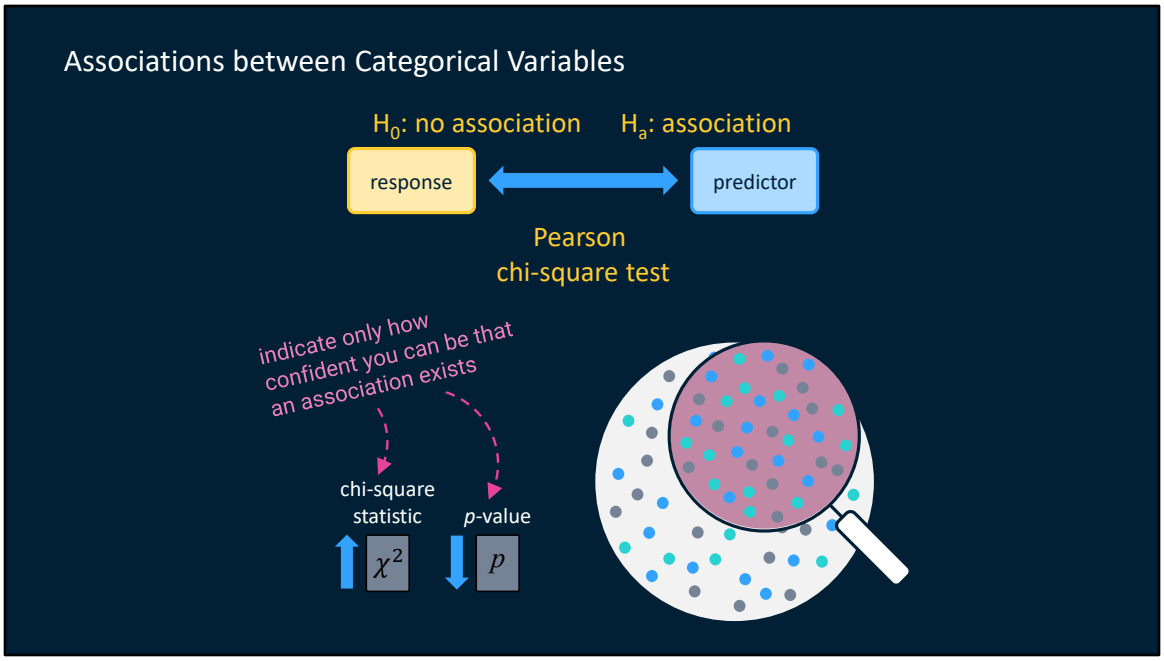
response ⟷ predictor

Pearson chi-square test

	1	2	Total
1	E O	E O	R
2	E O	E O	R
Total	C	C	T

$$\sum\sum\left(\frac{(\text{observed}_{rc}-\text{expected}_{rc})^2}{\text{expected}_{rc}}\right)$$

The overall chi-square, or the test statistic, is calculated by then adding the cell chi-square values over all cells.



Keep in mind that neither the chi-square statistic nor its *p*-value tells you the magnitude of an association. They indicate only how confident you can be that an association exists.

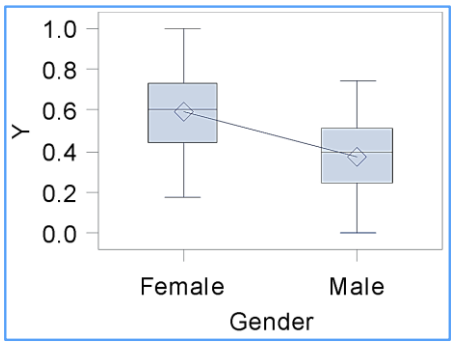
Chi-square statistics and their *p*-values depend on and reflect the sample size. A larger sample size yields a larger chi-square statistic and a smaller corresponding p-value, even though the association might not be strong.

For example, if you double the size of your sample by duplicating each observation, you double the value of the chi-square statistic, even though the strength of the association does not change.

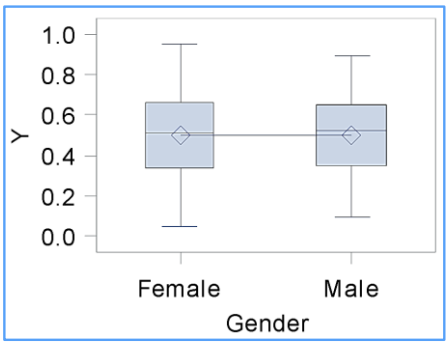
Question: Association

You have already seen association between two continuous variables (Lesson 3) and association between two categorical variables (this lesson). Below are illustrations of relationships between a continuous variable and a categorical variable. Which of the following represents an association?

a.



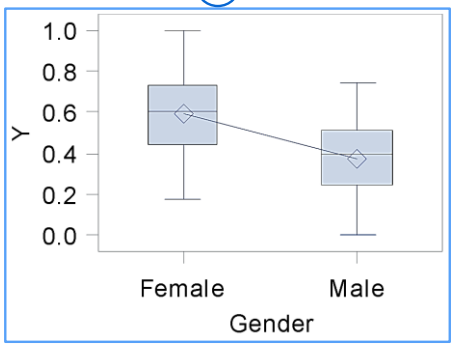
b.



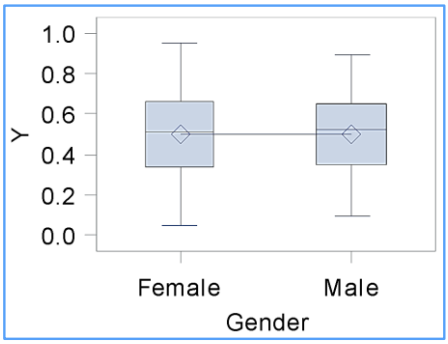
Answer: Association

You have already seen association between two continuous variables (Lesson 3) and association between two categorical variables (this lesson). Below are illustrations of relationships between a continuous variable and a categorical variable. Which of the following represents an association?

a.



b.

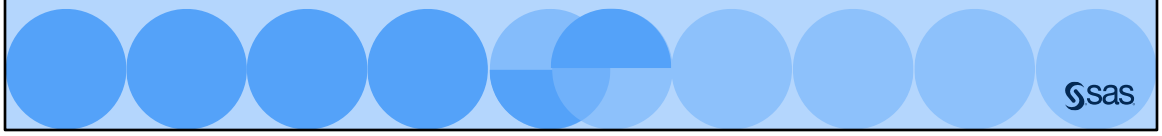


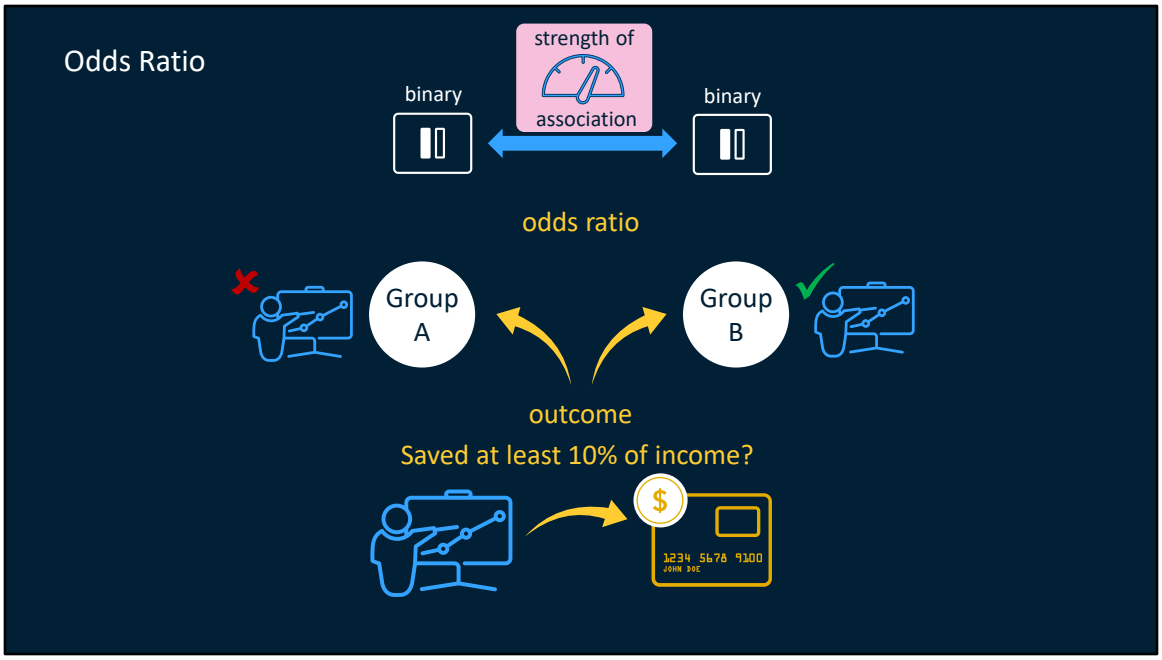
Correct answer: a

Where there is an association between the categorical predictor and the continuous response, a plot will show a non-horizontal line connecting the category-specific means of y. The category-specific means will be better predictions than the overall mean of y (option **a**).
A horizontal line through the means of all levels of the categorical predictor will indicate lack of association and therefore equal means (option **b**).

Cramer's V Statistic

To explore this learning resource in greater detail, please visit the course page on learn.sas.com. There, you'll find comprehensive course materials, tutorials, and updates to support your learning journey.





To measure the strength of the association between a binary predictor variable and a binary response variable, you can use odds ratio. An odds ratio indicates how much more likely it is that a certain event, or outcome, occurs in one group relative to its occurrence in another group.

For example, suppose you want to see the effect that fiscal training has on spending habits. The people in Group B attended a seminar about saving money, and the people in Group A did not. The outcome variable indicates whether each person saved at least 10% of income in the year following the seminar. In this example, you want to know how much more likely it is that a Yes outcome occurs in Group B relative to its occurrence in Group A, with respect to the odds.

Odds Ratio

$$\text{Odds} = \frac{P_{event}}{1 - P_{event}}$$

Group	Outcome		
	Yes	No	Total
A	60	20	80
B	90	10	100
Total	150	30	180

Probability of Yes in Group B

$$90/100 = .90 = 90\%$$

Probability of No in Group B

$$10/100 = .10 = 10\%$$

Probability of Yes in Group A

$$60/80 = .75 = 75\%$$

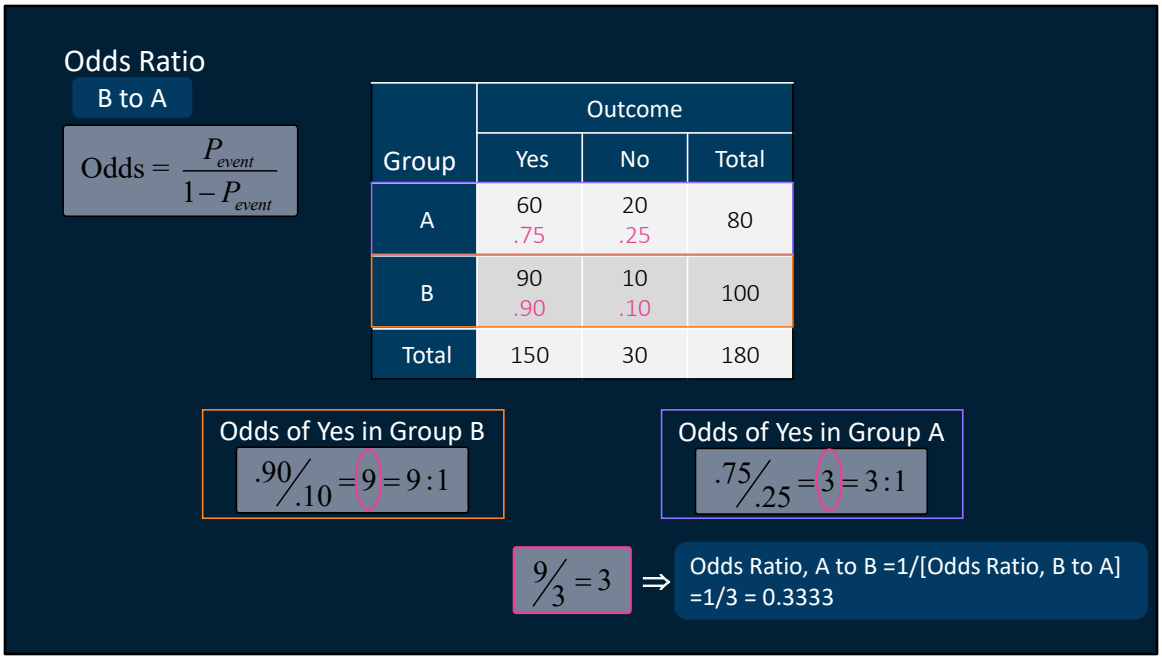
Probability of No in Group A

$$20/80 = .25 = 25\%$$

So, what are odds? Odds are not the same as probabilities. Instead, odds are calculated from probabilities. You divide the probability that the event occurs by the probability that the event does not occur.

To calculate the probability of a Yes outcome in Group B, you divide the number of Yes responses (90) by the total number of observations (100) to get 0.90, or 90%, the probability of having the Yes outcome in Group B. The probably of a No outcome in Group B is 10 divided by 100, which is 0.10, or 10%.

To calculate the probability of a Yes outcome in Group A, you divide 60 by 80, which is 0.75. The probability of a No outcome in Group A is 20 divided by 80, which is 0.25.

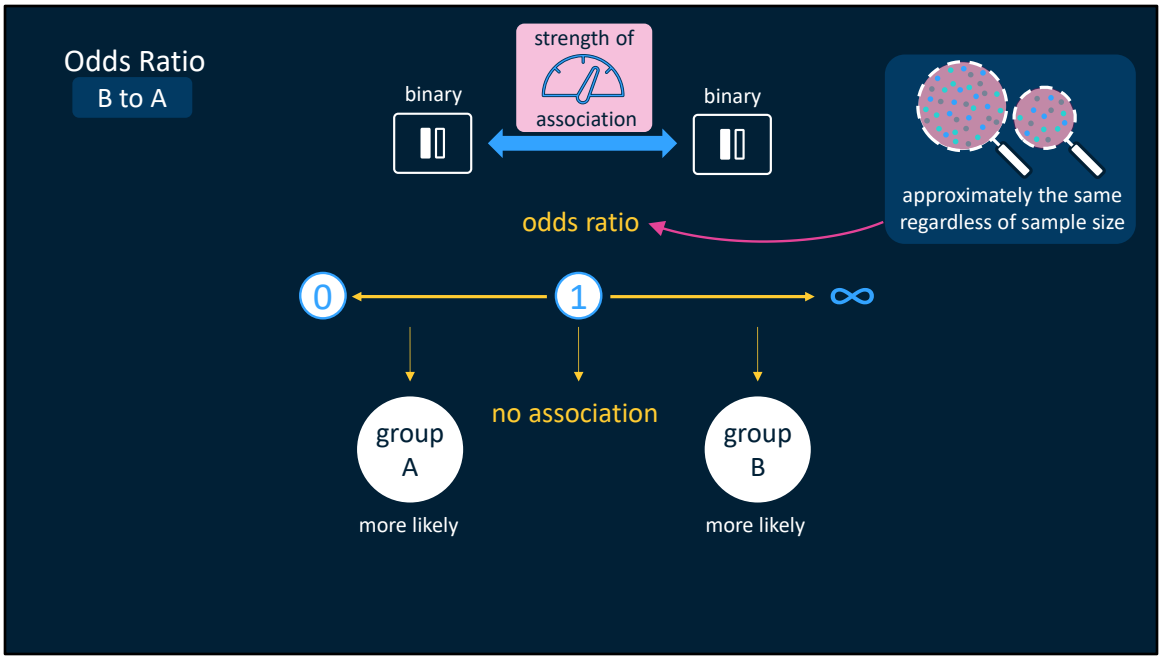


Now that we know the conditional probabilities of our response given the predictor, let's calculate the odds. The odds of the outcome occurring in Group B are the probability of a Yes outcome, 0.90, divided by the probability of a No outcome, 0.10. The odds are 9, or 9:1, which means that we expect nine occurrences to one non-occurrence in Group B.

To calculate the odds of the outcome occurring in Group A, you divide the probability of a Yes outcome, 0.75, by the probability of a No outcome, 0.25, which is 3, or 3:1. This means that you expect three times as many occurrences as non-occurrences in Group A.

Now that you know the odds of the outcome in both groups, you can compare the odds by calculating an odds ratio. You divide the odds of an outcome in Group B, 9, by the odds in Group A, 3, with a result of 3. An odds ratio of 3 means that the odds of getting the outcome in Group B are three times those of getting the outcome in Group A. So, in this example, the odds of saving at least 10% of income are three times higher for people who received training.

If you were interested in the odds ratio of group A to group B, you would simply take the multiplicative inverse (or reciprocal) of 3 to arrive at 0.3333, which indicates that the odds of getting the outcome in group A are one-third those in group B.



The odds ratio shows the strength of the association between the predictor variable and the outcome variable, and the value can range from 0 to infinity.

When the odds ratio is 1, there's no association between the predictor variable and the outcome variable. If the odds ratio is greater than 1, the group in the numerator (here Group B) is more likely to have the outcome. If the odds ratio is less than 1, the group in the denominator (here Group A) is more likely to have the outcome. The odds ratio is approximately the same regardless of sample size.

Question: Strength of an Association

A researcher wants to measure the strength of an association between two binary variables. Which statistic can she use?

a. Pearson Correlation
b. Odds Ratio
c. Coefficient of Determination (R2)
d. None of the above

Answer: Strength of an Association

A researcher wants to measure the strength of an association between two binary variables. Which statistic can she use?

a. Pearson Correlation
b. Odds Ratio
c. Coefficient of Determination (R2)
d. None of the above

The correct answer is b.

Odds ratio can be used to measure the strength of an association between two binary variables.

Pearson correlation is used to measure the strength of association between two interval variables. The coefficient of determination is the percentage of the variability of the dependent interval variable that is explained by the variation of the independent variables. The Spearman rank correlation is used to measure the strength of association between two ordinal variables. If both variables instead are binary, phi-coefficient can be used.

Demo: Examining Categorical Association

To explore this learning resource in greater detail, please visit the course page on learn.sas.com. There, you'll find comprehensive course materials, tutorials, and updates to support your learning journey.

4.3 Logistic Regression Model

Question: Regression Models

Which of the following statements is true regarding regression models? (Select all that apply.)

a. Linear regression can be used only for explanatory modeling.
b. Logistic regression can be used only for predictive modeling.
c. Linear regression can be used for explanatory modeling as well as for predictive modeling.
d. Logistic regression can be used for predictive modeling as well as for explanatory modeling.

Answer: Regression Models

Which of the following statements is true regarding regression models? (Select all that apply.)

a. Linear regression can be used only for explanatory modeling.

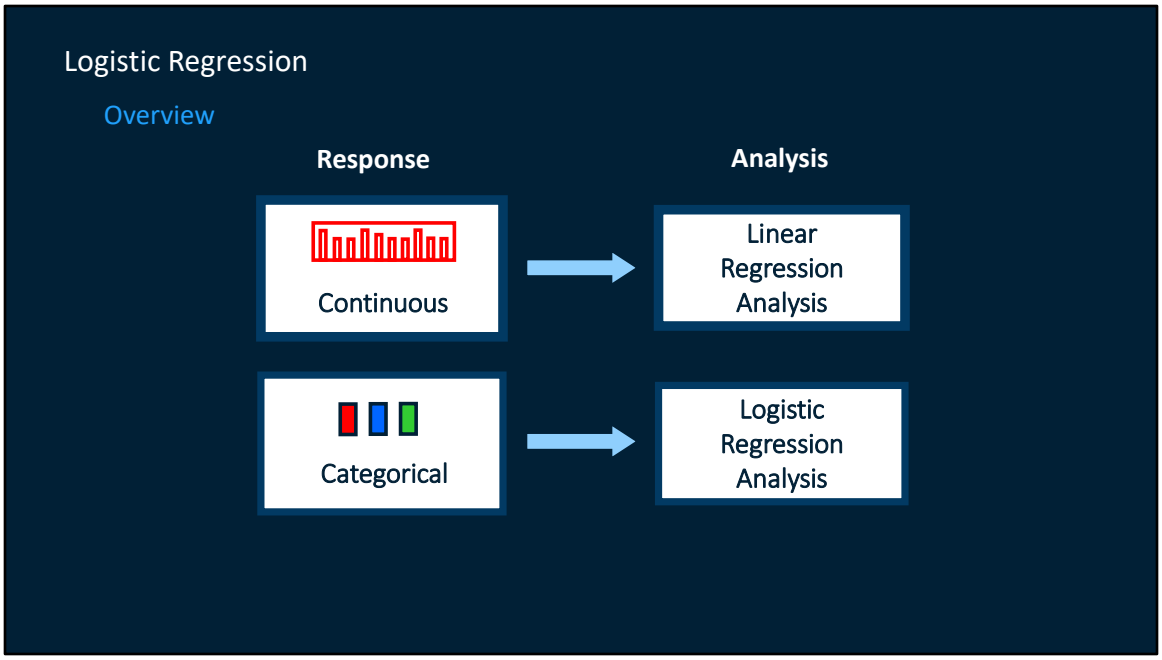
b. Logistic regression can be used only for predictive modeling.

c. Linear regression can be used for explanatory modeling as well as for predictive modeling.

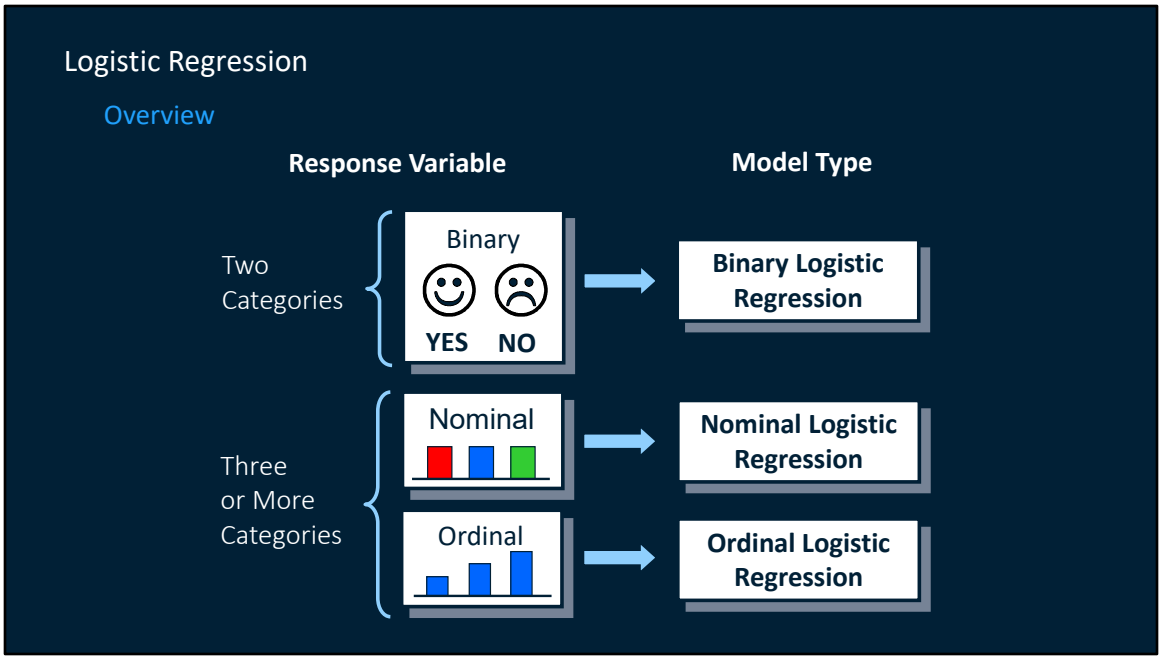
d. Logistic regression can be used for predictive modeling as well as for explanatory modeling.

ANSWER: c and d.

There are two main objectives of any regression model including linear and logistic regressions. First, assess the significance of the predictor variables in explaining the behavior of the response variable, and second, predict the values of the response variable given the values of the predictor variable. The distinction between using these regressions for explanatory modeling and predictive modeling is somewhat artificial. A model developed for prediction is probably a good explanatory model. Conversely, a model developed for an explanation is probably a good prediction model.



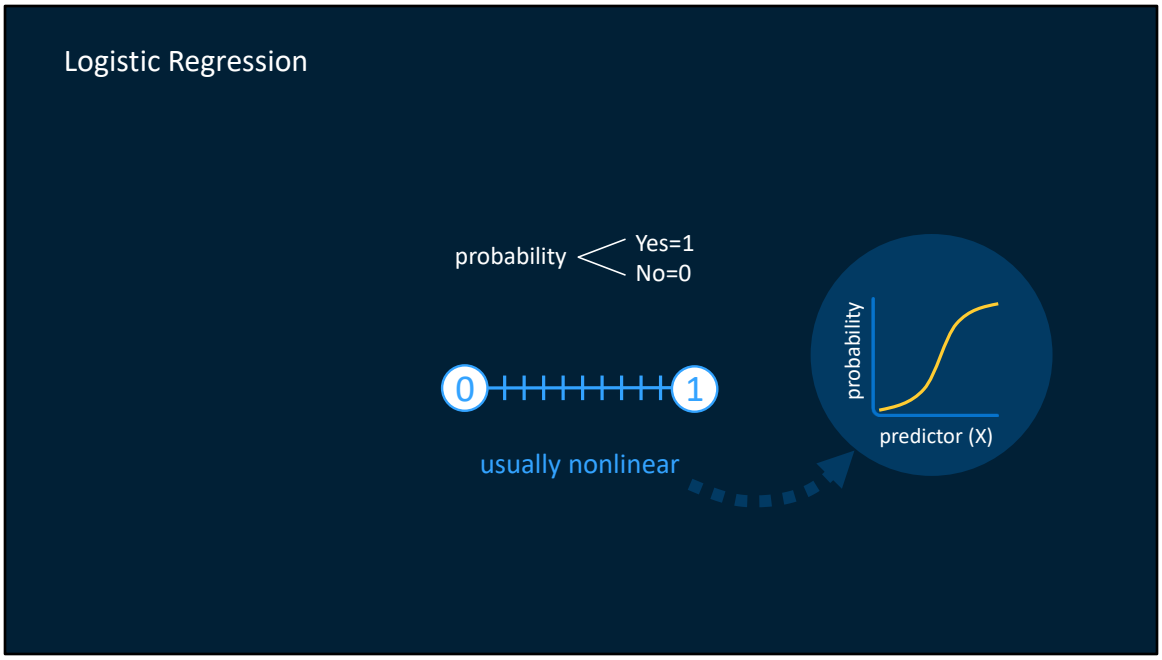
Regression analysis enables you to characterize the relationship between a response variable and one or more predictor variables. In linear regression, the response variable is continuous. In *logistic regression*, the response variable is categorical.



If the response variable is dichotomous (meaning it has two categories), the appropriate logistic regression model is binary logistic regression. This is the most common logistic regression model in data mining and machine learning applications, so it's the primary focus in this course.

If you have more than two categories (or levels) within the response variable, there are two possible logistic regression models:

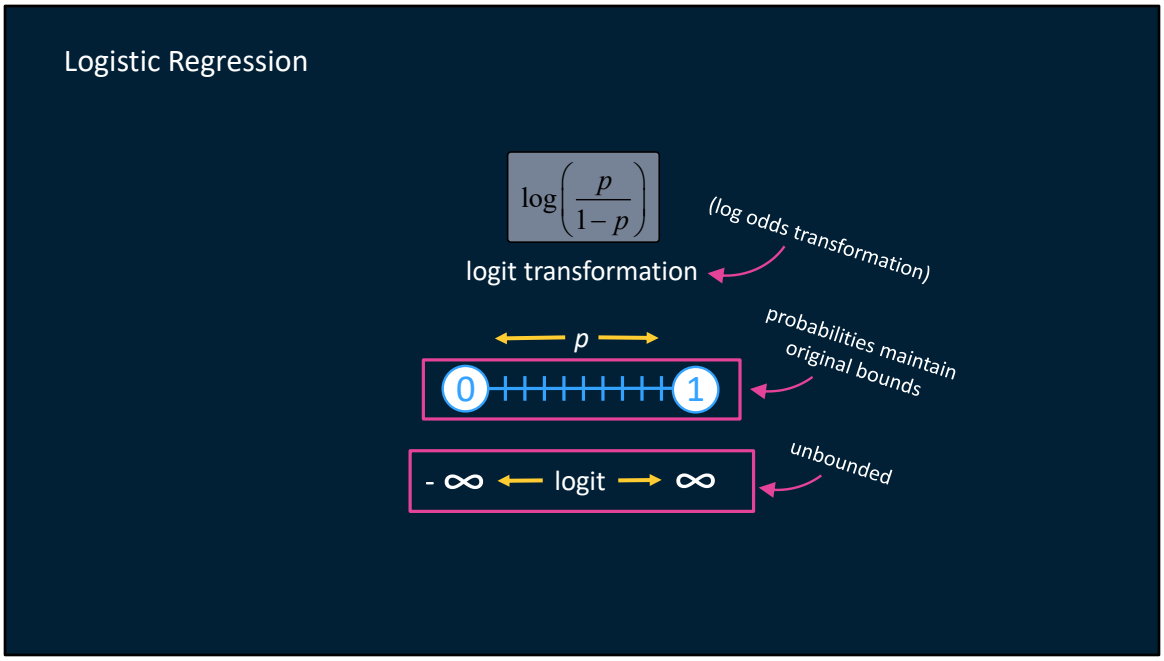
- If the response variable is nominal, you fit a nominal logistic regression model.
- If the response variable is ordinal, you fit an ordinal logistic regression model.



In logistic regression, we want to model the probability of both levels of the response. Although probabilities are continuous, they are bounded between 0 and 1. Also, the relationship between the probability of the outcome and a predictor variable is usually nonlinear rather than linear.

In fact, the relationship often resembles an S-shaped, or sigmoidal, curve.

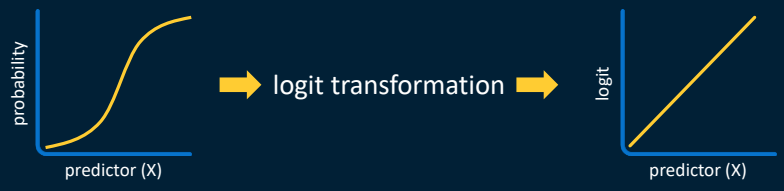
Logistic Regression



A logistic regression model applies a logit transformation, or simply the log odds transformation, to the probabilities. Here, log means 'natural' log, rather than log with a base 10. The logit effectively avoids the boundary problem for probabilities.

As *p* approaches its minimum value of 0, the logit approaches negative infinity, and as *p* approaches its maximum value of 1, the logit approaches positive infinity. So, the logit is unbounded, just like in linear regression, but the probabilities maintain the original bounds of 0 to 1.

Logistic Regression

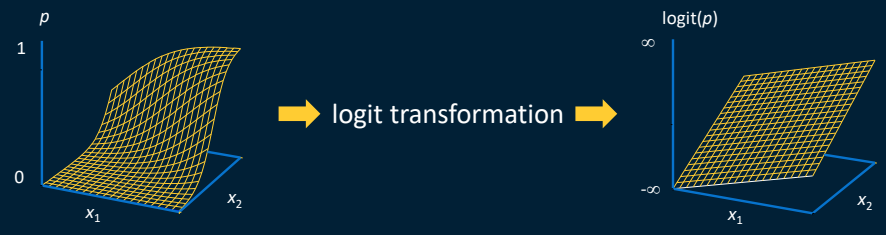


$$\text{logit}(p_i) = \beta_0 + \beta_1 X_i$$

In addition, the logit transformation enables us to move from modeling the probability with a sigmoidal nonlinear curve to modeling the logit with a linear function of the predictors. And modeling the logit allows you to indirectly model the probability of the response. Whatever the predicted value of the logit is, we can simply back-transform to the probability scale to get a value between 0 and 1.

The logistic model now looks quite familiar, compared to linear regression. In a two-dimensional space, with a binary target and one predictor (X), the logit is equal to a linear function of your parameters and predictor. Again, β_0 is the intercept of the regression equation, and β_1 is the slope of the model predictor. Unlike linear regression, we no longer assume normally distributed errors in logistic regression. Here the error is a function of p, the probability of the event.

Logistic Regression

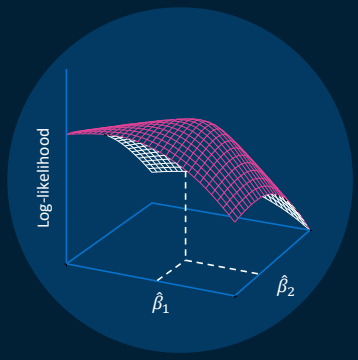


$$\text{logit}(p_i) = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i}$$

In a three-dimensional space, with a binary target and two predictors (X_1 and X_2), the logit is equal to a linear function of your parameters and predictors. Again, β_0 is the intercept of the regression equation, and the other betas are the slopes of the model predictors.

The graph of a linear combination on the logit scale is a (hyper)plane. Different parameter values give different surfaces with different slopes and different orientations. On the probability scale, it becomes a sigmoidal surface. The nonlinearity of the model is used to deal with the constrained scale of the target.

Logistic Regression



$$\text{logit}(p_i) = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i}$$

Method of Maximum Likelihood

To fit the model, logistic regression requires a more computationally complex estimation method, named the method of maximum likelihood, to estimate the parameters. This method finds the values of the parameters that make the observed data most likely. This is accomplished by maximizing the likelihood function that expresses the probability of the observed data as a function of the unknown parameters.

Question: Bounds for a Logit

What are the upper and lower bounds for a logit?

a. Lower=0, Upper=1
b. Lower=0, No upper bound
c. No lower bound, No upper bound
d. No lower bound, Upper=1

$$logit(p_i) = \ln\left(\frac{p_i}{(1-p_i)}\right)$$

Answer: Bounds for a Logit

What are the upper and lower bounds for a logit?

a. Lower=0, Upper=1
b. Lower=0, No upper bound
c. No lower bound, No upper bound
d. No lower bound, Upper=1

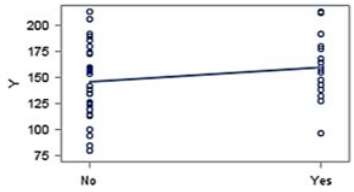
$$logit(p_i) = \ln\left(\frac{p_i}{(1-p_i)}\right)$$

The answer is c.
A probability is bounded by 0 and 1. The logit of the probability transforms the probability into a linear function, which has no lower or upper bounds. So a logit has no lower or upper bounds.

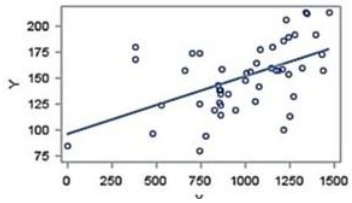
Question: Logistic Regression Model

With response plotted on the Y-axis and the predictor on the X-axis, which of the following is a **logistic regression model**?

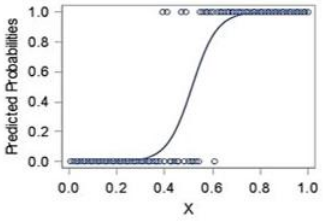
a.



b.



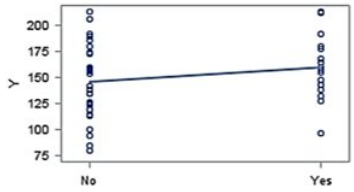
c.



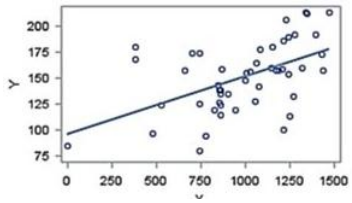
Answer: Logistic Regression Model

With response plotted on the Y-axis and the predictor on the X-axis, which of the following is a **logistic regression model**?

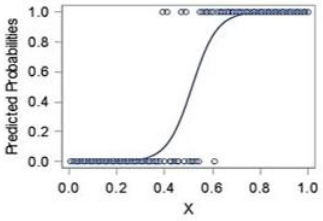
a.



b.



c.

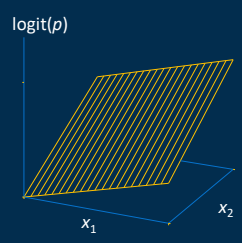


The answer is c.
Picture a is an ANOVA (analysis of variance) model. Here the response variable is continuous, and you have a binary predictor. This is typically a case of t test or ANOVA.
Picture b is a linear regression model. Here the response variable is continuous, and you have a continuous predictor. You perform ordinary least squares regression.
Picture c is a logistic regression model. When the response variable is binary, there can be an indirect modeling of the variable, by use of a logit link function.

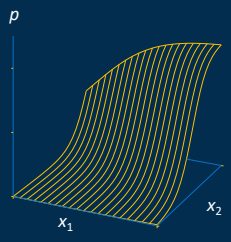
Interpreting the Odds Ratio

$$logit(\hat{p}) = log\left(\frac{\hat{p}}{1-\hat{p}}\right) = \hat{\beta}_0 + \hat{\beta}_1 \cdot x_1 + \hat{\beta}_2 \cdot x_2$$

Unit change in $x_2 \Rightarrow$



$\hat{\beta}_2$ change in logit

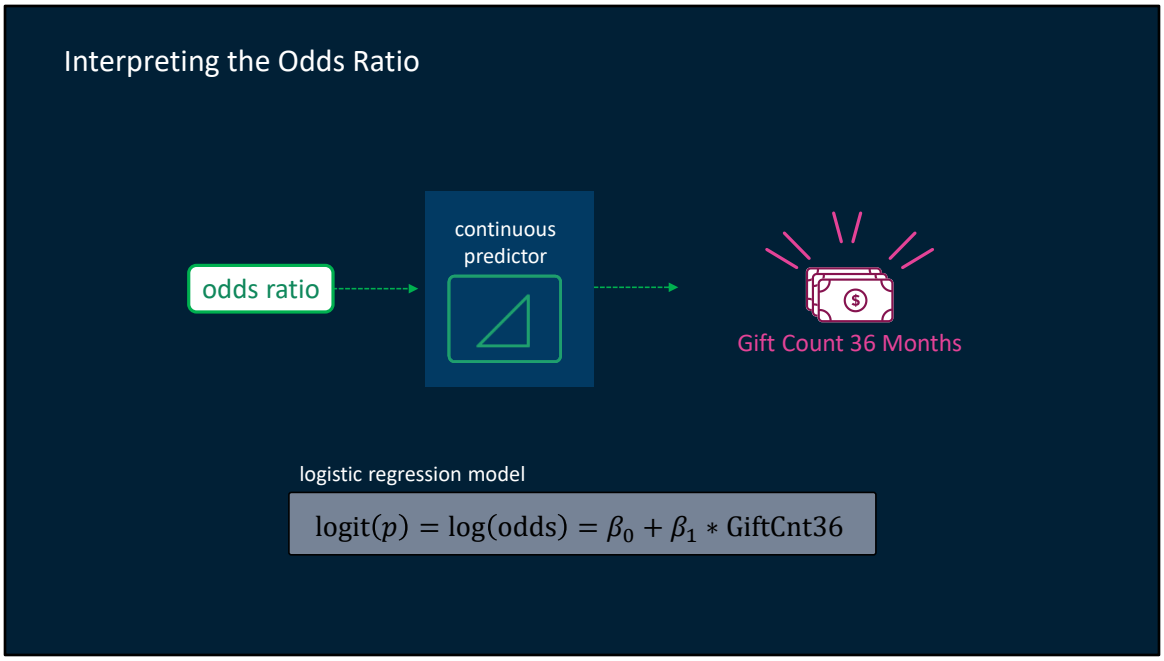


$100\ (exp(\hat{\beta}_2) - 1)\%$ change in the odds

$\hat{\beta}_2$ = expected change in the *Logit* for one unit change in x_2

$e^{\hat{\beta}_2}$ = **odds ratio** for x_2

A linear-additive model is particularly easy to interpret since each input variable affects the logit linearly. The coefficients are the slopes. Exponentiating each parameter estimate gives the odds ratios, which compares the odds of the event in one group to the odds of the event in another group.



To help interpret the odds ratio, let's see how to calculate the odds and the odds ratio from the logistic regression model. For a continuous predictor variable such as **GiftCnt36**, the odds ratio measures the increase or decrease in odds associated with a one-unit difference of the predictor variable.

Remember, the logit is the natural log of the odds. Because you can calculate an estimated logit from the logistic model, the odds can be calculated by simply exponentiating that value. An odds ratio for a one-unit difference is then the ratio of the exponentiated predicted logits that are one unit apart.

Interpreting the Odds Ratio

Odds Ratio Estimate				
Effect	**Unit**	**Point Estimate**	**95% Wald Confidence Limits**	
GiftCnt36	1	1.143	1.121	1.165

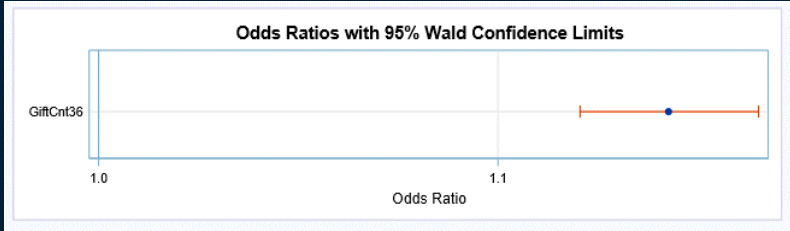
large samples

Wald confidence interval

versus

Profile likelihood confidence interval

small samples

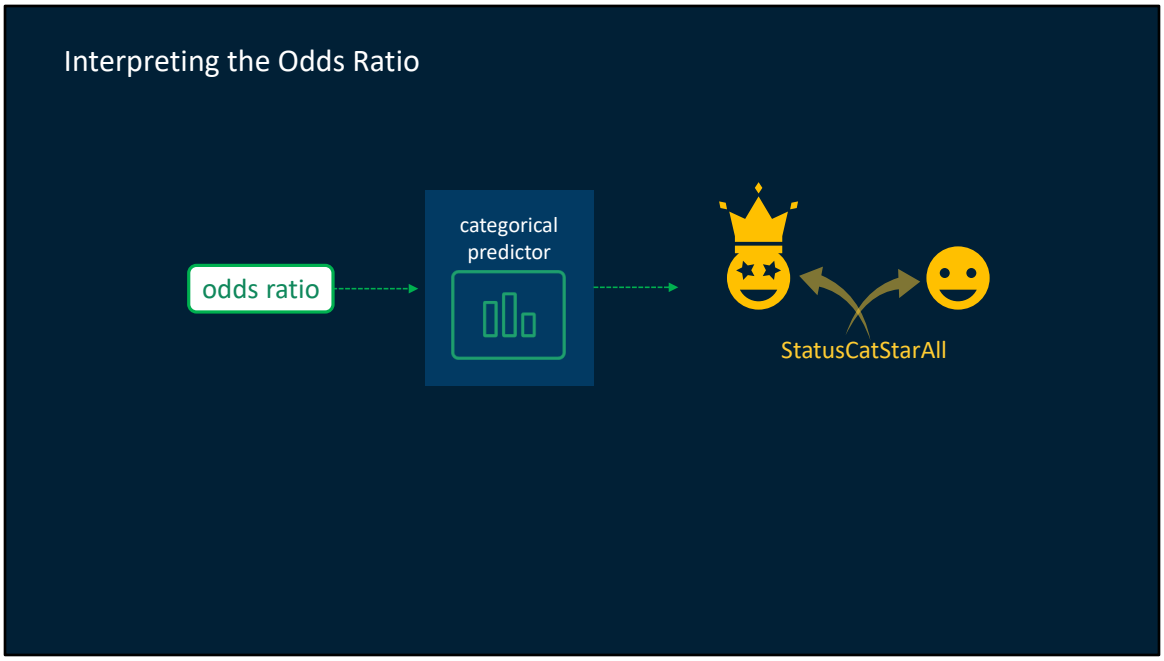


For **GiftCnt36**, the odds ratio estimate equals 1.143. This means that for each additional donation in the past 36 months, the odds of donation during the 97NK campaign change by a factor of 1.143, a 14.3% increase.

Because the 95% confidence interval, 1.121 to 1.165, does not include 1.000, the odds ratio is significant at the 0.05 alpha level, and therefore, the predictor **GiftCnt36** is significantly different from 0.

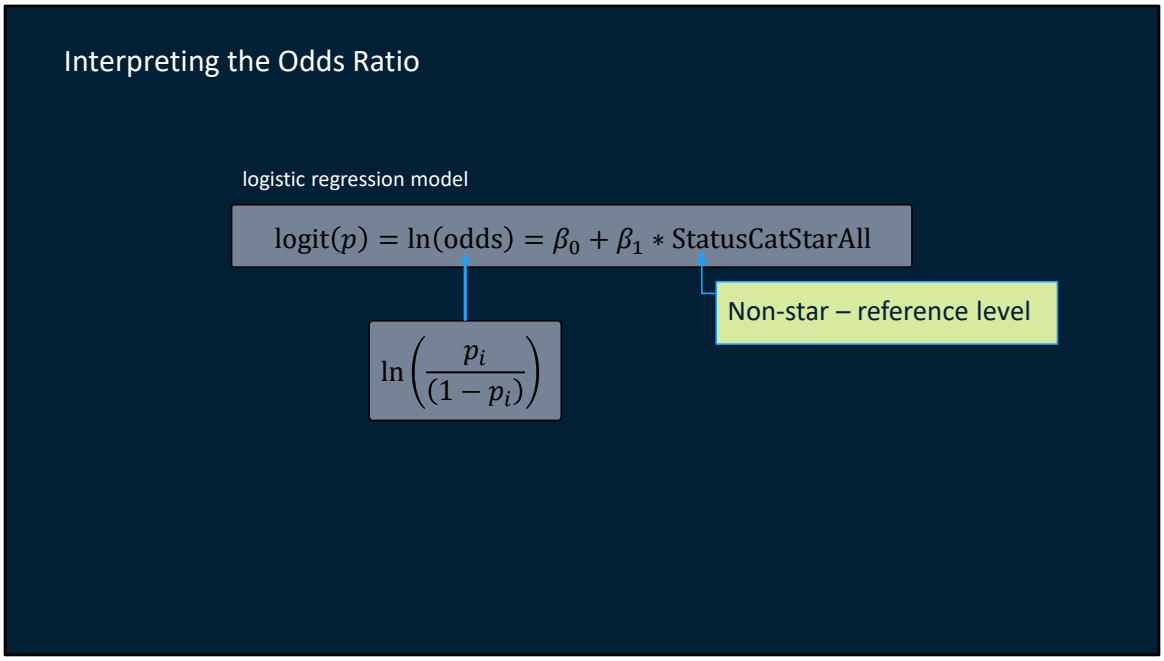
Note that the Wald-based confidence intervals are particularly suitable in the machine learning world, but different from profile likelihood confidence intervals you had seen earlier. This difference is because the Wald confidence intervals use a normal error approximation, whereas the profile likelihood confidence intervals are based on the value of the log-likelihood. These likelihood-ratio confidence intervals require a much greater number of computations, but are generally preferred to the Wald confidence intervals, especially for sample sizes less than 50.

The Odds Ratio plot displays the results of the Odds Ratio table graphically. This plot is obtained by applying the parameter estimates from the logistic model to values of the predictors, and then converting the predictions to the probability scale. A reference line shows the null hypothesis, an odds ratio equal to 1. When the confidence interval crosses the reference line, the effect of the variable is not significant.



Calculating and interpreting odds ratios for categorical variables is similar to that of continuous variables. Consider a categorical predictor variable such as **StatusCatStarAll** that has two levels, star donors and non-star donors.

Interpreting the Odds Ratio



Imagine now that we fit a logistic regression model with the predictor **StatusCatStarAll** instead of **GiftCnt36**.

The logit of p is also equal to the linear predictor for our model: $\beta_0+\beta_1$*StatusCatStarAll.

In this case, we use the level non-star to represent the reference level.

Interpreting the Odds Ratio

logistic regression model

$$\text{logit}(p) = \ln(\text{odds}) = \beta_0 + \beta_1 * \text{StatusCatStarAll}$$

Non-star=0
Star=1

$$\text{odds}_{\text{star}} = e^{\beta_0+\beta_1}$$

$$\text{odds}_{\text{non-star}} = e^{\beta_0}$$

$$\text{odds ratio} = \frac{e^{\beta_0+\beta_1}}{e^{\beta_0}} = e^{\beta_1}$$

So, individuals who belong to the non-star category are coded as 0 and those who belong to star category are coded as 1.

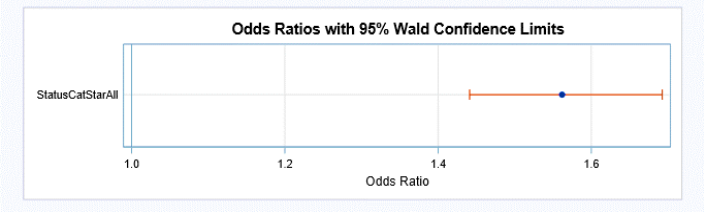
To obtain the odds for a star category, we exponentiate the linear predictor for the level. First, we substitute 1 for StatusCatStarAll to get $\beta_0+\beta_1$ as the linear predictor. Then, we add the parameter estimates that we got for β_0 and β_1 and exponentiate the sum.

To obtain the odds for a non-star category, we follow the same process. First, we substitute 0 for StatusCatStarAll to get β_0 as the linear predictor. Then we take the parameter estimate that we got for β_0 and exponentiate it.

The odds ratio is then the odds for the star category divided by the odds for a non-star category. Mathematically, this is equivalent to e^{β_1}, so we can divide the two values we just calculated, or we can simply take the parameter estimate that we got for β_1 and exponentiate it.

Question: Odds Ratio (Setup)

Odds Ratio Estimate				
Effect	Unit	Point Estimate	95% Wald Confidence Limits	
StatusCatStarAll (Star vs Non-star)	1	1.562	1.441	1.693



Question: Odds Ratio

The odds ratio of 1.562 for **StatusCatStarAll (Star vs Non-star)** indicates which of the following? (Hint: 1/1.562=0.6402)

a. Individuals with star category had 1.562 times odds of response than individuals without star category.
b. Individuals with star category had 56.2% higher odds of response than individuals without star category.
c. Individuals with non-star category had 35.97% lower odds of response than individuals with star category.
d. All the above.

Answer: Odds Ratio

The odds ratio of 1.562 for **StatusCatStarAll (Star vs Non-star)** indicates which of the following? (Hint: 1/1.562=0.6402)

a. Individuals with star category had 1.562 times odds of response than individuals without star category.

b. Individuals with star category had 56.2% higher odds of response than individuals without star category.

c. Individuals with non-star category had 35.97% lower odds of response than individuals with star category.

d. All the above.

ANSWER: d.

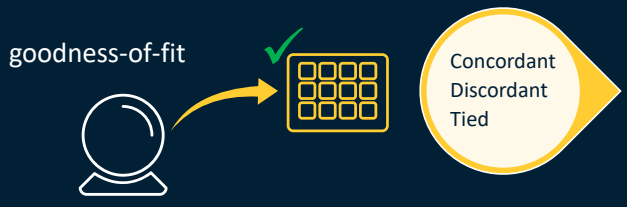
An odds ratio of 1.562 means that the odds are multiplied by 1.562 for each one-unit increase in the predictor. Another way to state the effect is to use the formula100*(Odds Ratio-1). This produces the percent increase in the odds for a unit increase in the predictor. An odds ratio of 1.562 corresponds to a 56.2% increase in the odds for a unit increase in X. If the odds ratio for Star = 1.562, then the odds ratio for Non-Star = 1/1.562 = 0.64. Then 100*(Odds Ratio-1)=-35.97% change in odds for unit increase in X.

Assessing the Model Fit

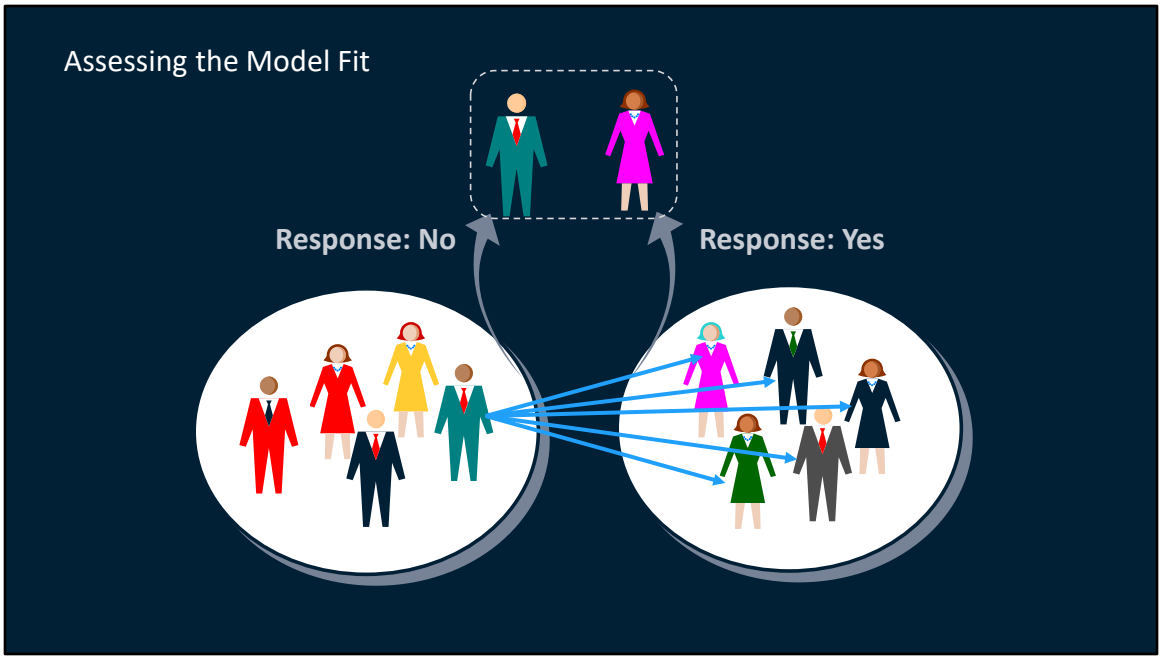
Assessing the Fit of a Logistic Regression Model

logistic regression model

$$\text{logit}(p_i) = \ln(\text{odds}) = \beta_0 + \beta_1 X_i$$



If a logistic regression model predicts its own data accurately, then we say that the model fits the data well. Counting concordant, discordant, and tied pairs is a way to assess how well the model predicts its own data and therefore how well the model fits. In general, you want a high percentage of concordant pairs and low percentages of discordant and tied pairs.



To find concordant, discordant, and tied pairs, we compare everyone who had the outcome of interest against everyone who did not.

SAS creates two groups of observations, one for each value of **Response**. One group contains all the observations in which the value of **Response** is 0, meaning the non-responders. The other group contains all the observations in which the value of **Response** is 1, the responders.

Each member in each group is paired to every member in the other group. One such pair is shown here. SAS then selects every pair and then determines whether each pair is concordant, discordant, or tied. Let's look at each type of pair.

Assessing the Model Fit

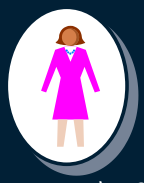
Concordant Pair

A pair is *concordant* if the observation with the outcome has a **higher** predicted outcome probability than the observation without the outcome.

Response: No

P(Response) = 0.32

Response: Yes



P(Response) = 0.42

The actual sorting agrees with the model.
This is a **concordant** pair.

Suppose that a pair consists of a woman who responded to the campaign and a man who did not. The man who did not respond to the campaign has a lower predicted probability of response, .32, than the woman who responded to the campaign, 0.42. A pair is concordant when the observation with the event, in this case, response=Yes, has a higher predicted probability of having the event than the observation without the event. That is, the model assigns a higher probability of being a responder than to the non-responder. When the model sorts the pair correctly, as in this case, the pair is concordant.

Assessing the Model Fit

Discordant Pair

A pair is *discordant* if the observation with the outcome has a **lower** predicted outcome probability than the observation without the outcome.

Response: No

P(Response) = 0.42

Response: Yes

P(Response) = 0.32

The actual sorting disagrees with the model.
This is a **discordant** pair.

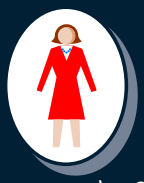
Let's look at another pair. This pair compares a man who responded to the campaign and a woman who did not. From our model, we know that the non-responder woman have a higher predicted probability, 0.42, of being a responder, than the man who actually responded to the campaign, 0.32. However, our model did not sort this pair correctly. A pair is discordant if the observation with the desired outcome has a *lower* predicted probability than the observation without the outcome.

Assessing the Model Fit

Tied Pair

A pair is *tied* if it is neither concordant nor discordant. The predicted outcome probabilities are the **same**.

Response: No



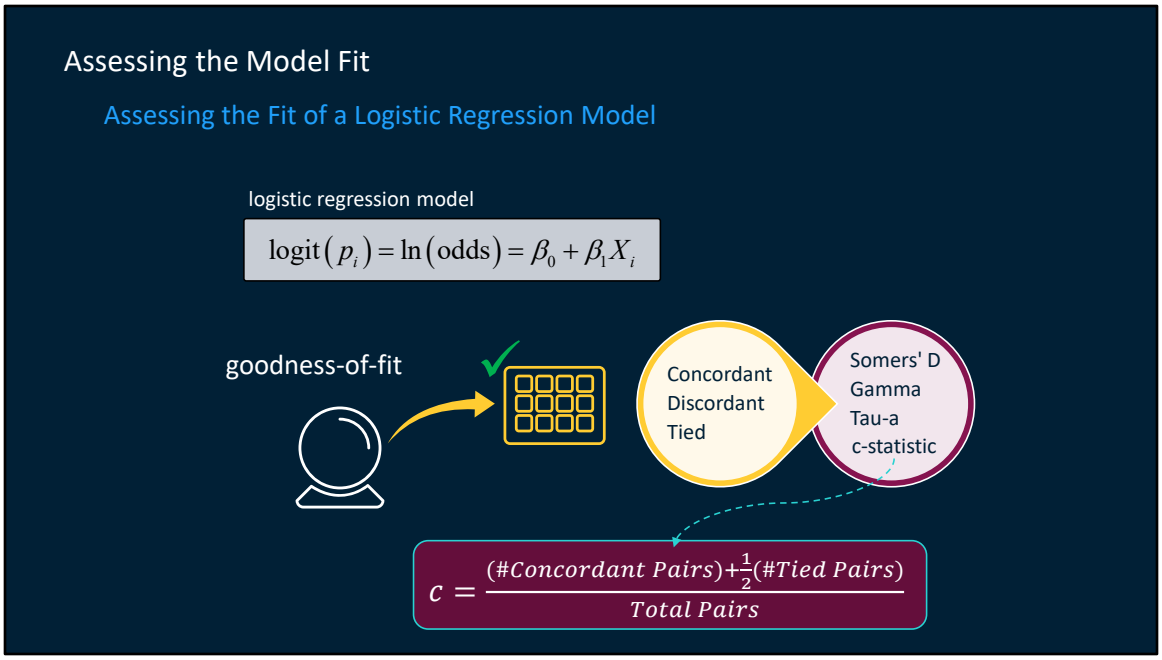
P(Response) = 0.42

Response: Yes

P(Response) = 0.42

The model cannot distinguish between the two.
This is a **tied** pair.

The last pair compares two women, one who responded to the campaign and the other who did not. According to our model, both have a predicted probability of response as 0.42, so our model cannot distinguish between them. A pair is tied if it is neither concordant nor discordant, that is, the probabilities are the same.



Several rank correlation indices are also computed from the numbers of concordant, discordant, and tied pairs of observations: Somers' D, Gamma, Tau-a, and c. In general, a model with higher values for these indices has better predictive ability than a model with lower values.

The c value is the most commonly used. The c, concordance statistic, estimates the probability of an observation with the event having a higher predicted probability than an observation without the event. The c value is calculated as the number of concordant outcomes plus one-half times the number of ties divided by the total number of pairs. The range of possible values is 0.5 to 1.0, where 1.0 is perfect prediction. It can be shown that the area under the ROC curve for a model equals the c-statistic. Thus, the c-statistic also called the ROC index.

Question: Comparing Pairs for Model Assessment

While you compare pairs for model assessment, which of the following statements is True?

a. Concordant pairs indicate predictions that match with what has been observed.
b. Discordant pairs indicate predictions that do not match with what has been observed.
c. Tied pairs suggest your model is not able to differentiate well between events and non-events.
d. All the above.

Answer: Comparing Pairs for Model Assessment

While you compare pairs for model assessment, which of the following statements is True?

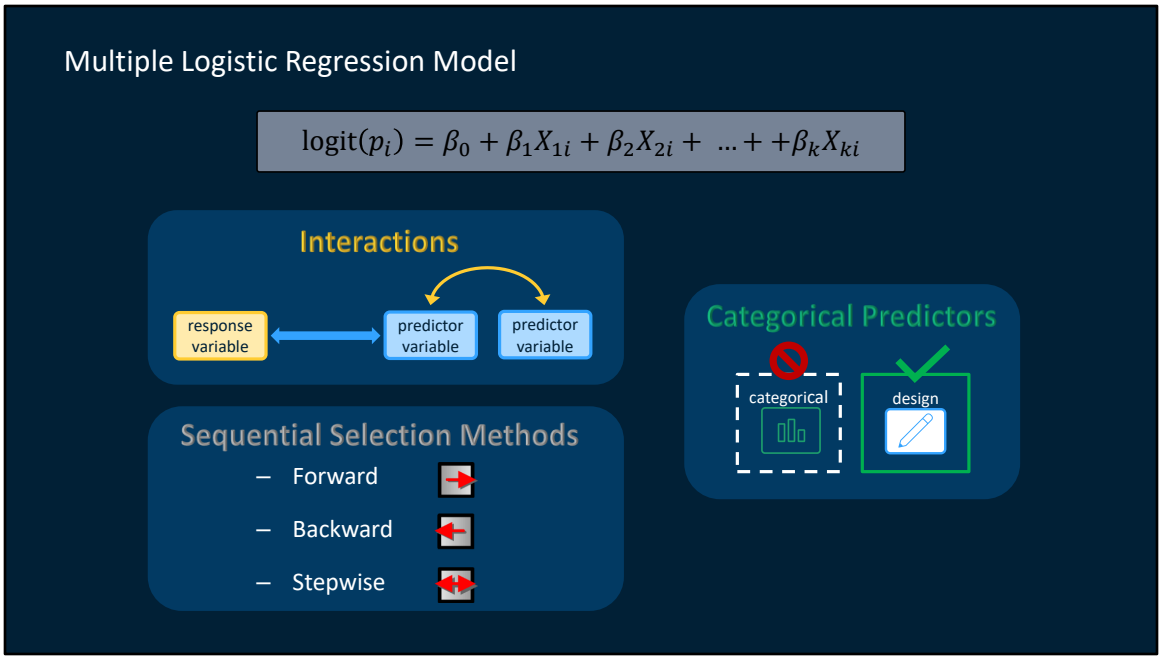
a. Concordant pairs indicate predictions that match with what has been observed.

b. Discordant pairs indicate predictions that do not match with what has been observed.

c. Tied pairs suggest your model is not able to differentiate well between events and non-events.

d. All the above.

ANSWER: d.

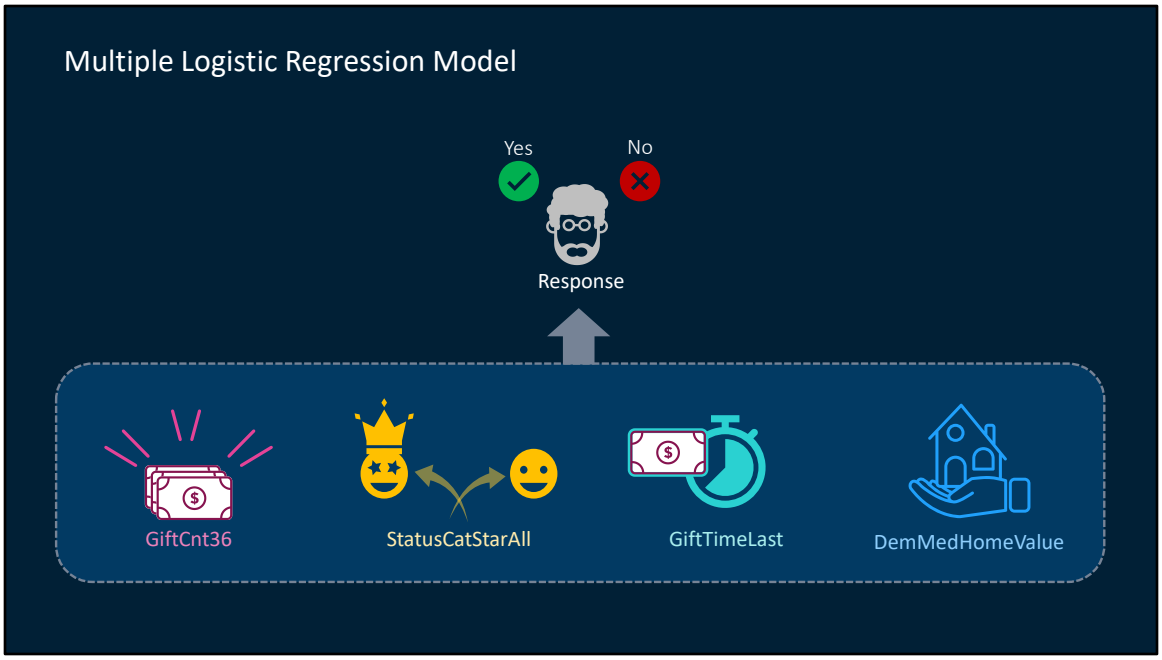


Just as in linear regression, if you have many predictors in logistic regression, it is called multiple logistic regression model. A multiple logistic regression model with *k* predictors is shown here. The number of parameters in the logistic model consider the intercept and the number of predictors.

Just as in linear regression, if you suspect that there are interactions between predictors, you can fit a more complex logistic regression model by including interaction effects. Remember, an interaction is present when the effect of one variable on the outcome depends on, or changes, due to another variable.

Just as in linear regression, a set of design variables, also known as dummy variables, is created for categorical predictors that represent the information in each classification variable. There are various methods of parameterizing the classification variables.

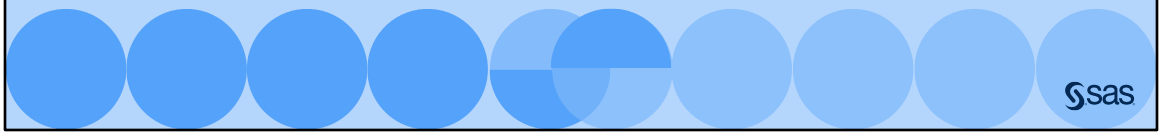
Just as in linear regression, you have similar model selection options in logistic regression that commonly includes forward, backward, and stepwise methods.



Now, let's fit a logistic regression model using some continuous and categorical predictors. You already saw that Response and StatusCatStarAll have a significant association. In addition to this, you use three more continuous variables: Gift count 36 months, Gift time last, and Dem Median home value to fit a multiple logistic regression model. However, you can always use the available model selection methods to select inputs for model building.

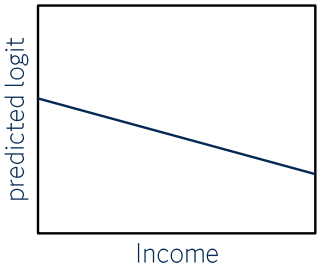
Demo: Fitting a Multiple Logistic Regression Model

To explore this learning resource in greater detail, please visit the course page on learn.sas.com. There, you'll find comprehensive course materials, tutorials, and updates to support your learning journey.



Question: Default and Income

Here's a graph of the logit of the predicted probability for defaulting on a loan by the customer's income.



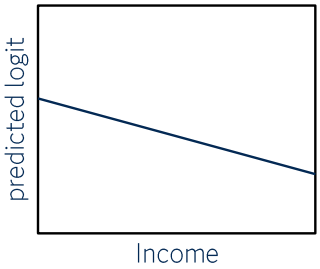
As a customer's income increases, the probability of defaulting on a loan decreases.

a. True

b. False

Answer: Default and Income

Here's a graph of the logit of the predicted probability for defaulting on a loan by the customer's income.

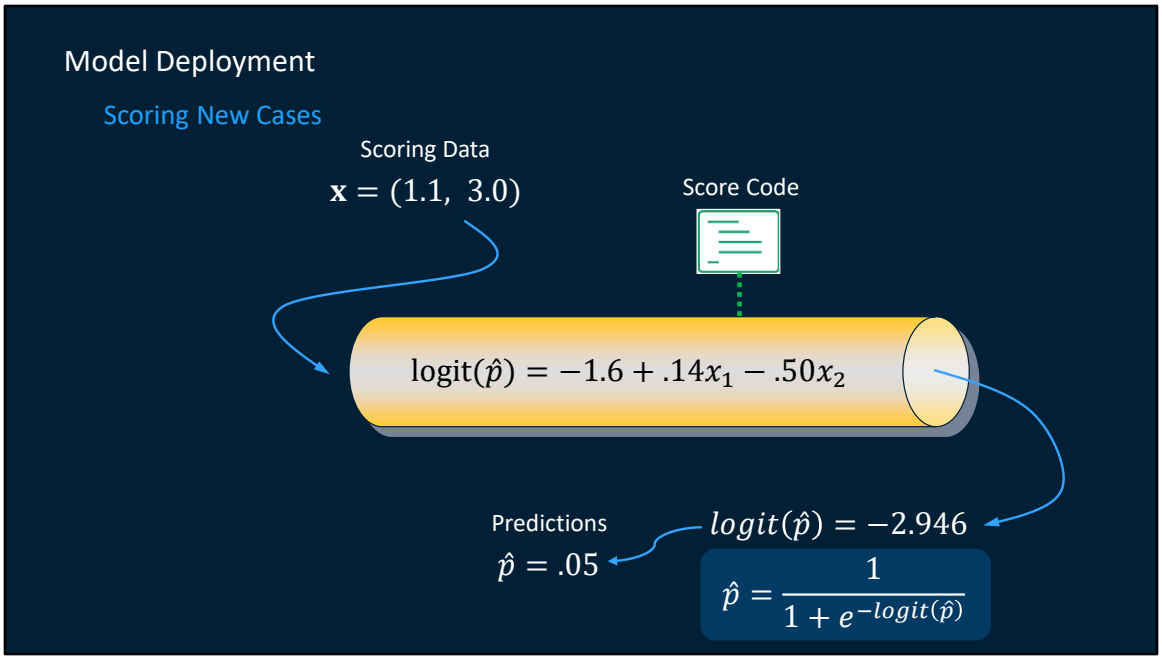


As a customer's income increases, the probability of defaulting on a loan decreases.

a. True
b. False

The correct answer is True.
Notice that as a customer's income increases, the probability of defaulting on a loan decreases.

4.4 Model Deployment



The overriding purpose of predictive modeling is to score new cases.

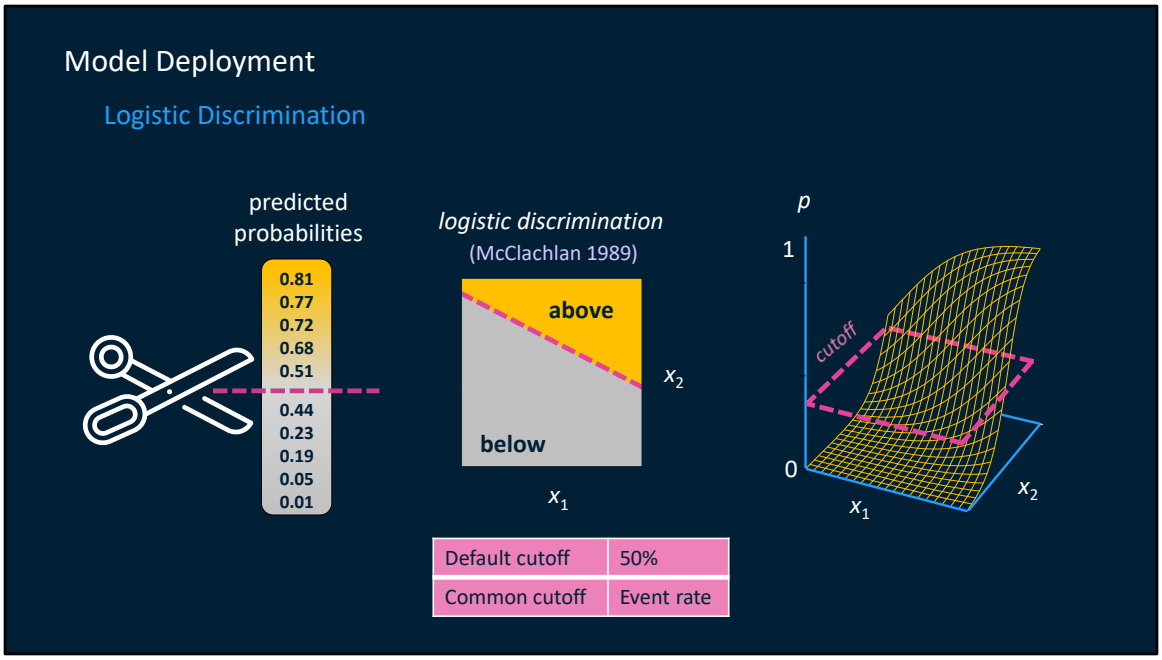
Predictions can be made by simply plugging in the new values of the inputs from the scoring data.

For scoring, the model is first translated into another format (typically, score code).

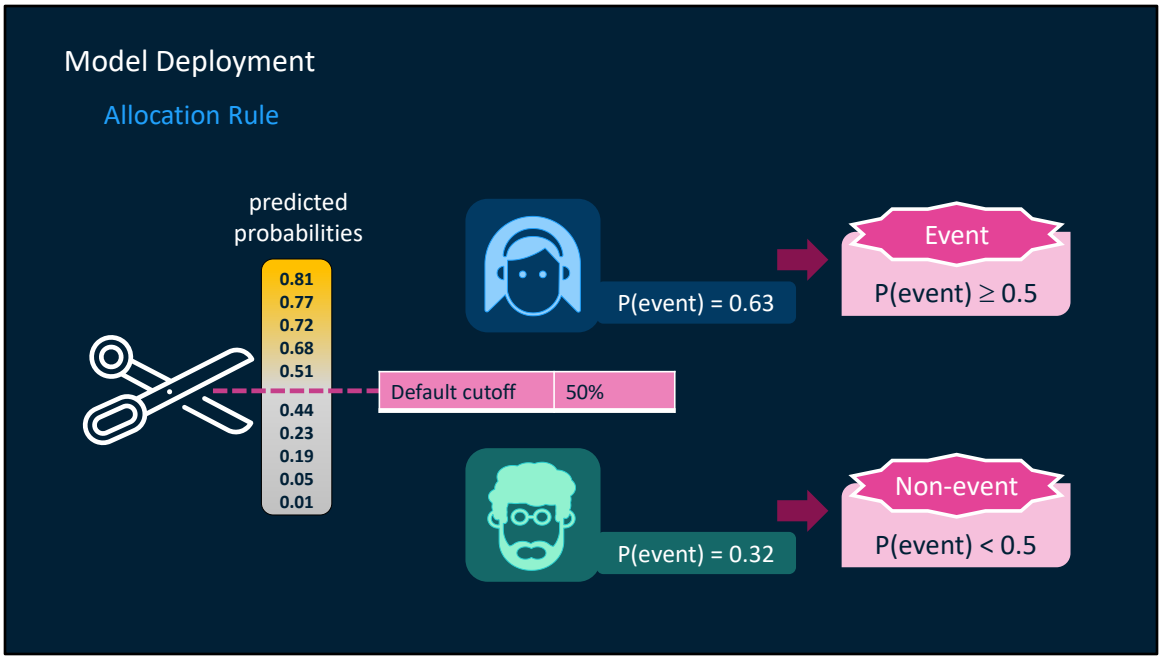
The logit equation produces a ranking or logit score.

You can obtain a prediction estimate using a straightforward transformation of the logit score, the logistic function. The *logistic function* is simply the inverse of the logit function. You can obtain the logistic function by solving the logit equation for *p*.

The most common predictions are the predicted probabilities.



In supervised classification, the ultimate use of logistic regression is to allocate cases to classes. This is more correctly termed logistic discrimination An allocation rule is merely an assignment of a cutoff probability, where cases above the cutoff are allocated to class 1 (event) and cases below the cutoff are allocated to class 0 (non-event). The standard logistic discrimination model separates the classes by a linear surface ((hyper)plane). The decision boundary is always linear. Determining the best cutoff is a fundamental concern in logistic discrimination. A logical cutoff is 50%, which is the default in SAS Viya, but typically a much better way to choose a cutoff is to use the percentage of events in the original data.



Thus, assuming a default cutoff of 50%, if your model predicts an event probability of 0.63 for someone, she would be predicted as an event, because the predicted probability is greater than 0.5. Whereas if someone is predicted with an event probability of 0.32 by your model, he would be predicted as a non-event, because the predicted probability is smaller than 0.5.

Demo: Scoring a Logistic Regression Model

To explore this learning resource in greater detail, please visit the course page on learn.sas.com. There, you'll find comprehensive course materials, tutorials, and updates to support your learning journey.

Question: Allocation Rule

In the previous demonstration, if you decide to have an allocation rule based on a 0.45 cutoff, which of the following cases would be predicted as a response? (Select all that apply.)

	DemMedHomeValue	P_Response
a.	$114,300	0.4543596915
b.	$51,100	0.4359083123
c.	$65,000	0.4175748543
d.	$59,200	0.6024474572

Answer: Allocation Rule

In the previous demonstration, if you decide to have an allocation rule based on a 0.45 cutoff, which of the following cases would be predicted as a response? (Select all that apply.)

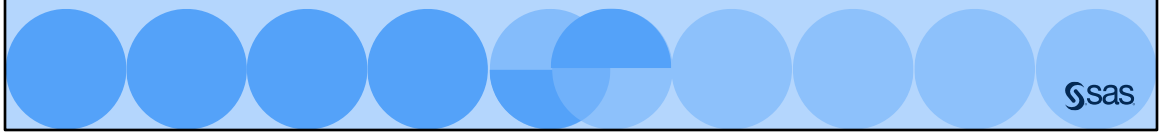
	DemMedHomeValue	P_Response
a.	$114,300	0.4543596915
b.	$51,100	0.4359083123
c.	$65,000	0.4175748543
d.	$59,200	0.6024474572

Correct answer: a and d

Because the predicted probabilities (0.4543 and 0.6024) of response are greater than 0.45 (cutoff), they would be predicted as response. The other two have predicted probabilities (0.4359 and 0.4175) smaller than 0.45, thus predicted as non-response.

Lesson Summary

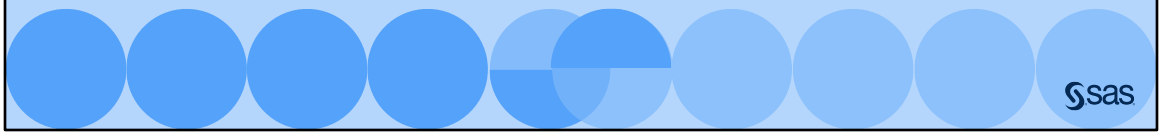
To explore this learning resource in greater detail, please visit the course page on learn.sas.com. There, you'll find comprehensive course materials, tutorials, and updates to support your learning journey.



Lesson 5: Statistical Foundations of Machine Learning

Lesson Overview

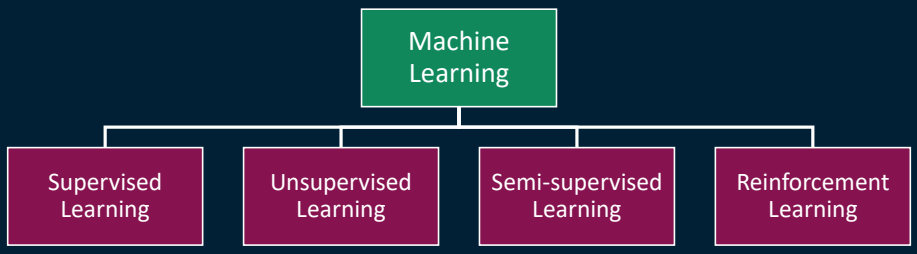
To explore this learning resource in greater detail, please visit the course page on learn.sas.com. There, you'll find comprehensive course materials, tutorials, and updates to support your learning journey.



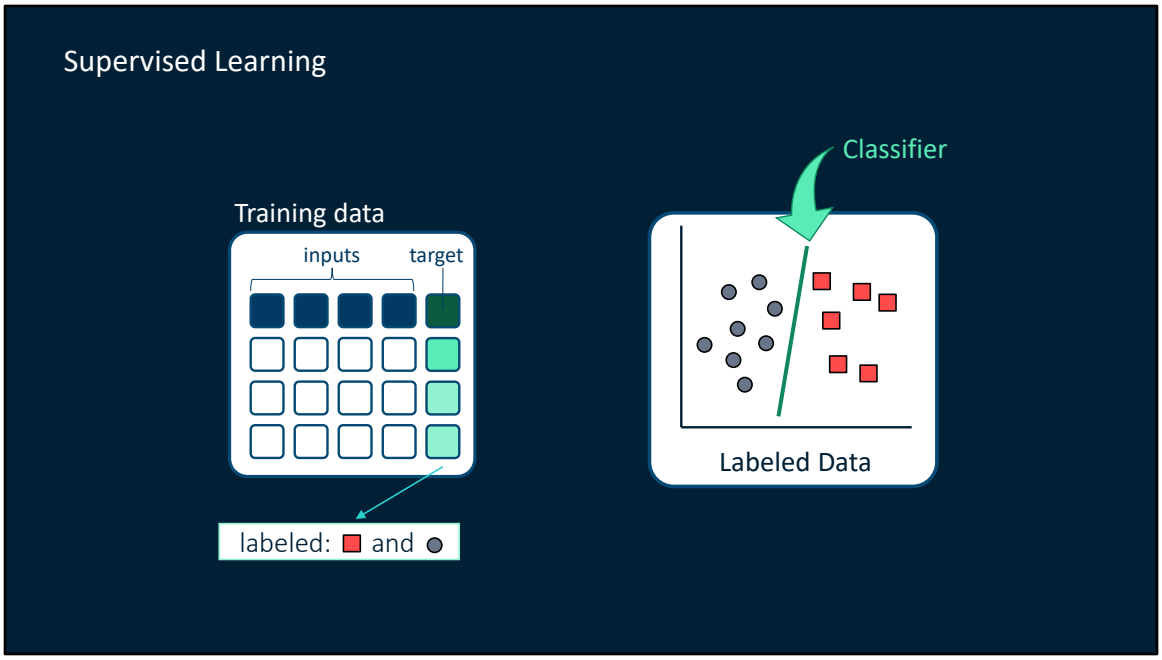
5.1 Overview of Machine Learning

Machine Learning

- Nature of the business problem?
- Type and volume of data?

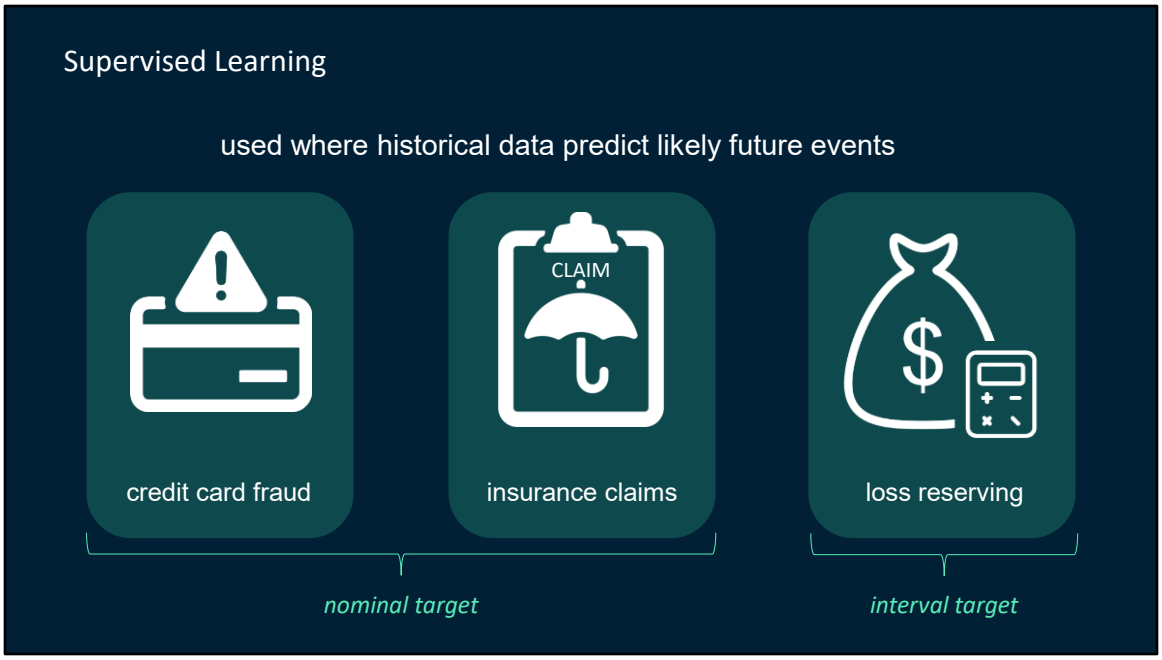


Depending on the nature of the business problem being addressed, and the type and volume of data, there are different approaches to machine learning. Two of the most widely used machine learning methods are supervised learning and unsupervised learning, but there are also other methods of machine learning. This lesson provides an overview of the most popular types of machine learning: supervised learning, unsupervised learning, semi-supervised learning, and reinforcement learning.

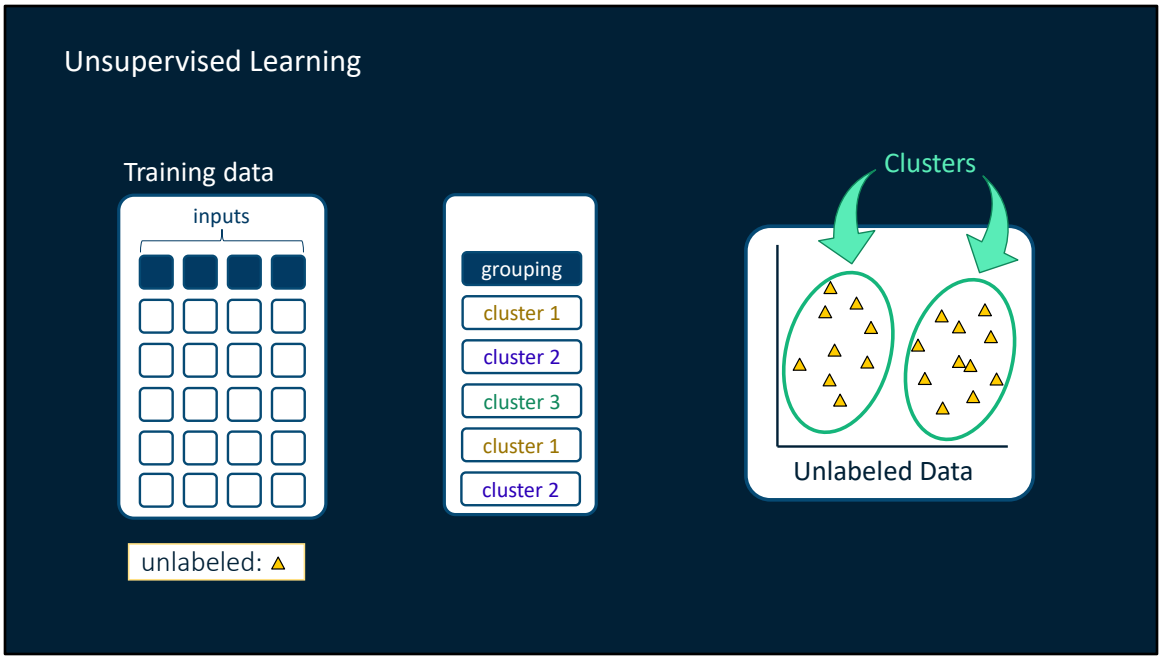


Supervised learning algorithms are trained using labeled examples, such as an input where the desired target is known. Here the value of the target is used as the label for an observation in the table. The learning algorithm receives a set of inputs along with the corresponding correct targets. Target variable could either be categorical or continuous. The algorithm learns by comparing its actual target with predicted target to find errors. It then modifies the model accordingly and produces a classifier.

For example, here you have a binary target: red squares and gray circles, and a linear classifier is represented here with a green line that differentiates red squares with gray circles.



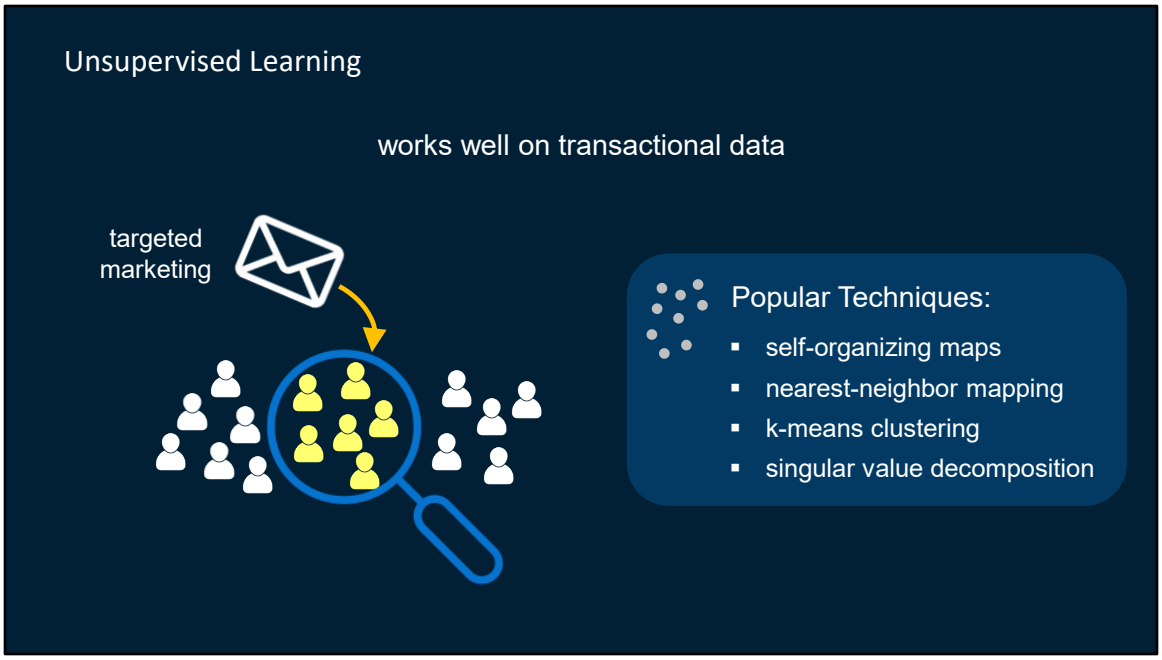
Supervised learning is commonly used in applications where historical data predict likely future events. For example, supervised learning can anticipate when credit card transactions are likely to be fraudulent or which insurance customer is likely to file a claim or a loss reserve, which is an estimate of an insurer's liability from future claims that it will have to pay out on.



Unsupervised learning is used against data that have no historical labels. The system is not told the "right answer." The algorithm must figure out what is being shown. The goal is to explore the data and find some structure within.

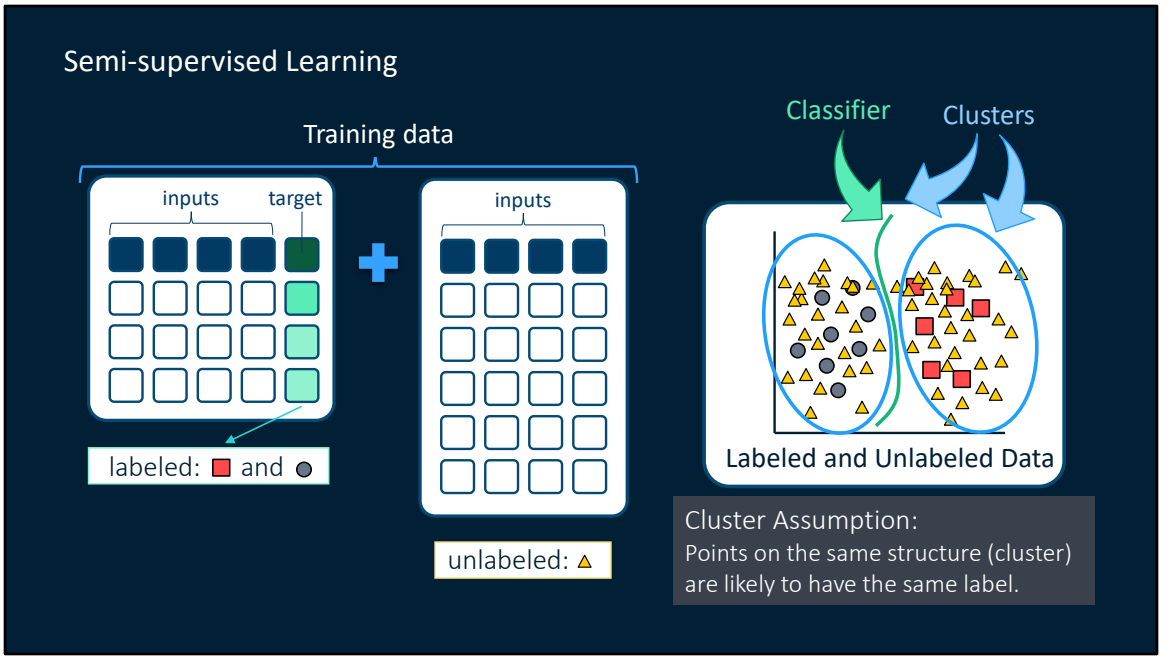
Unsupervised learning commonly attempts to group training data set cases based on similarities in input variables. It is a data reduction method because an entire training data set can be represented by a small number of groups. The groupings are known as clusters or segments, and they can be applied to other data sets to classify new cases.

For example, here you have all the observations represented by yellow triangles as no labels are available. A cluster structure is represented here with green ellipses that group the yellow triangles into two mutually exclusive clusters.



Unsupervised learning works well on transactional data. For example, it can identify segments of customers with similar attributes who can then be treated similarly in marketing campaigns.

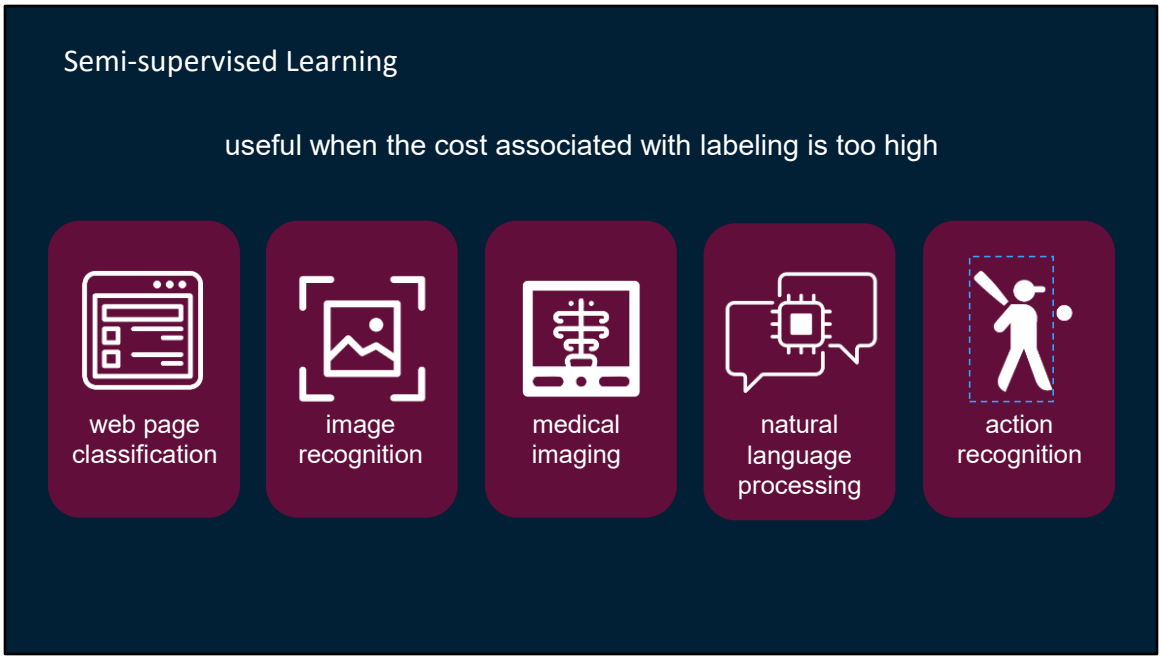
Popular techniques include self-organizing maps, nearest-neighbor mapping, k-means clustering, and singular value decomposition. These algorithms are also used to segment text topics, recommend items, and identify data outliers.



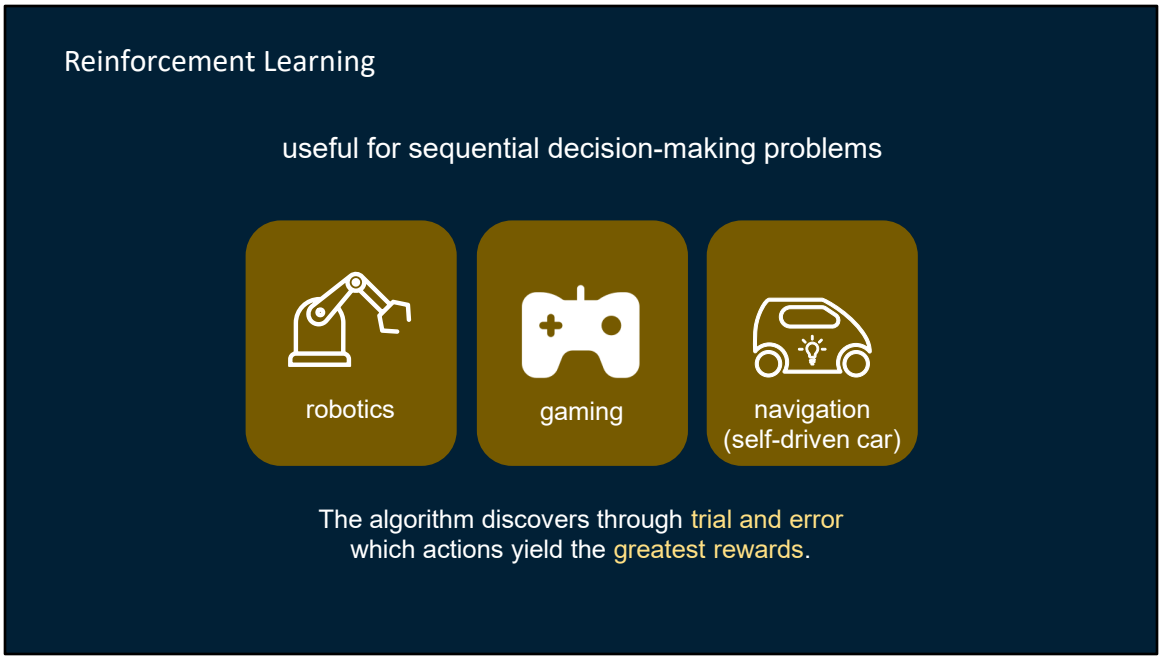
Semi-supervised learning is used for the same applications as supervised learning. It uses both labeled and unlabeled data for training, typically a small amount of labeled data with a large amount of unlabeled data (because unlabeled data are less expensive and take less effort to acquire). It's kind of a marriage between supervised and unsupervised learning. This type of learning can be used with methods such as classification, regression, and prediction. Input variables are used for distance calculation.

For example, here you have a binary target: red squares and gray circles available for a small number of observations and most of the other observations represented by yellow triangles as no labels are available. Based on the labeled data, a nonlinear classifier is represented here with a green curve that differentiates red squares with gray circles. A cluster structure is represented here with blue ellipses that groups the yellow triangles into two mutually exclusive clusters.

Points on the same structure (typically referred to as a *cluster*) are likely to have the same label. This assumption is often called the *cluster assumption.*



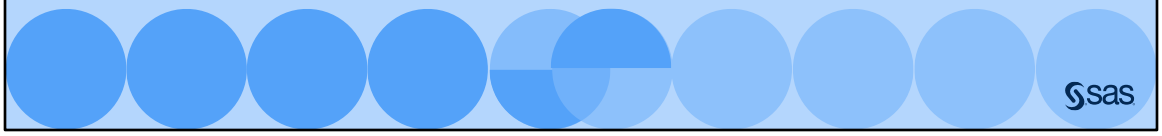
Semi-supervised learning is useful when the cost associated with labeling is too high to allow for a fully labeled training process. The semi-supervised learning algorithm has numerous applications, including web page classification, image recognition, medical imaging, natural language processing, and action recognition.

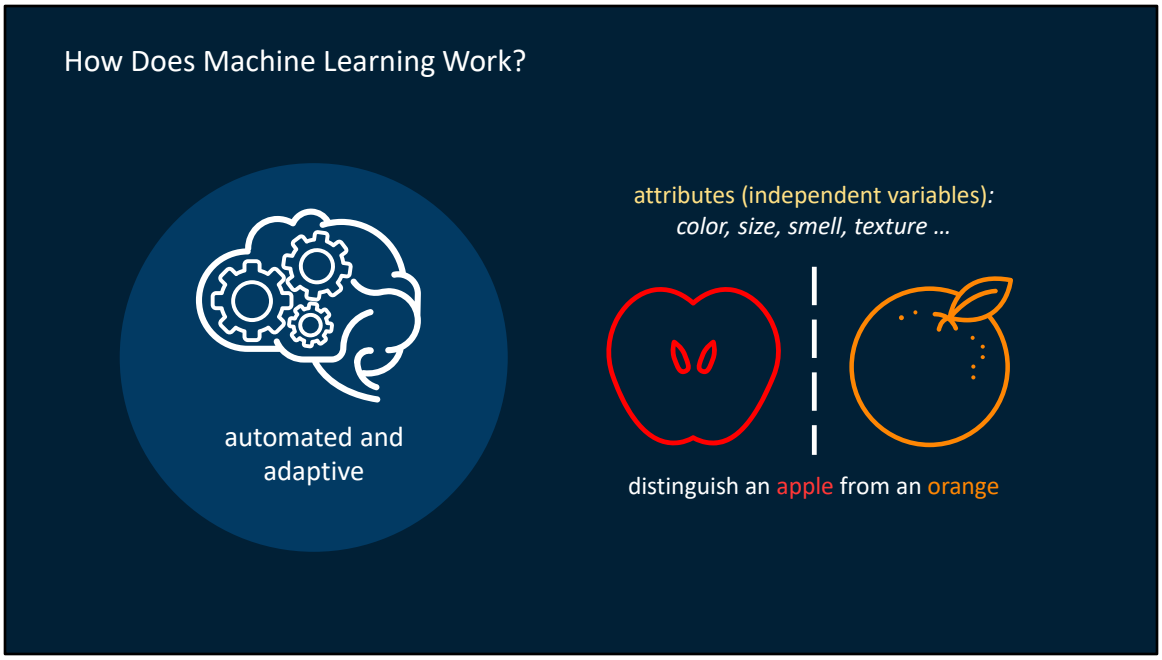


Reinforcement learning is often used for robotics, gaming, and navigation like a self-driven car. With reinforcement learning, the algorithm discovers through trial and error which actions yield the greatest rewards. Thus, it is mainly utilized for sequential decision-making problems.

More about Reinforcement Learning

To explore this learning resource in greater detail, please visit the course page on learn.sas.com. There, you'll find comprehensive course materials, tutorials, and updates to support your learning journey.



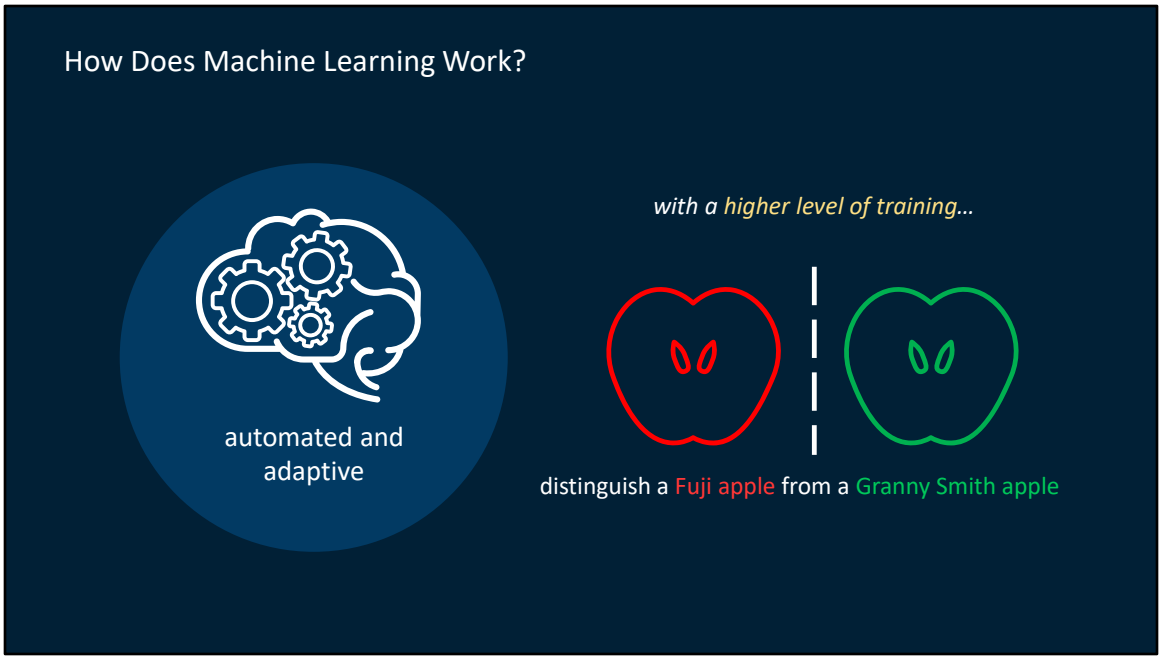


There are many different methods of machine learning. What they all have in common is that they are automated and adaptive. Machine learning techniques are sometimes modeled after the human brain.

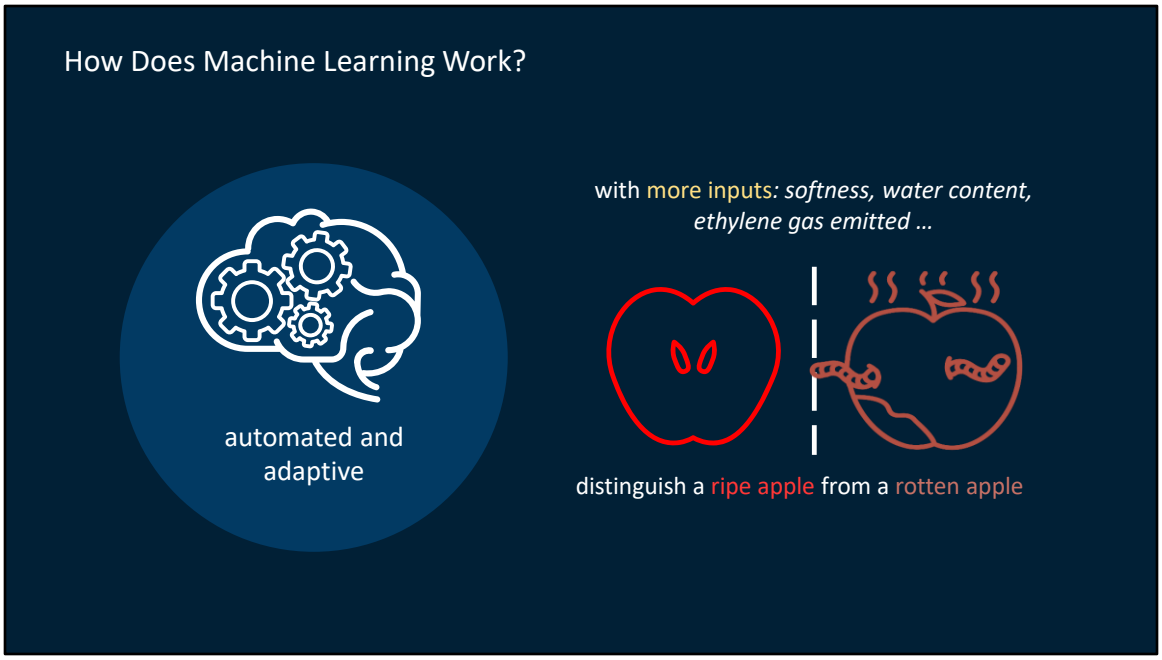
I like to use a simple example of distinguishing an apple from an orange. If I want to make an apple pie and I ask you to bring home apples from the market, you need to be able to tell an apple from an orange.

If you come home with oranges, well…best case scenario, there will be no dessert. Luckily, you can do this easily, because over time as a child, you learned to differentiate an apple from an orange based on color, size, smell, and texture. Someone would tell you that yes, that is an apple, or no, that is not an apple. This is called "supervised learning."

Color, size, smell, and texture are all "features" or "attributes" or as statisticians like to call them, independent variables. These features help determine the outcome. In this case, the outcome is binary. It's an apple, or it's not an apple. In your brain, these different attributes are weighted differently. For example, texture and color might be weighted higher than size because it is better at differentiating apples from oranges.



And likewise, with a higher level of training, I can distinguish a Granny Smith apple from a Fuji apple.



And given more inputs, like softness, water content, and ethylene gas emitted, I could determine not only if an apple is rotten, but how likely it is to become rotten even before it has become brown and squishy.

Question: Machine Learning Methods

Which of the following statements is true regarding machine learning approaches? (Select all that apply)

a. Unsupervised learning discovers underlying patterns based on input associations.
b. Supervised learning predicts the output based on input-target associations.
c. The input data in supervised learning is unlabeled whereas in unsupervised learning the data is labeled.
d. Reinforcement learning follows a trial-and-error method, and the data is not predefined.

Answer: Machine Learning Methods

Which of the following statements is true regarding machine learning approaches? (Select all that apply)

a. Unsupervised learning discovers underlying patterns based on input associations.

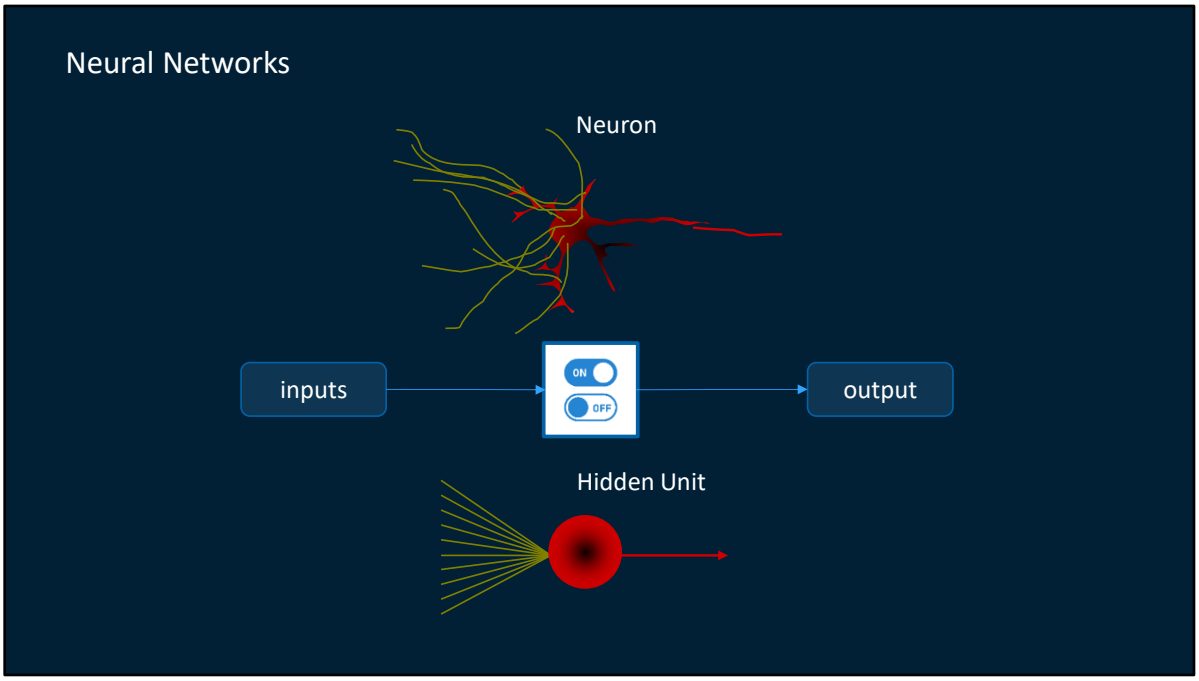
b. Supervised learning predicts the output based on input-target associations.

c. The input data in supervised learning is unlabeled whereas in unsupervised learning the data is labeled.

d. Reinforcement learning follows a trial-and-error method, and the data is not predefined.

ANSWER: a, b and d.

The input data in supervised learning is labeled whereas in unsupervised learning the data is unlabeled. Based on input associations, unsupervised learning discovers underlying data structures such as clusters, whereas based on association of inputs and target, supervised learning predicts the target. With reinforcement learning, the algorithm discovers through trial and error which actions yield the greatest rewards.

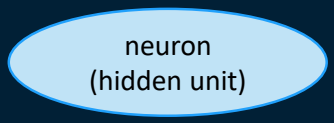


Neural networks, also known as artificial neural networks (ANNs), are at the heart of machine learning. Their name and structure are inspired by the human brain, mimicking the way that biological neurons signal to one another.

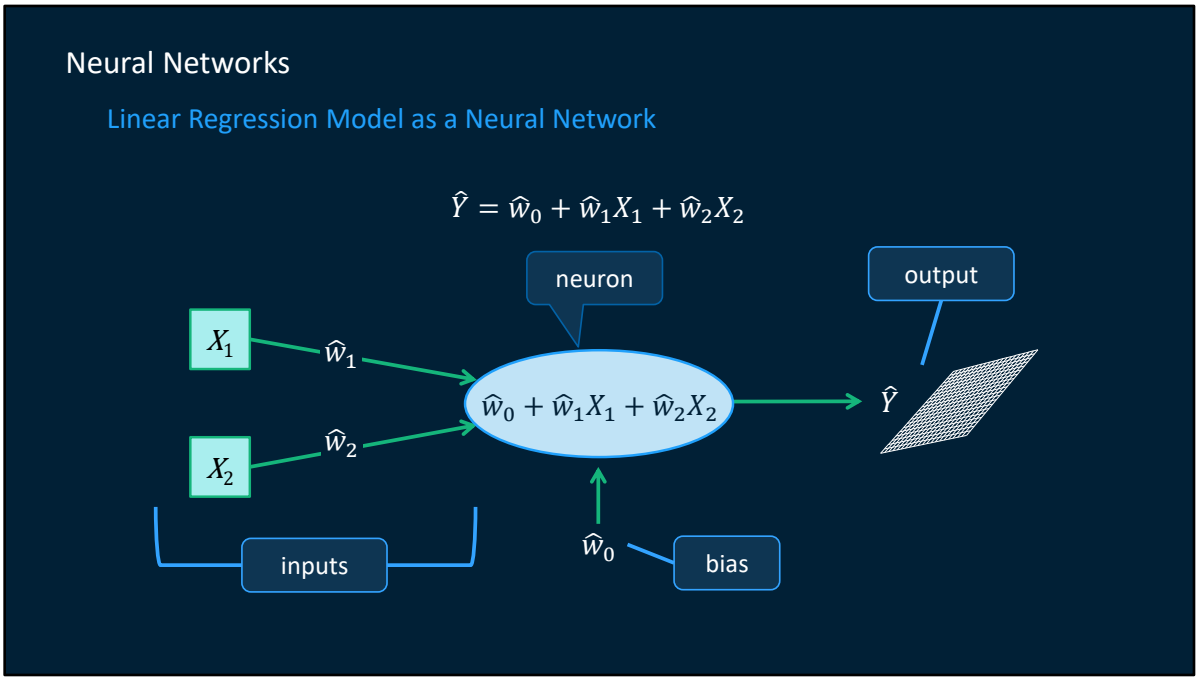
A neural network consists of many simple processing elements called neurons or hidden units. Each neuron is connected to other neurons by means of directed communication links, each with an associated weight. A neuron is nothing more than a switch with information input and output. The switch will be activated if there are enough stimuli of other neurons hitting the information input. Then at the information output, a pulse is sent to other neurons.

Neural Networks

What Is a Neuron?

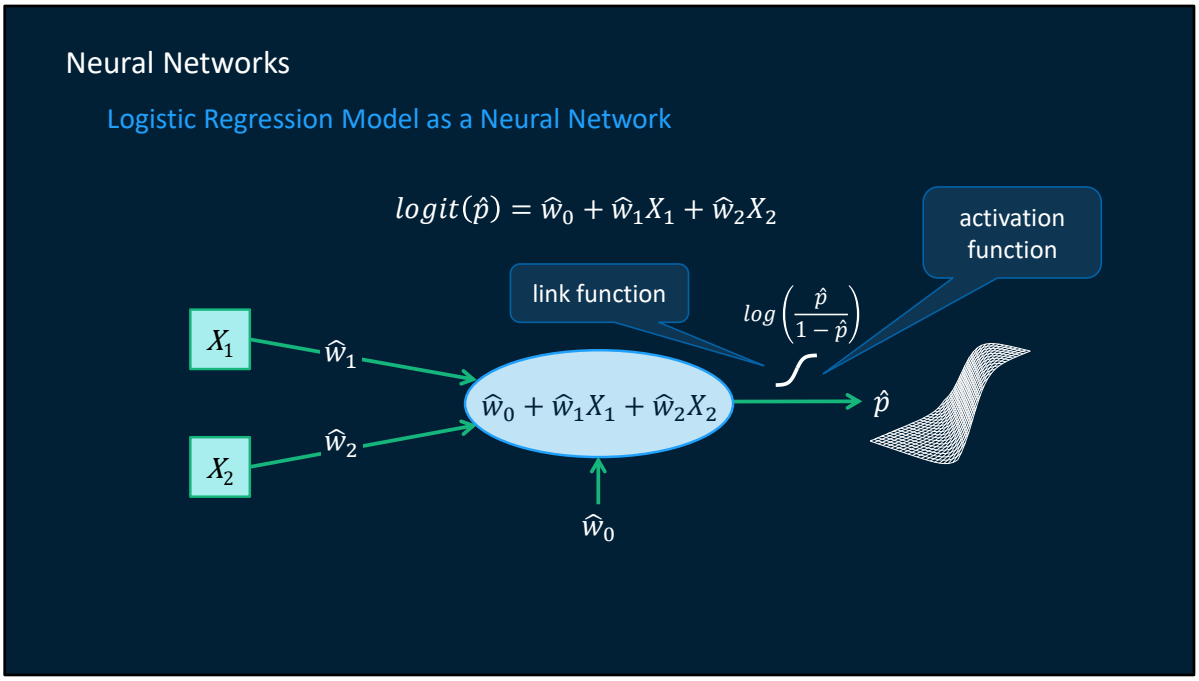


The basic unit of computation in a neural network is the *neuron*, often called a *hidden unit*. In order to understand neurons, let's look at regression models represented as neural networks.



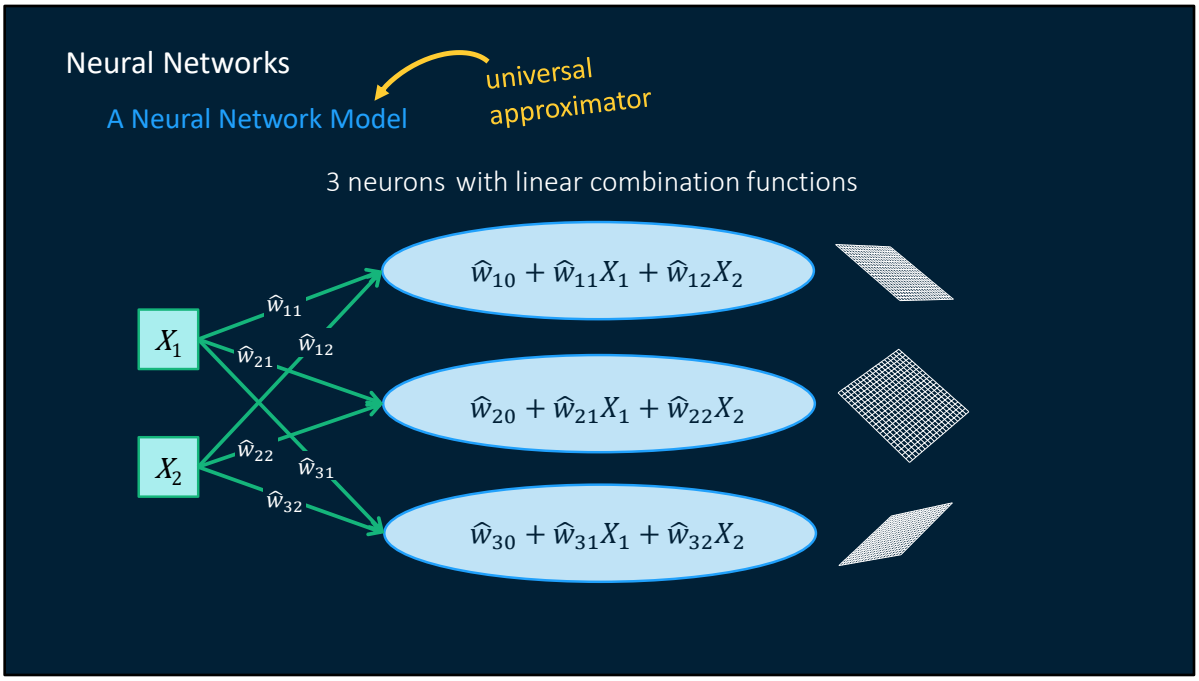
The classic linear regression equation is probably well known to you. Consider a simple linear regression equation with two inputs. The same can be represented in the form of a neural network.

In a neuron, typically, each input (x_i) is first weighted (w_1 and w_2) and then summed. To mimic a neuron's threshold functionality, a bias value (w_0) is added to the weighted sum, predisposing the neuron to give the output value. Here, the bias term is the y-intercept, and the weight associated with each input is the input's slope parameter. A neural network with no hidden layer, only one output neuron, and a linear function of x_i resembles a regression model. In this case, the model is easy to understand and to fit but has a downside – it can model only linear relationships!

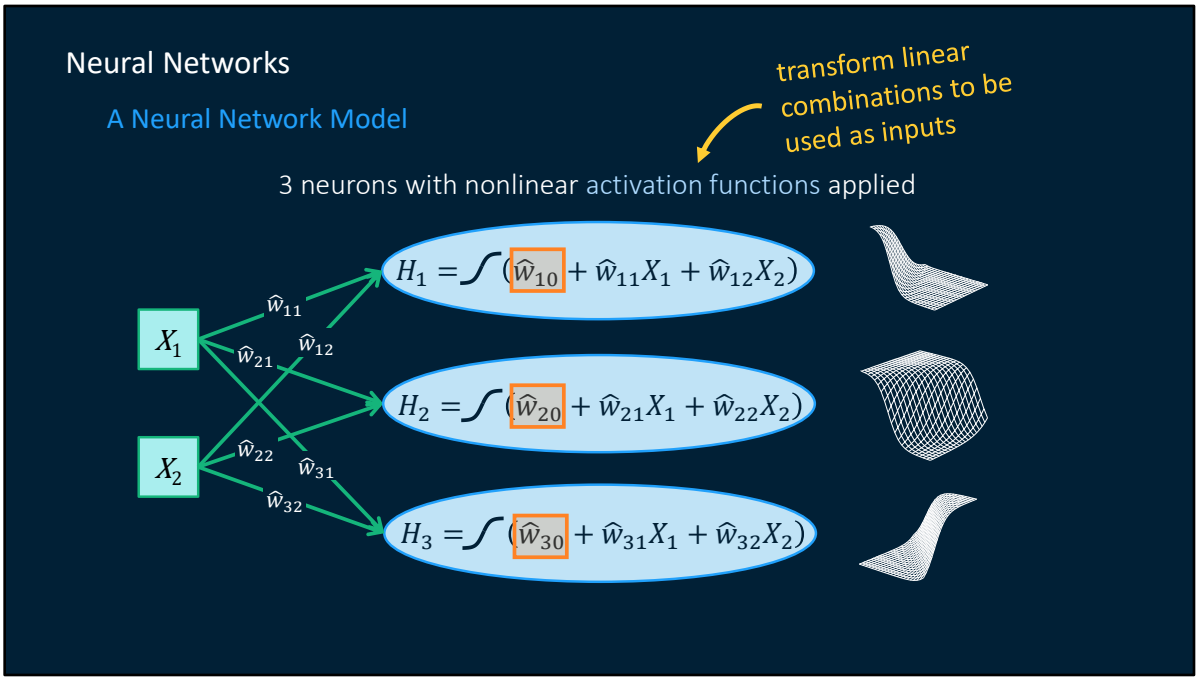


A little modification in the previous linear regression model is made. Represented here is a logistic regression model for two inputs X_1 and X_2 with a binary target. An output *activation function* is used to transform the output into a suitable scale for the expected value of the target. In statistics, this function is called the inverse *link function*. For binary targets, the logistic function is suitable because it constrains the output to be between zero and one.

Notice that the neuron is still linear, but as we make this value go through a non-linear function, the overall result is no longer linear. This model is still too simple and might not handle the assumed underlying complexity of the relationship between inputs and output. We can take a step further and deepen the model.

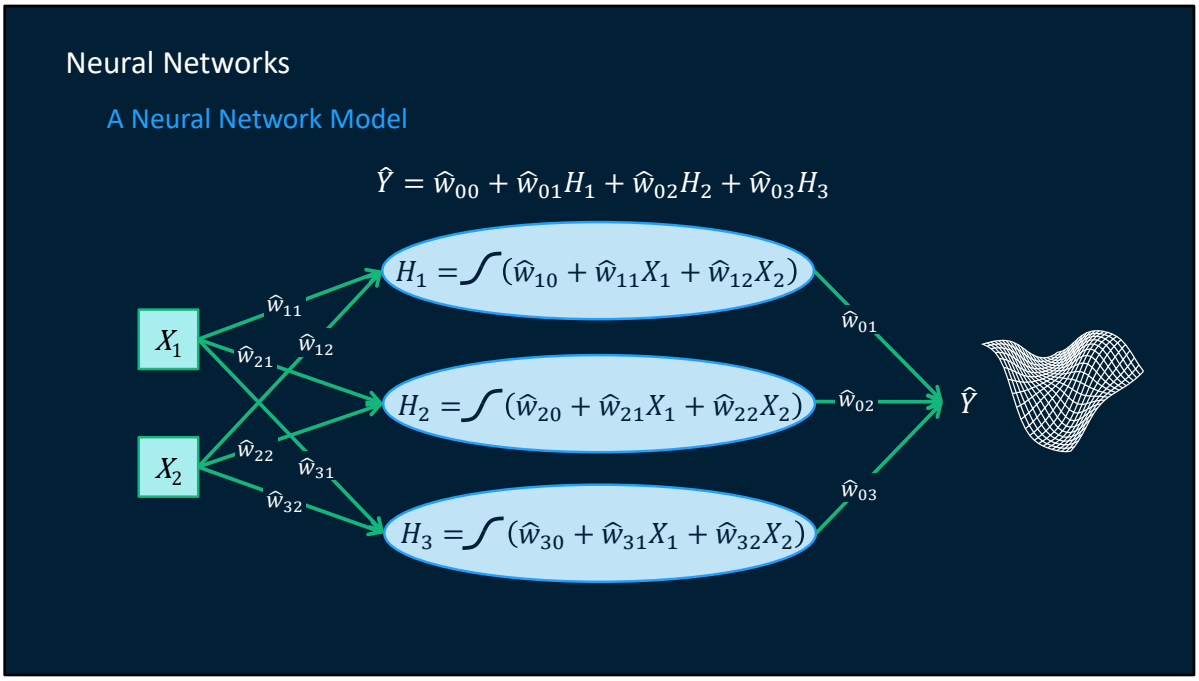


The foremost benefit of neural networks is their unlimited flexibility. A neural network is a *universal approximator*, which means that, with enough neurons and enough time, a neural network can model any input-output relationship, no matter how complex. Let's consider a simple example with three neurons. Each neuron receives a linear combination of input variables, but possibly with different weights. The weights from the linear combinations of the input variables are labeled in the diagram.



To capture progressively more nonlinearities, linear combinations in each neuron go through a nonlinear *activation function*. An activation function transforms the linear combinations and then converts them to another unit that can then use them as inputs. Different activation functions can be used to produce different nonlinear functions.

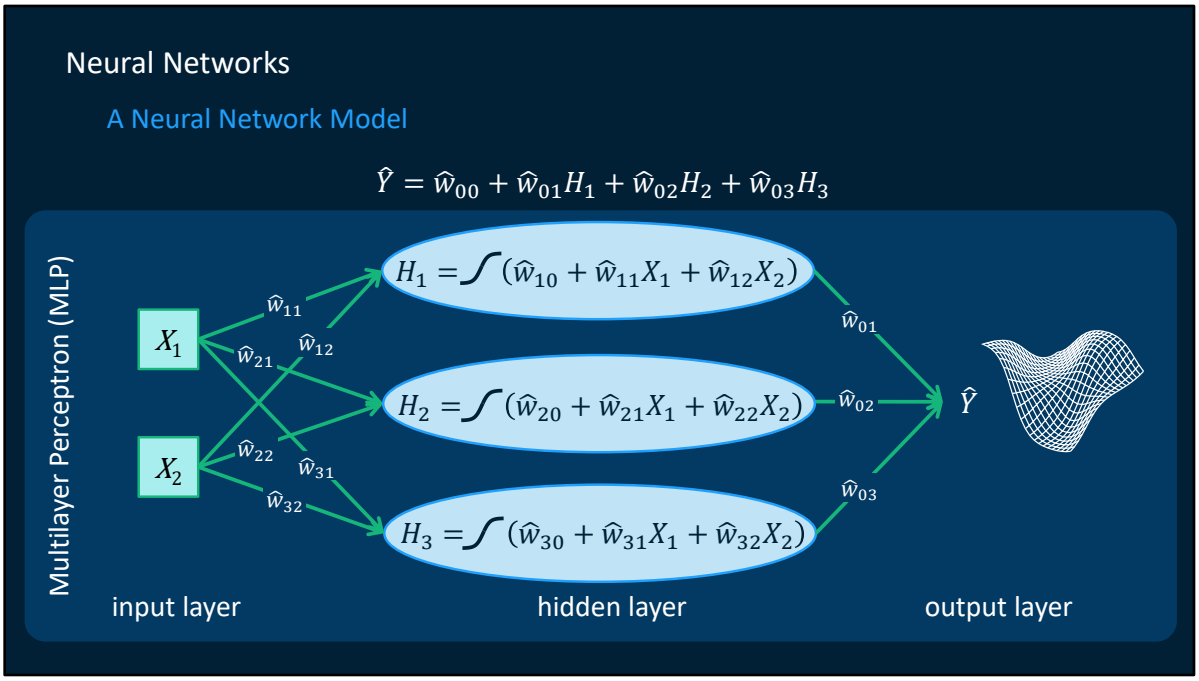
When there is a single hidden layer, the hidden units can be thought of as regressions applied to linear combinations of the input variables. In this example, an equation for each of the three hidden units is shown. In each equation, the first w-hat term is the bias term, which is not represented in the diagram. The second subscript on a bias term is always zero.



Finally, a neural network can be thought of as a regression model on a set of derived inputs, called *hidden units*. The functional form (or the prediction formula) for this neural network model is shown at the top.

Each linear combination of the input variables contains additional weights and a bias term. Remember that the lines in the neural network diagram represent the weights. The weights from the linear combinations of the input variables and the weights from the hidden units are now all labeled in the diagram. For example, the line connecting the hidden unit H1 to the input x2 represents the weight w-hat sub one two. And the line connecting the target to the hidden unit H1 represents the weight w-hat sub zero one. An output activation function is used to transform the output into a suitable scale for the expected value of the target to produce a nonlinear model.

To summarize, each neuron in a network transforms data using a series of computations: a neuron multiplies an initial value by some weight, sums results with other values coming into the same neuron, adjusts the resulting number by the neuron's bias, and then normalizes the output with an activation function.



Represented here is the most widely used neural network known as the *multilayer perceptron* (MLP). An MLP consists of an input layer, one or more hidden layers composed of hidden units, and an output layer.

Question: Neural Networks

Linear regression and logistic regression models can be viewed as neural networks with no hidden layers.

a. True
b. False

Answer: Neural Networks

Linear regression and logistic regression models can be viewed as neural networks with no hidden layers.

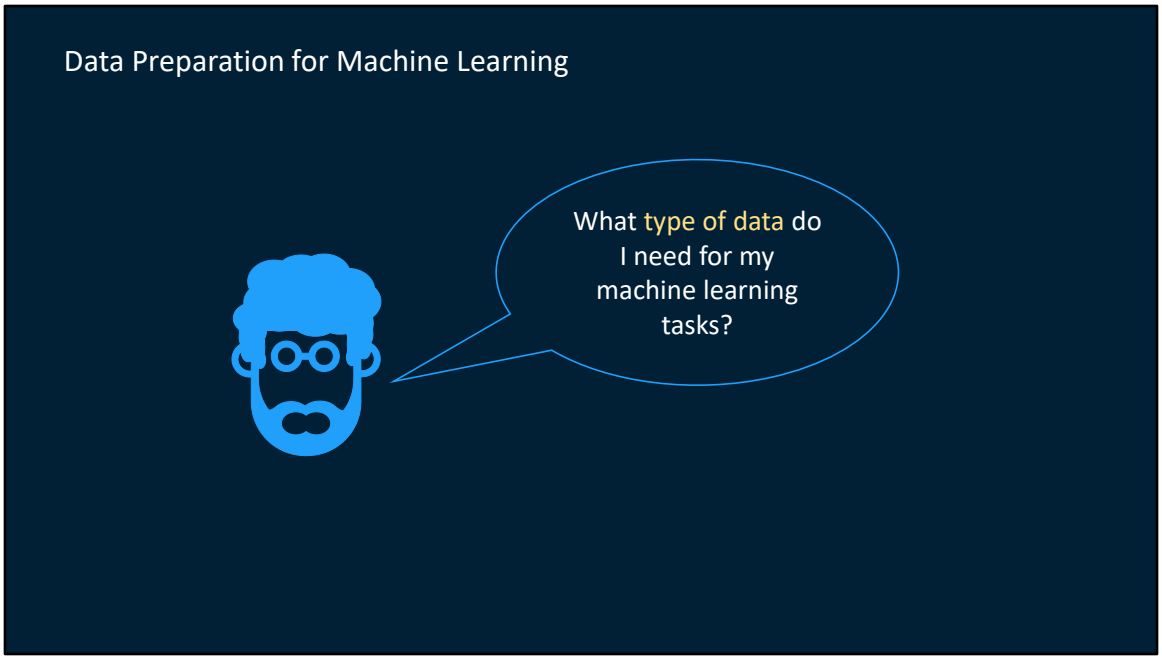
a. True
b. False

ANSWER: True

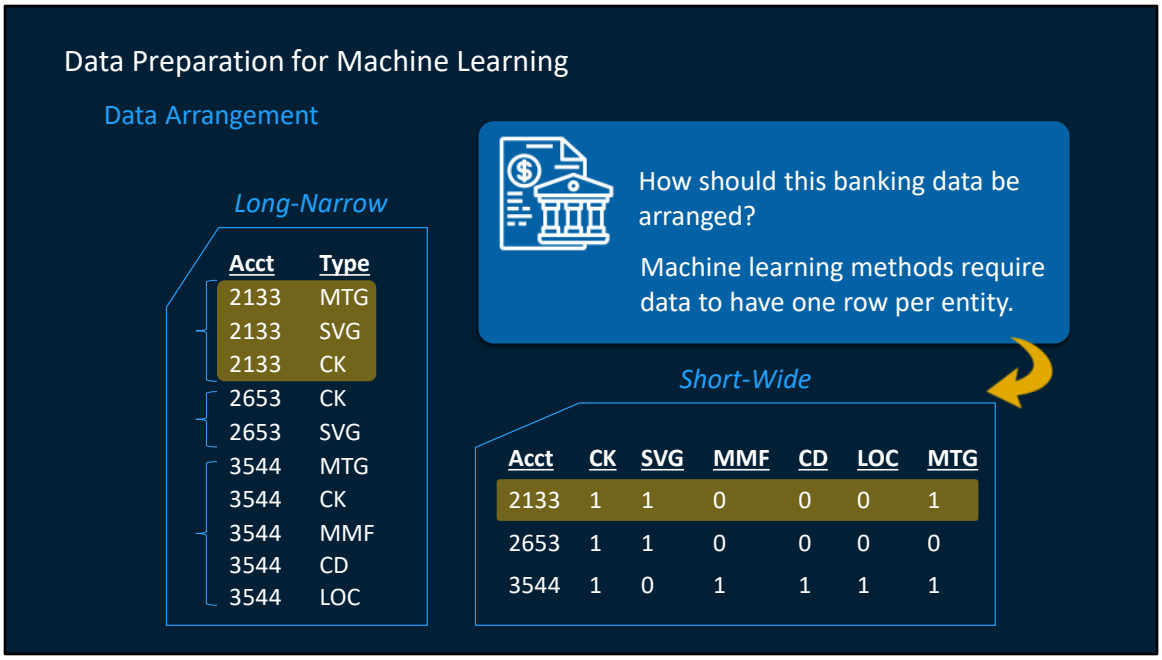
A neural network with no hidden layer, only one output neuron, and a linear function of input variables resembles a regression model.

More Information: Common Algorithms

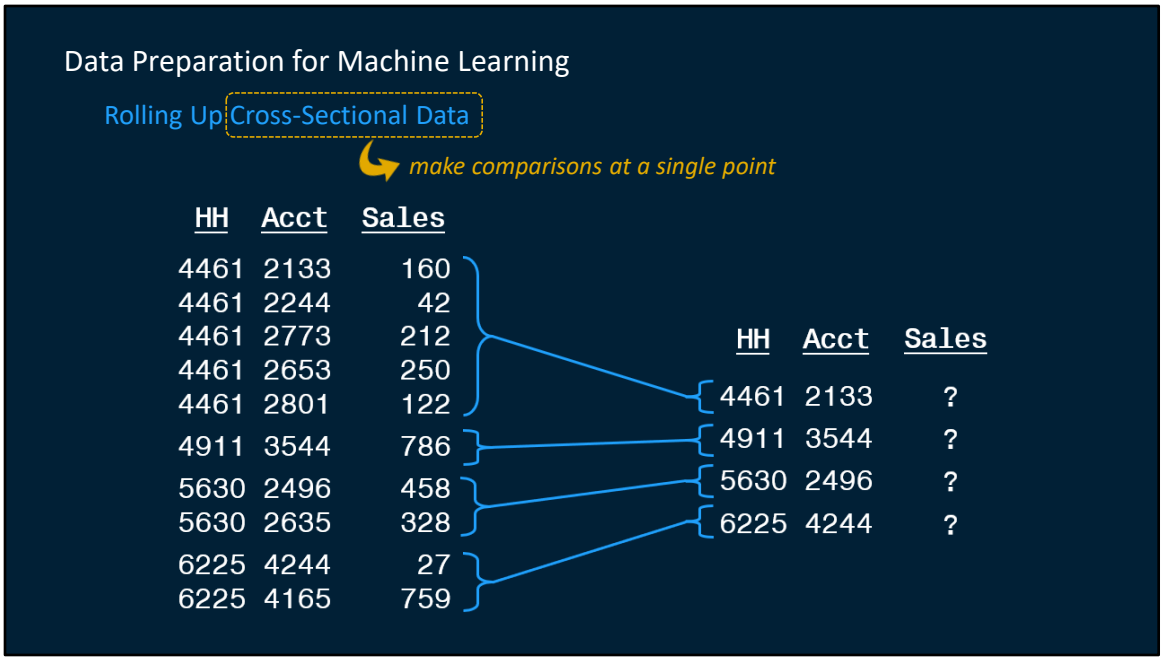
To explore this learning resource in greater detail, please visit the course page on learn.sas.com. There, you'll find comprehensive course materials, tutorials, and updates to support your learning journey.



So, by now you might be wondering: *what type of data do I need for machine learning tasks*?



The data often must be manipulated into a structure suitable for a particular analysis-by-software combination. For example, should this banking data be arranged with multiple rows for each account-product combination (long-narrow form) or with a single row for each account and multiple columns for each product (short-wide form)?For example, account ID 2133 has multiple rows in long-narrow table. It can be arranged in a short-wide form in a single row, but multiple columns of binary type. Machine learning methods require data to have one row per entity.



Business requirements often dictate rolling up accounts from a single household into a single record (case). This process usually involves creating new summary data. How should the sales figures for multiple accounts in a household be summarized? Using the sum, the mean, the variance, or all three?

Given here is an example of *cross-sectional* studies that make comparisons at a single point in time. Next is an example of *longitudinal* studies that make comparisons over time.

Data Preparation for Machine Learning

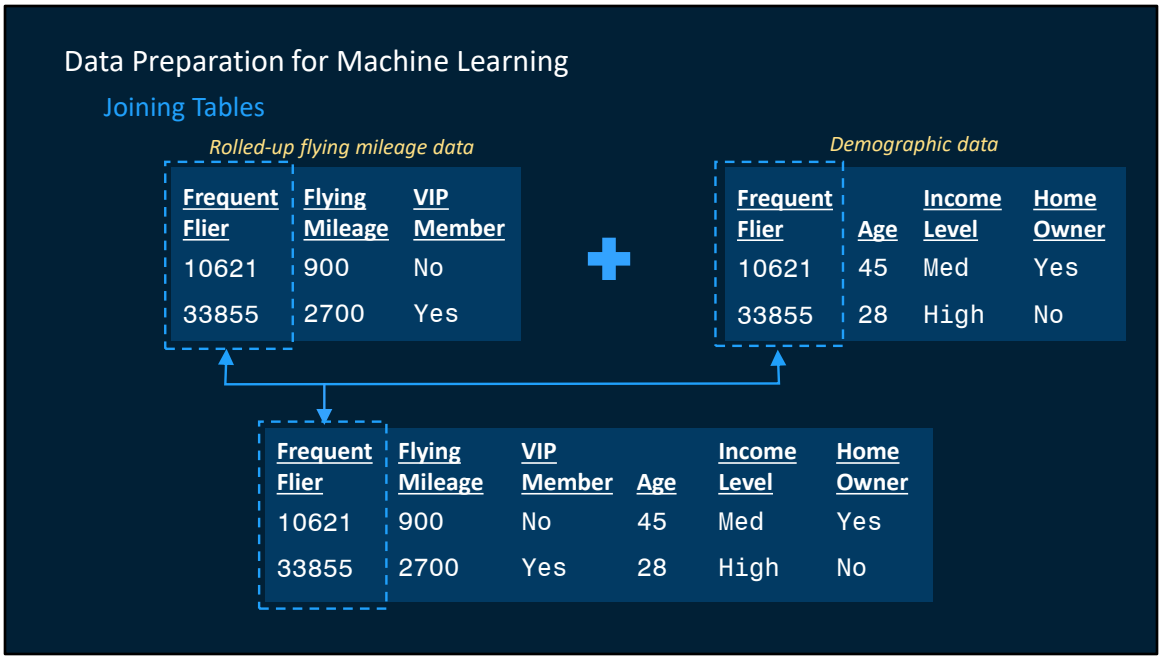
Rolling Up Longitudinal Data

make comparisons over time

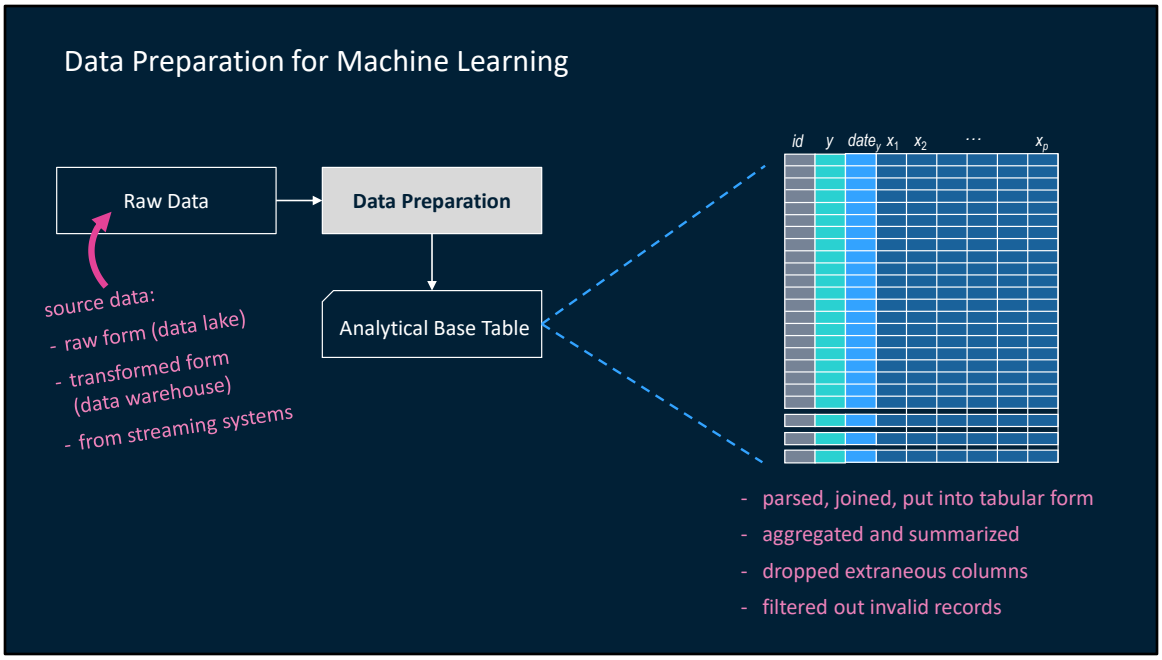
How should flying mileage be consolidated?

Frequent Flier	Month	Flying Mileage	VIP Member
10621	Jan	650	No
10621	Feb	0	No
10621	Mar	0	No
10621	Apr	250	No
33855	Jan	350	No
33855	Feb	300	No
33855	Mar	1200	Yes
33855	Apr	850	Yes

In some situations, it might be necessary to roll up longitudinal data into a single record for everyone. Longitudinal data, sometimes referred to as panel data, track the same sample at different points in time. For example, suppose an airline wants to build a prediction model to target current frequent fliers for a membership offer in the "Very Important Passenger" club. One record per passenger is needed for supervised classification. How should the flying mileage be consolidated if it is to be used as a predictor of club membership?



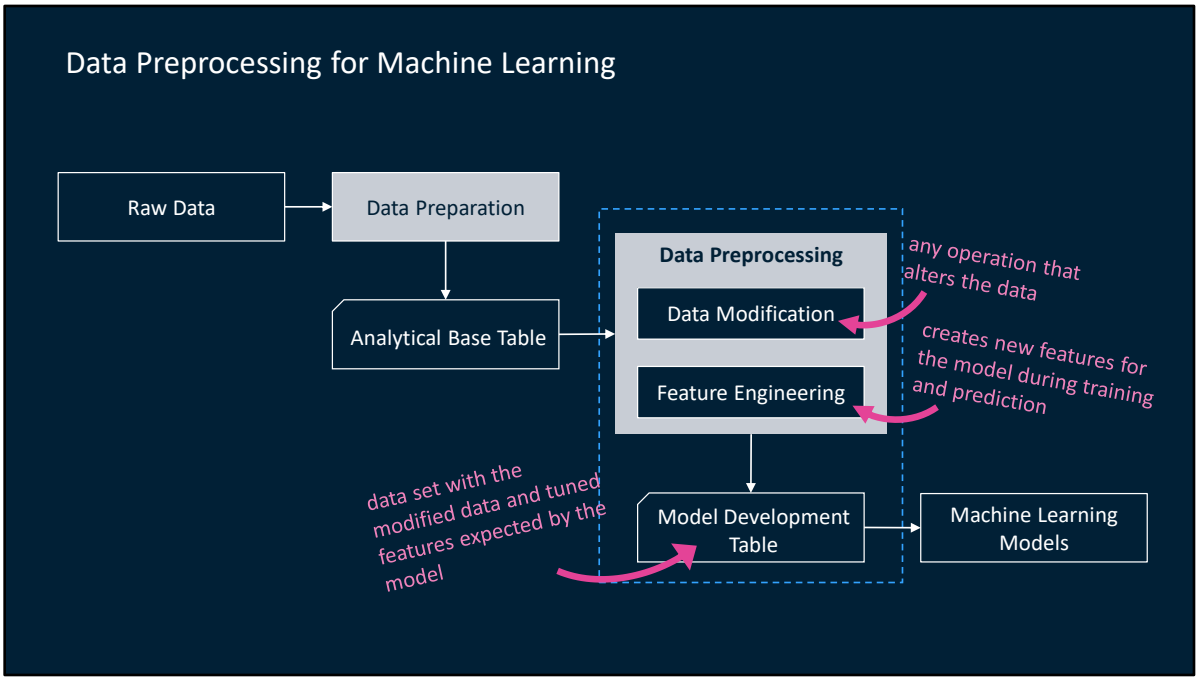
In addition to rolling-up, we need to join several tables to stitch together different types of data. For example, after flying mileage is consolidated as the sum of the four months, it can be joined with the demographic data. Frequent flyer column value can be used as a key to match the two tables.



As you've just seen in the previous examples, the data used for machine learning involves a grueling data preparation phase. Data preparation is the process of converting *raw data* into prepared data commonly known as *analytical base table*.

Raw data refers to the data in its source form, without any prior preparation for machine learning. This type of data might be in its raw form (in a data lake) or in a transformed form (in a data warehouse). In addition, data sent from streaming systems that eventually call machine learning models for predictions is data in its raw form only.

Analytical Base Table (ABT) is usually a wide table that refers to the data set in the form ready for your machine learning task. Data sources have been parsed, joined, and put into a tabular form. Data has been aggregated and summarized to the right granularity—for example, each row or an ID column in the data set represents a unique customer, and each column represents summary information for the customer, like the total spent in the last two months. In the case of supervised learning tasks, the target feature is present along with a target date, which is optional. Extraneous columns have been dropped, and invalid records have been filtered out.



Even when you get to the analytical base table, you are yet not ready to apply machine learning models. The data now need to be preprocessed for achieving certain modeling requirements to yield a model development table. The model development table refers to the data set with the modified data and tuned features expected by the model.

Data modification is a broad preprocessing category. Any operation that alters the data or data roles can be considered as a modification. *Feature engineering* is another broad preprocessing category that performs certain machine learning-specific operations on the columns in the prepared data set and creating new features for your model during training and prediction. Data preprocessing examples include case filtering, replacement of values, imputation, scaling numerical columns to a value between 0 and 1, truncating values, and one-hot-encoding categorical features.

Data preprocessing is discussed in detail in the next section.

Question: Data Preparation and Data Preprocessing

Which of the following statements is True regarding machine learning data? (Select all that apply.)

a. Source data yields a **raw table** without any prior preparation.
b. Data preparation yields an **analytical base table** in the form ready for your machine learning task, such as feature engineering.
c. Data preprocessing yields a **model development table** that is suitable to apply to machine learning models.
d. Machine learning data is generally clean and does not require any preparation and processing.

Answer: Data Preparation and Data Preprocessing

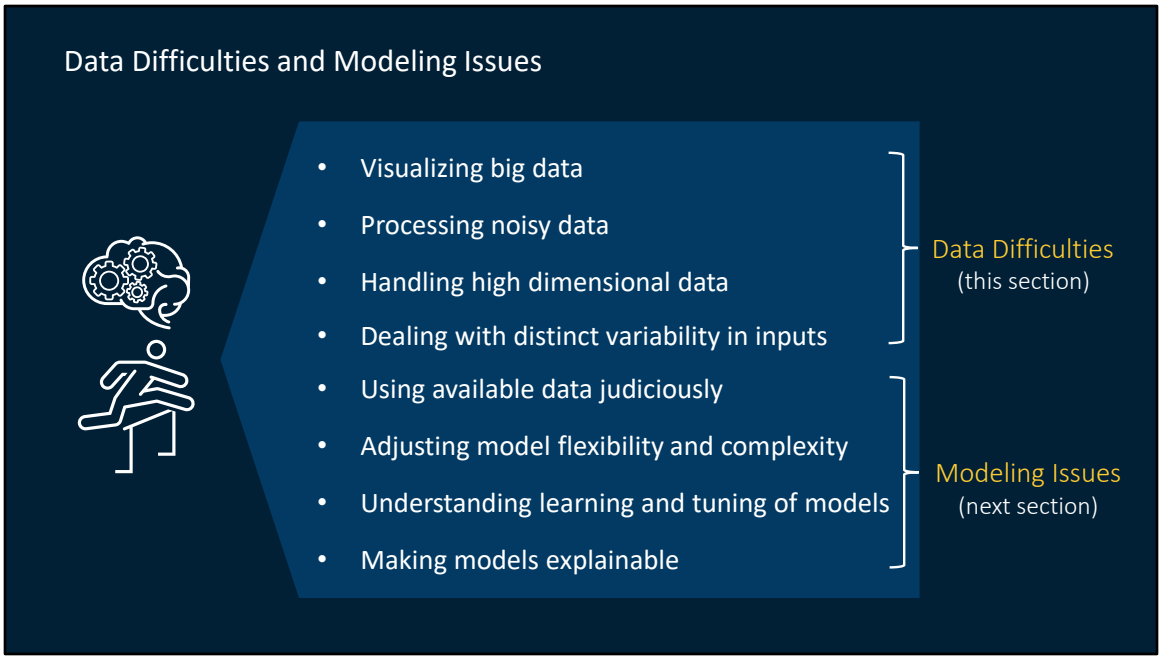
Which of the following statements is True regarding machine learning data? (Select all that apply.)

a. Source data yields a **raw table** without any prior preparation.
b. Data preparation yields an **analytical base table** in the form ready for your machine learning task, such as feature engineering.
c. Data preprocessing yields a **model development table** that is suitable to apply to machine learning models.
d. Machine learning data is generally clean and does not require any preparation and processing.

ANSWER: a, b, and c.

Raw table refers to the data in its source form. *Analytical base table* is usually a wide table that refers to the data set in the form ready for your machine learning task. The data now need to be preprocessed for achieving certain modeling requirements to yield a *model development table*. Machine learning data is generally dirty and thus involves grueling preparation and processing phases.

5.2 Data Pre-processing for Machine Learning Models



The benefits of machine learning translate to innovative applications that can improve the way processes and tasks are accomplished. However, despite its numerous benefits, there are still risks and challenges. Listed here are some of the challenges that we encounter in data mining and machine learning.

The first four: visualizing high-dimensional data, processing noisy data, handling too many variables, and dealing with distinct variability in inputs, are essentially related with machine learning data and will be discussed in this section of the lesson.

The remaining four: using available data economically, addressing model flexibility and complexity, finding best values of parameters and making models explainable, are related with core modeling and will be discussed in the next section of the lesson.

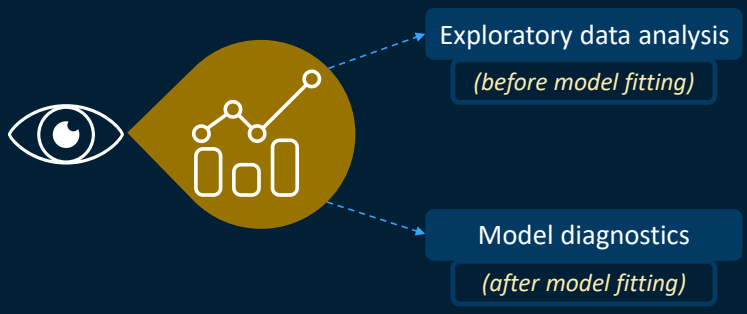
Data Visualization

- Visualizing big data
- Processing noisy data
- Handling high dimensional data
- Dealing with distinct variability in inputs
- Using available data judiciously
- Adjusting model flexibility and complexity
- Understanding learning and tuning of models
- Making models explainable

Let's begin our discussion of data difficulties with visualizing big data.

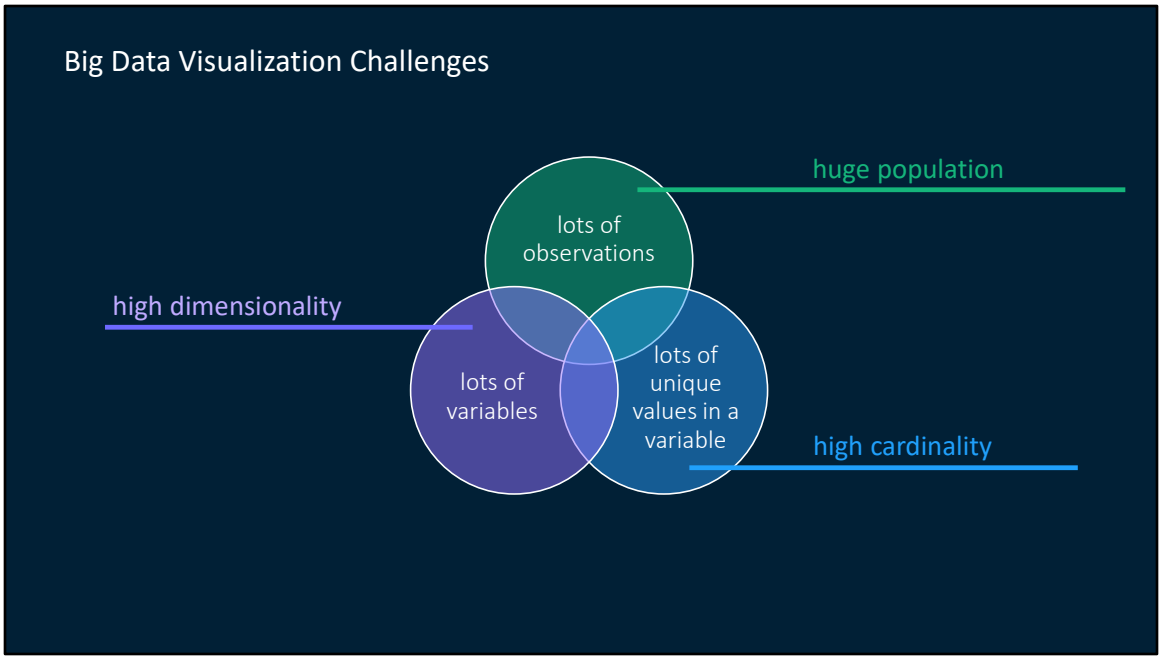
Data Visualization

"A picture is worth a thousand words."



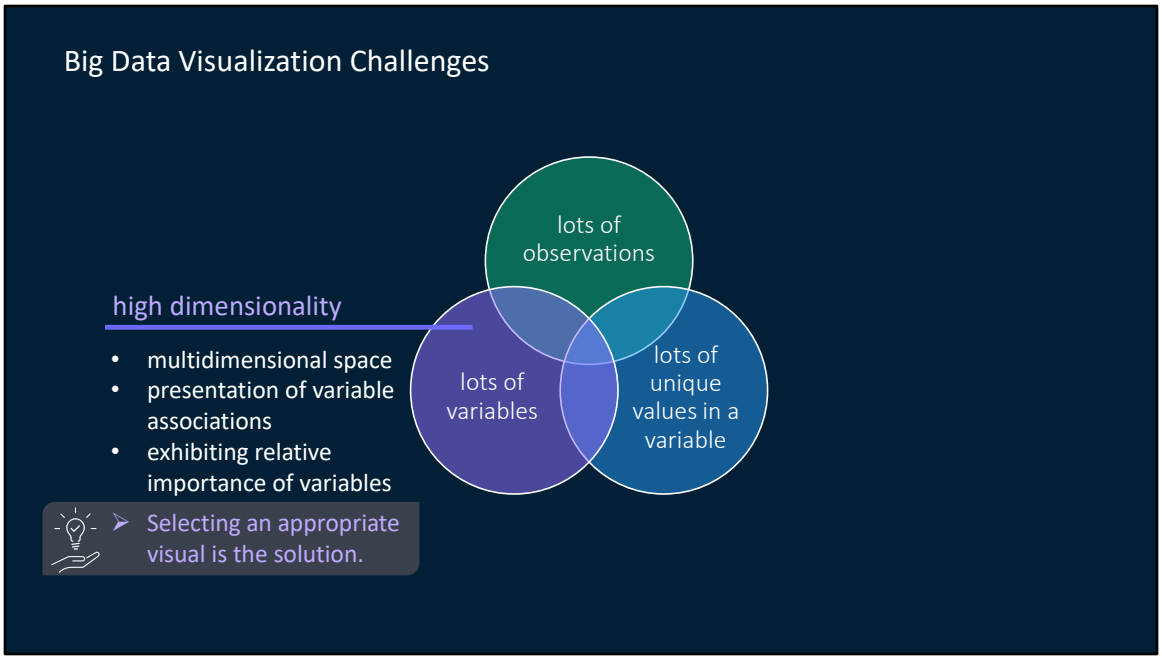
A picture is worth a thousand words – especially when you are trying to understand and discover insights from data. It is much easier to understand information in a visual than in a large table with lots of rows and columns. You can choose the most appropriate visualization by understanding the data and its composition, what information you are trying to convey to your audience, and how viewers process visual information.

You need effective data visualizations while exploring the data before fitting models. You also need effective data visualizations for model diagnostics after fitting models. Visualizing your data can be both fun and challenging.

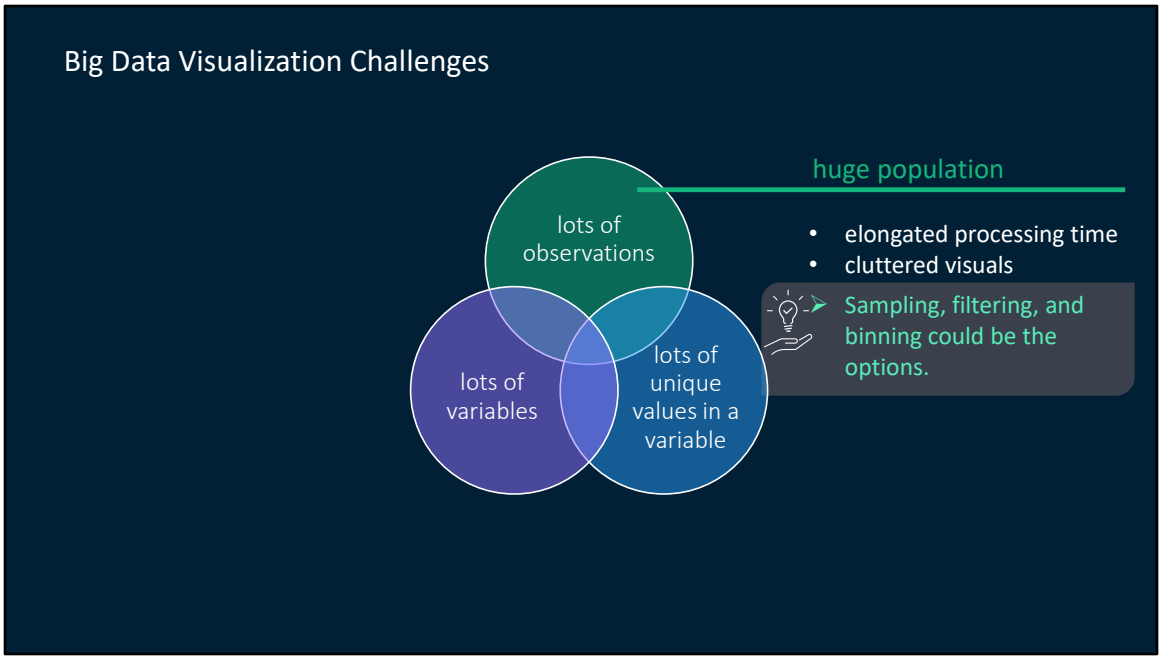


Big data brings new challenges to visualization because of the speed, size, and diversity of data that must be considered. To create meaningful visuals of your data, consider the following questions. Do you have a lot of observations (or *rows*)? Do you have a lot of variables (or *columns*)? Do your variables have a lot of unique values? Having lots of observations, lots of variables, or lots of unique values in a variable is a challenge for visualization.

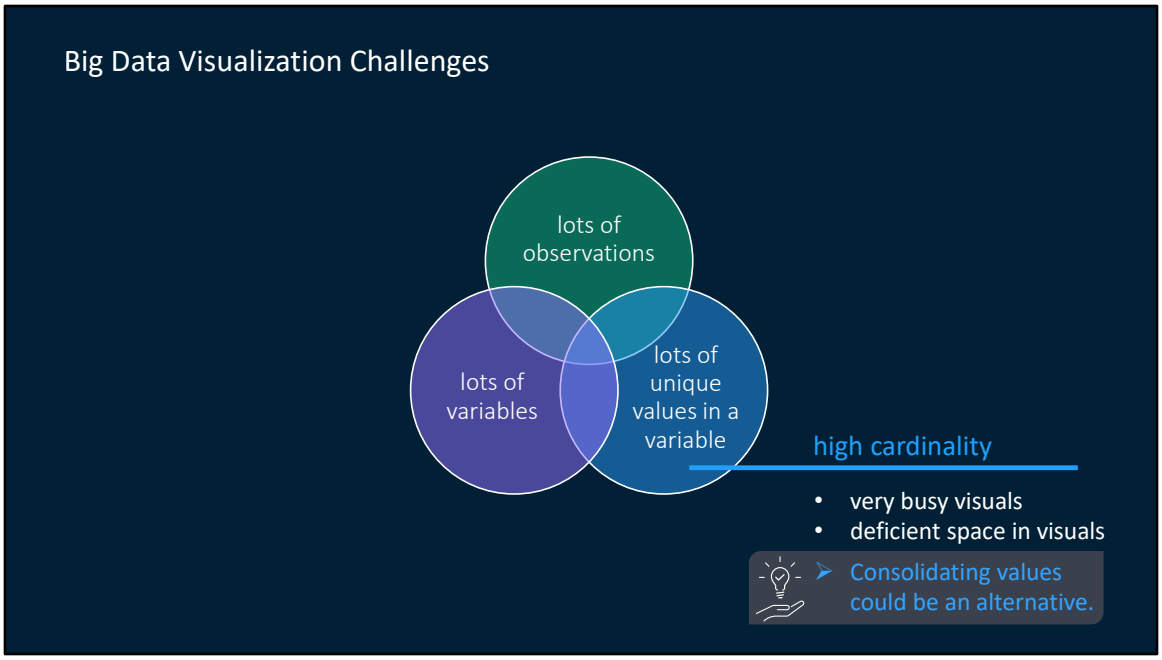
Large studies are generally more reliable than smaller studies, and there is a growing interest in the analysis of *huge population* sizes that integrates information from millions of cases or different data sources, or both. The *dimension* of your data refers to the number of input variables. If you have a lot of variables, we call it *high dimensionality*. *Cardinality* is the uniqueness of data values contained in a column. *High cardinality* means there is a large percentage of unique values (such as bank account numbers, because each item should be unique). Low cardinality means a column of data contains a large percentage of repeat values (such as a "gender" column). Population size, dimensionality, and cardinality composition play an important role when selecting graphs to represent your data.



The analysis of high-dimensional data offers a great challenge because the human intuition about geometry of high dimensions fails. One of the biggest challenges in data visualization is to find general representations of data that can display the multivariate structure of more than two variables. Appropriate visuals are especially helpful when you're trying to find relationships among hundreds or thousands of variables to determine their relative importance – or if they are important at all.



Extremely large populations for visualization pose risk of slower execution. The visuals might be messy and cluttered with countless observations. For example, what if you have a billion rows in a data set and want to create a scatter plot on two measures? It would be impossible to see so many data points. And the application creating the visual might not be able to plot a billion points in a timely or effective manner. In such a scenario perhaps, you can explore the data on a workable sample or use *binning* (the grouping together of data) on both axes so that you can effectively visualize the big data. When working with large amounts of data, being able to quickly and easily filter your data is important. What if you want to view data only for a certain region, product line, or some other variable?



Another concern with big data is cardinality because the data might have many unique values per column. If there are too many columns in your bar chart, you cannot see the labels for each bar, and the graph becomes less meaningful. The resulting visuals might be very busy and chaotic. If you are working with high cardinality data, one challenge is displaying output in a way that's not overwhelming. You might need to collapse and condense the data while still providing graphs and charts that decision makers are accustomed to seeing.

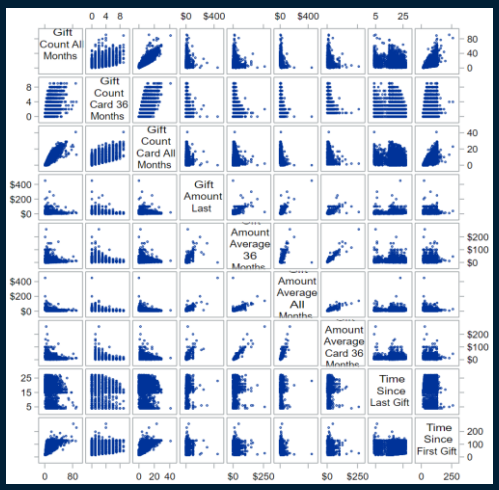
Visualization Examples

Traditional Statistical Plots

As the number of variables and observations increases…

…the plot can become more cluttered

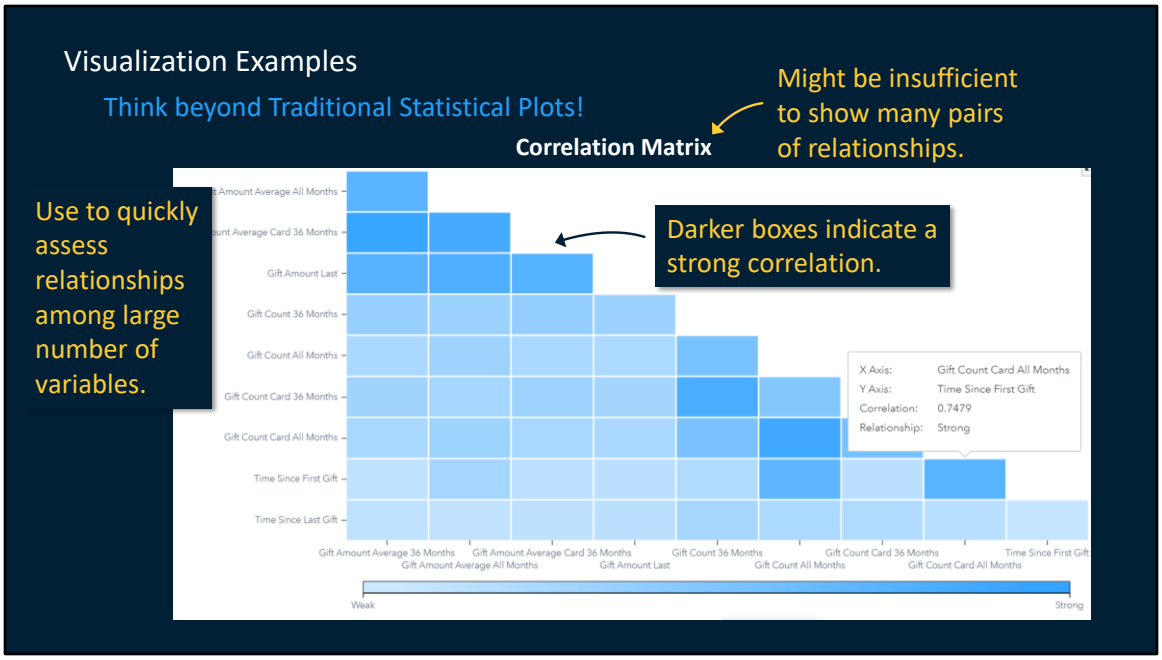
Scatter Plot Matrix



Traditional tools, might take significant time.

When analyzing relationships between many numeric variables, scatter plots and correlation coefficients can quickly uncover patterns to a subset of interesting relationships. Suppose you want to examine the relationship between nine continuous variables. Of course, the first choice would be to create a scatter plot matrix that displays every possible pairing of the measures that are assigned to the visualization. However, as the number of variables and observations increases, the plot can become more cluttered. Also, in a big data scenario, traditional tools, or algorithms, might take significant time to return results.

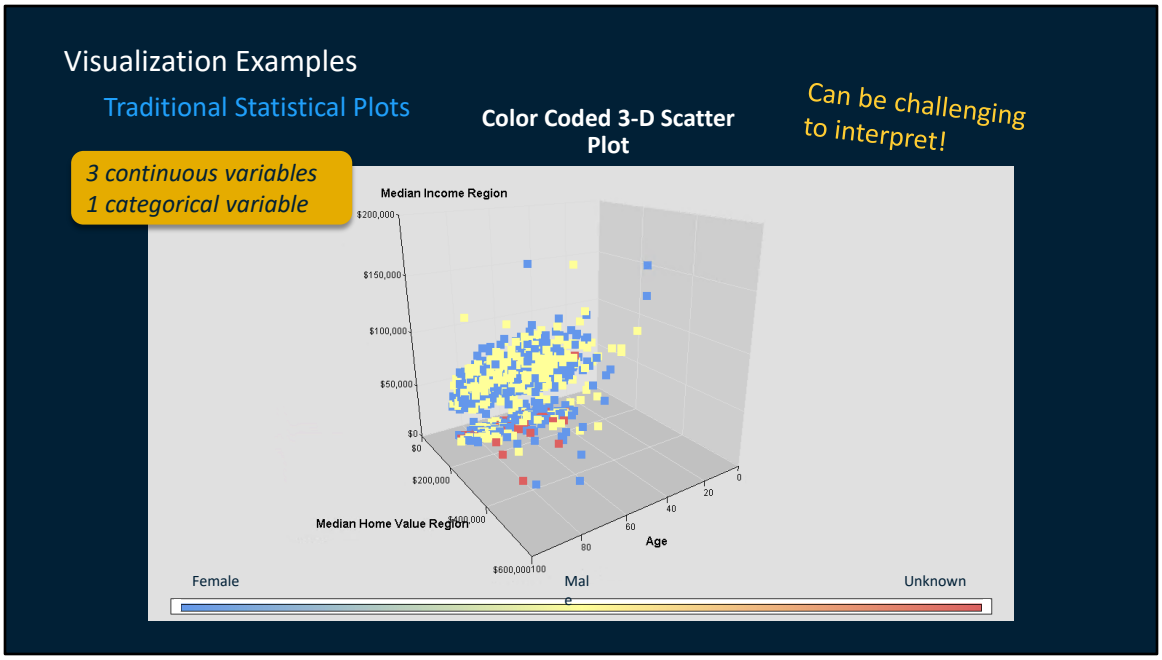
Given this scatter plot, can you determine the strength of correlation between pairs of variables? Is there an efficient way to analyze the relationships between these continuous variables while accounting for data size and ease of interpretation?



Yes, there are new sophisticated tools and techniques that can address the challenges associated with visualizing big data. The output shown here is from SAS Visual Analytics that uses intelligent autocharting to help business analysts and nontechnical users easily visualize their data.

So, instead of using the traditional scatter plot matrix, you use a correlation matrix plot, which makes it easy to quickly assess relationships among large number of variables. It also shows how strong the relationship is between variables. A color-coded correlation matrix quickly identifies strong and weak relationships between the variables. Darker boxes indicate a strong correlation and lighter boxes indicate a weaker correlation. Also, if you hover your mouse pointer over a box, a summary of the relationship is shown.

So, even if you have more than nine variables, this chart is still relevant. However, a scatter plot matrix might be insufficient to show many pairs of relationships.



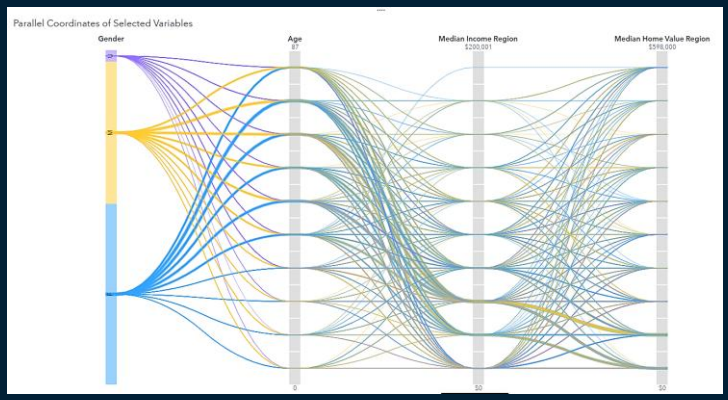
Consider another scenario where you want to explore the relationship between three numeric variables and a categorical variable. The first visualization technique that you might think of is a 3-D scatter plot, which can be used to show the relationship between three numeric variables. A fourth variable can be added by matching the color of the markers, as shown in this chart where we are trying to understand the relationship between median income region, median home value region, age, and gender. Needless to say, a three-dimensional plot is more challenging to interpret than a 2-dimensional plot. Also, if you needed to add few more variables in this exploration, a 3-D scatter plot wouldn't allow this.

Visualization Examples

Think beyond Traditional Statistical Plots!

efficient for analyzing multivariate numerical data

Parallel Coordinate Plot



use to compare multiple numerical variables with
- different magnitudes and
- different units of measurement

In big data, you quite often come across situations where you need to explore relationships between many variables (say, more than three). Well, the good news is that there are efficient visualization techniques for analyzing multivariate numerical data. One such technique is the *parallel coordinate plot*, as shown in this example.

A parallel coordinate plot displays data as lines moving through categories and binned measures. These lines are color coded and often are called *polylines*. The thickness of a polyline indicates the relative number of observations in that bin. You can restrict the active polylines to one or more bins in order to focus on only the data that interests you, which means that the user can interact with the plot.

A parallel coordinate plot enables you to compare many quantitative variables together, looking for patterns and relationships between them. They are appropriate for simultaneously comparing multiple numerical variables when those variables have different magnitudes (or different scales) and different units of measurement. The idea is to find patterns, similarities, clusters, and positive, negative, or no particular relationships in multidimensional data sets. And most importantly, you can explore the relationship between multiple variables while using a two-dimensional plot.

More about Visualization Examples

To explore this learning resource in greater detail, please visit the course page on learn.sas.com. There, you'll find comprehensive course materials, tutorials, and updates to support your learning journey.

Question: Data Visualization Challenges

Which of these terms describes a situation with ***lots of unique values in a variable***?

a. Huge population
b. High dimensionality
c. High cardinality
d. Low variability

Answer: Data Visualization Challenges

Which of these terms describes a situation with ***lots of unique values in a variable***?

a. Huge population
b. High dimensionality
c. High cardinality
d. Low variability

The correct answer is c.

When you have lots of unique values in a variable, it is typically termed as *high cardinality*. Huge population is a situation with lots of observations. High dimensionality is a situation with lots of variables. Low variability means that the values are more consistent.

Dirty Data: Errors, Missing Values, and Outliers

- Visualizing big data
- **Processing noisy data**
- Handling high dimensional data
- Dealing with distinct variability in inputs
- Using available data judiciously
- Adjusting model flexibility and complexity
- Understanding learning and tuning of models
- Making models explainable

Data encryption errors, missingness, and outliers are commonly present in noisy data.

Dirty Data: Errors, Missing Values, and Outliers

Checking Account	No. of Checking Accounts	Average Daily Balance	Non Sufficient Funds	Direct Deposit for Sure	Savings Account	Closing Balance
cking	#cking	ADB	NSF	dirdep	SVG	bal
Y	1	468.11	1	1876	Y	1208
Y	1	68.75	0	0	Y	0
Y	1	212.04	0	6		0
	.	.	0	0	Y	4301
y	2	585.05	0	7218	Y	234
Y	1	47.69	2	1256		238
	1	4687.7	0	0		0
Y	.	.	1	0	Y	1208
	.	.	.	1598		0
	1	0.00	0	0		0
Y	3	89981.12	0	0	Y	45662
Y	2	585.05	0	7218	Y	234

Shown here is a simple example of dirty data from banking domain. Variable labels are listed for better understanding. You will need them for upcoming student activities.

Inadequate data scrutiny is a common oversight. Errors, outliers, and missing values must be detected, investigated, and corrected (if possible). The basic data scrutiny tools are raw listings, if/then subsetting functions, exploratory graphics, and descriptive statistics such as frequency counts and minimum and maximum values.

Question: Errors

What do you see in the data below that looks wrong or even just out of place? There really are no wrong answers here. No matter how trivial, what should be investigated?

	C1	C2	C3	C4	C5	C6	C7
	cking	#cking	ADB	NSF	dirdep	SVG	bal
R1	Y	1	468.11	1	1876	Y	1208
R2	Y	1	68.75	0	0	Y	0
R3	Y	1	212.04	0	6		0
R4		.	.	0	0	Y	4301
R5	y	2	585.05	0	7218	Y	234
R6	Y	1	47.69	2	1256		238
R7		1	4687.7	0	0		0
R8	Y	.	.	1	0	Y	1208
R9		.	.	.	1598		0
R10		1	0.00	0	0		0
R11	Y	3	89981.12	0	0	Y	45662
R12	Y	2	585.05	0	7218	Y	234

Possible Answer: Errors

There is a lowercase y in column 1 [R5C1] versus all the uppercase Ys.

	C1	C2	C3	C4	C5	C6	C7
	cking	#cking	ADB	NSF	dirdep	SVG	bal
R1	Y	1	468.11	1	1876	Y	1208
R2	Y	1	68.75	0	0	Y	0
R3	Y	1	212.04	0	6		0
R4		.	.	0	0	Y	4301
R5	y	2	585.05	0	7218	Y	234
R6	Y	1	47.69	2	1256		238
R7		1	4687.7	0	0		0
R8	Y	.	.	1	0	Y	1208
R9		.	.	.	1598		0
R10		1	0.00	0	0		0
R11	Y	3	89981.12	0	0	Y	45662
R12	Y	2	585.05	0	7218	Y	234

Live Web students: Type in the chat what you think should be investigated.

Possible Answer: Errors

In the sixth row, values in C3-C7 are not in line with the others.

	C1	C2	C3	C4	C5	C6	C7
	cking	#cking	ADB	NSF	dirdep	SVG	bal
R1	Y	1	468.11	1	1876	Y	1208
R2	Y	1	68.75	0	0	Y	0
R3	Y	1	212.04	0	6		0
R4		.	.	0	0	Y	4301
R5	y	2	585.05	0	7218	Y	234
R6	Y	1	47.69	2	1256		238
R7		1	4687.7	0	0		0
R8	Y	.	.	1	0	Y	1208
R9		.	.	.	1598		0
R10		1	0.00	0	0		0
R11	Y	3	89981.12	0	0	Y	45662
R12	Y	2	585.05	0	7218	Y	234

Live Web students: Type in the chat what you think should be investigated.

Possible Answer: Errors

For the **SVG** column, a row is blank [R6C6] when its corresponding **bal** column has a nonzero number.

	C1	C2	C3	C4	C5	C6	C7
	cking	#cking	ADB	NSF	dirdep	SVG	bal
R1	Y	1	468.11	1	1876	Y	1208
R2	Y	1	68.75	0	0	Y	0
R3	Y	1	212.04	0	6		0
R4		.	.	0	0	Y	4301
R5	y	2	585.05	0	7218	Y	234
R6	Y	1	47.69	2	1256		238
R7		1	4687.7	0	0		0
R8	Y	.	.	1	0	Y	1208
R9		.	.	.	1598		0
R10		1	0.00	0	0		0
R11	Y	3	89981.12	0	0	Y	45662
R12	Y	2	585.05	0	7218	Y	234

Live Web students: Type in the chat what you think should be investigated.

Possible Answer: Errors

Rows 5 and 12 are the same except for the lowercase y in R5C1. Are these the same observation where someone corrected the y but forgot to remove the original row?

	C1	C2	C3	C4	C5	C6	C7
	cking	#cking	ADB	NSF	dirdep	SVG	bal
R1	Y	1	468.11	1	1876	Y	1208
R2	Y	1	68.75	0	0	Y	0
R3	Y	1	212.04	0	6		0
R4		.	.	0	0	Y	4301
R5	y	2	585.05	0	7218	Y	234
R6	Y	1	47.69	2	1256		238
R7		1	4687.7	0	0		0
R8	Y	.	.	1	0	Y	1208
R9		.	.	.	1598		0
R10		1	0.00	0	0		0
R11	Y	3	89981.12	0	0	Y	45662
R12	Y	2	585.05	0	7218	Y	234

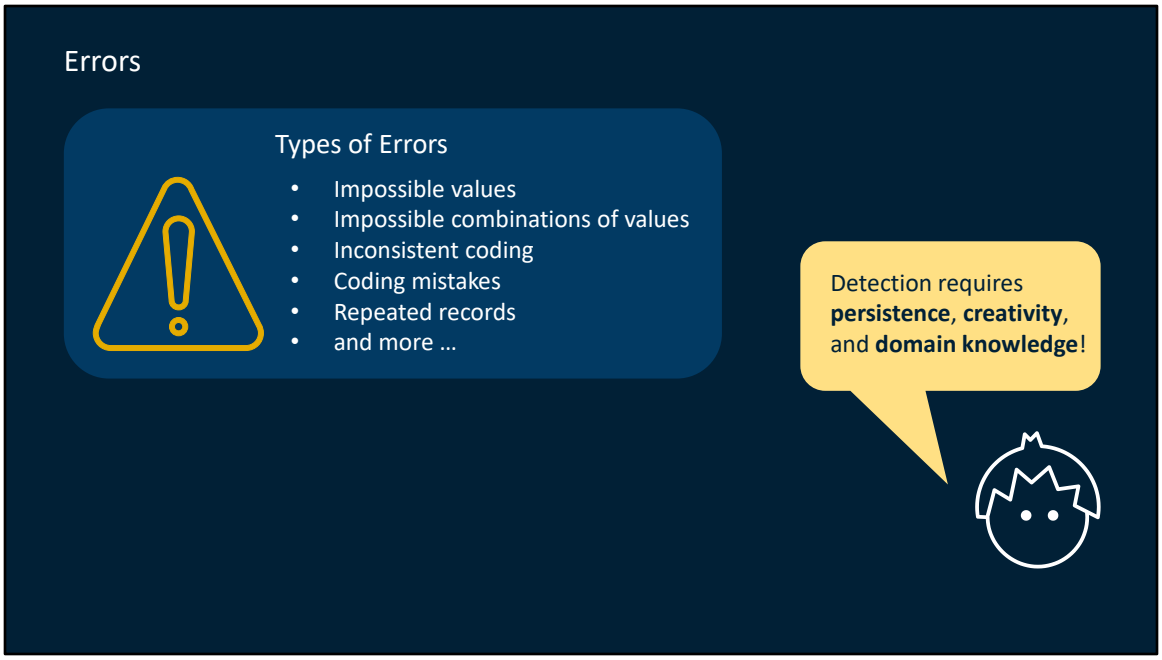
Live Web students: Type in the chat what you think should be investigated.

Possible Answer: Errors

What does a 2 in the **NSF** column [R6C4] mean? Is NSF (non-sufficient funds) binary?

	C1	C2	C3	C4	C5	C6	C7
	cking	#cking	ADB	NSF	dirdep	SVG	bal
R1	Y	1	468.11	1	1876	Y	1208
R2	Y	1	68.75	0	0	Y	0
R3	Y	1	212.04	0	6		0
R4		.	.	0	0	Y	4301
R5	y	2	585.05	0	7218	Y	234
R6	Y	1	47.69	2	1256		238
R7		1	4687.7	0	0		0
R8	Y	.	.	1	0	Y	1208
R9		.	.	.	1598		0
R10		1	0.00	0	0		0
R11	Y	3	89981.12	0	0	Y	45662
R12	Y	2	585.05	0	7218	Y	234

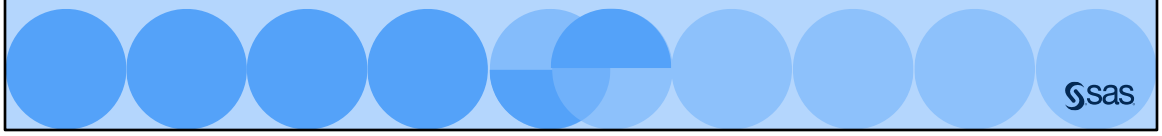
Live Web students: Type in the chat what you think should be investigated.



Detection of such errors as impossible values, impossible combinations of values, inconsistent coding, coding mistakes, and repeated records require persistence, creativity, and domain knowledge. You have already seen most of these errors in the preceding activity.

Demo: Modifying and Correcting Data

To explore this learning resource in greater detail, please visit the course page on learn.sas.com. There, you'll find comprehensive course materials, tutorials, and updates to support your learning journey.



Question: Missing Values

Are there any missing values?

	C1	C2	C3	C4	C5	C6	C7
	cking	#cking	ADB	NSF	dirdep	SVG	bal
R1	Y	1	468.11	1	1876	Y	1208
R2	Y	1	68.75	0	0	Y	0
R3	Y	1	212.04	0	6		0
R4		.	.	0	0	Y	4301
R5	y	2	585.05	0	7218	Y	234
R6	Y	1	47.69	2	1256		238
R7		1	4687.7	0	0		0
R8	Y	.	.	1	0	Y	1208
R9		.	.	.	1598		0
R10		1	0.00	0	0		0
R11	Y	3	89981.12	0	0	Y	45662
R12	Y	2	585.05	0	7218	Y	234

Missing values occur for a variety of reasons. They often represent unknown but knowable information. Structural missing data represent values that logically could not have a value. Missing values are often coded in different ways - sometimes by a question mark, or -9999, sometimes by "n/a" or by some other specific number or character; and sometimes miscoded as zeros. The reasons for the coding and the consistency of the coding must be investigated.

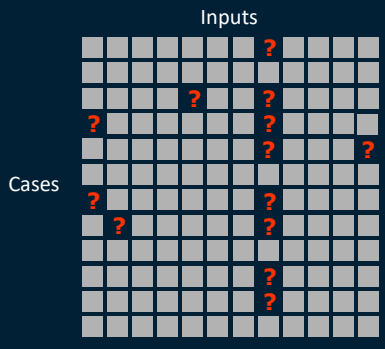
Answer: Missing Values

Are there any missing values?

	C1	C2	C3	C4	C5	C6	C7
	cking	#cking	ADB	NSF	dirdep	SVG	bal
R1	Y	1	468.11	1	1876	Y	1208
R2	Y	1	68.75	0	0	Y	0
R3	Y	1	212.04	0	6		0
R4		.	.	0	0	Y	4301
R5	y	2	585.05	0	7218	Y	234
R6	Y	1	47.69	2	1256		238
R7		1	4687.7	0	0		0
R8	Y	.	.	1	0	Y	1208
R9		.	.	.	1598		0
R10		1	0.00	0	0		0
R11	Y	3	89981.12	0	0	Y	45662
R12	Y	2	585.05	0	7218	Y	234

Missing values occur for a variety of reasons. They often represent unknown but knowable information. Structural missing data represent values that logically could not have a value. Missing values are often coded in different ways - sometimes by a question mark, or -9999, sometimes by "n/a" or by some other specific number or character; and sometimes miscoded as zeros. The reasons for the coding and the consistency of the coding must be investigated.

Missing Data



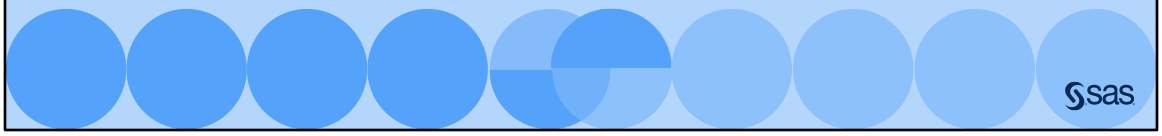
- Missing data problems
- Missing data analysis strategies
- Missing data mechanisms
- Missing data causes

(Self-Study)

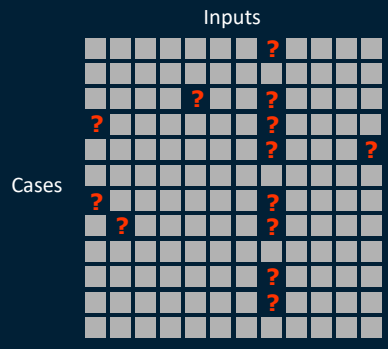
Nearly all the real-world data have missing values, and dealing with missing values or incomplete observations is critical in any data mining and machine learning analysis. The term missing data is defined here as a statistical problem characterized by an incomplete data matrix where there's one or multiple variables where certain observations are missing for those variables. It is important to understand:
What are problems of missing data?
How to handle missing values?
What are the types of missingness? And why is your data missing? (You can learn the last two in a self-study section of this lesson.)

Missing Data Mechanisms

To explore this learning resource in greater detail, please visit the course page on learn.sas.com. There, you'll find comprehensive course materials, tutorials, and updates to support your learning journey.



Missing Data Problems



Major problems caused by missing values:

- Parameter bias
  - overestimation or underestimation
- Inferential error
  - prediction error or hypothesis testing errors

missing data problems (Newman 2014)

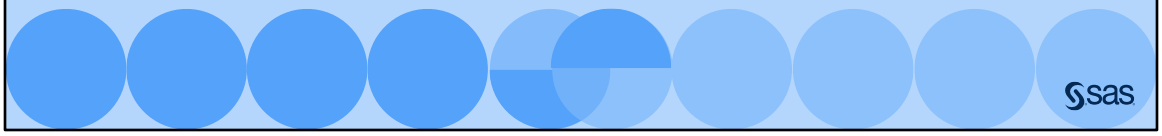
The two major problems caused by missing data are parameter bias and inferential error.

Bias refers to the systematic overestimation or underestimation of a parameter (such as underestimated mean, correlation, or regression coefficient). Missing data bias typically occurs when the missingness mechanism is systematic or nonrandom - under MAR or MNAR.

Error refers to prediction error or hypothesis testing errors of inference, such as Type I error (also called *false positive*) and Type II error (or *low power*, also called *false negative*).

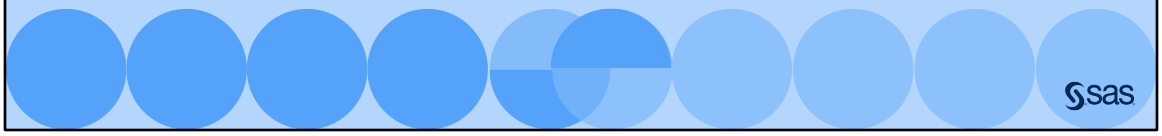
More about Missing Data Problems

To explore this learning resource in greater detail, please visit the course page on learn.sas.com. There, you'll find comprehensive course materials, tutorials, and updates to support your learning journey.



Missing Data Causes

To explore this learning resource in greater detail, please visit the course page on learn.sas.com. There, you'll find comprehensive course materials, tutorials, and updates to support your learning journey.



Question: Missing Value Representation

How are missing values in a numeric variable represented in SAS?

a. By a blank (‘ ‘)
b. By a period (.)
c. By a comma (,)
d. By a semicolon (;)

Answer: Missing Value Representation

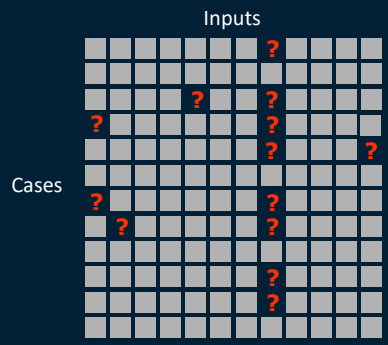
How are missing values in a numeric variable represented in SAS?

a. By a blank (‘ ‘)
b. By a period (.)
c. By a comma (,)
d. By a semicolon (;)

ANSWER: b

In SAS, numeric missing values are represented by a period (.), whereas character missing values are represented by a blank (‘ ').

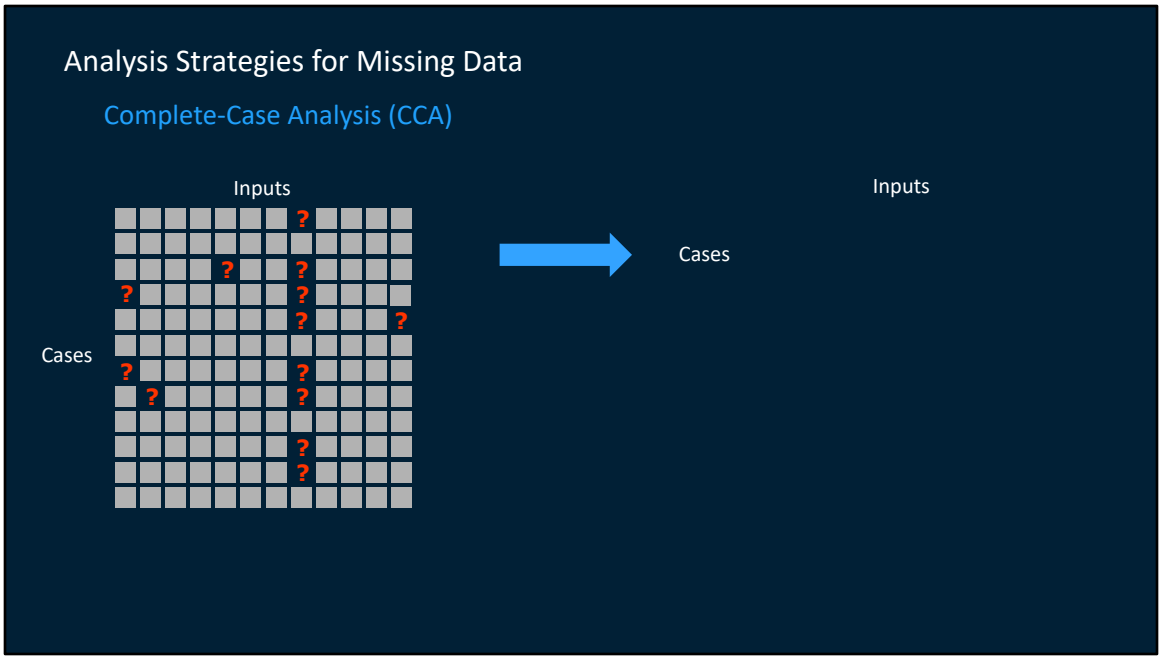
Analysis Strategies for Missing Data



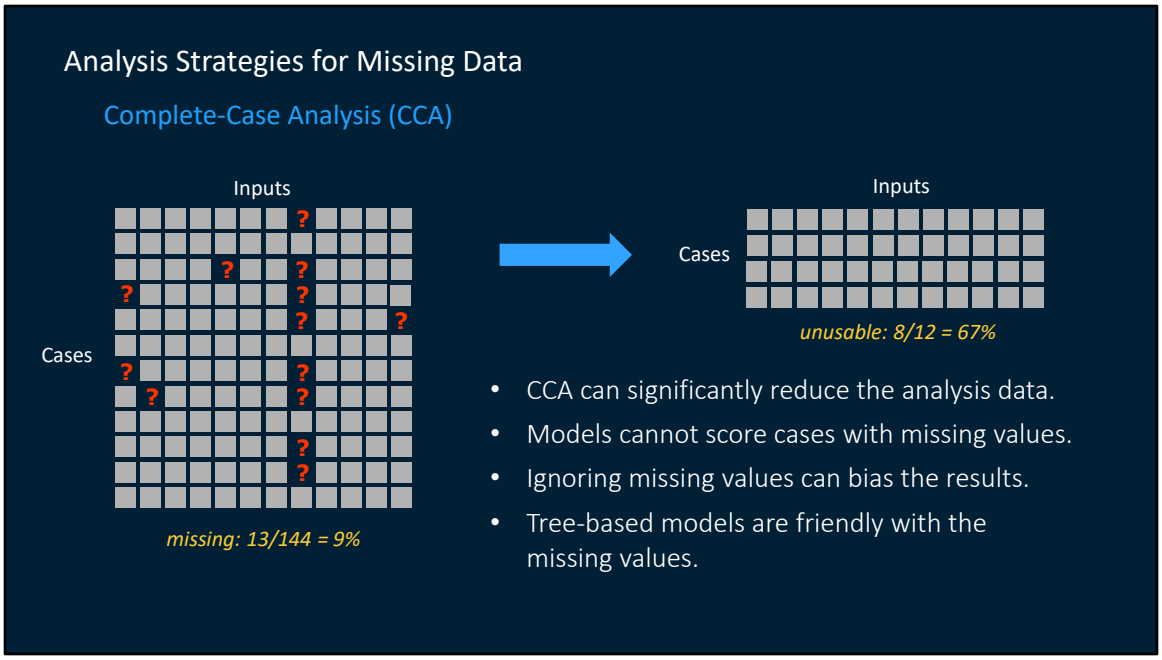
Analysis strategies for handling missing value

- Complete-case analysis (CCA)
- Imputation
- Missing indicators

There are predominantly three analysis strategies for handling missing values.



The simplest strategy is 'do nothing' when you encounter missing values. However, missing values can be theoretically and practically problematic for many machine learning models such as **Neural Network** or **Regression** or **SVM**. This will lead you to use only the cases that have complete records in the analysis, commonly known as *complete-case analysis*.

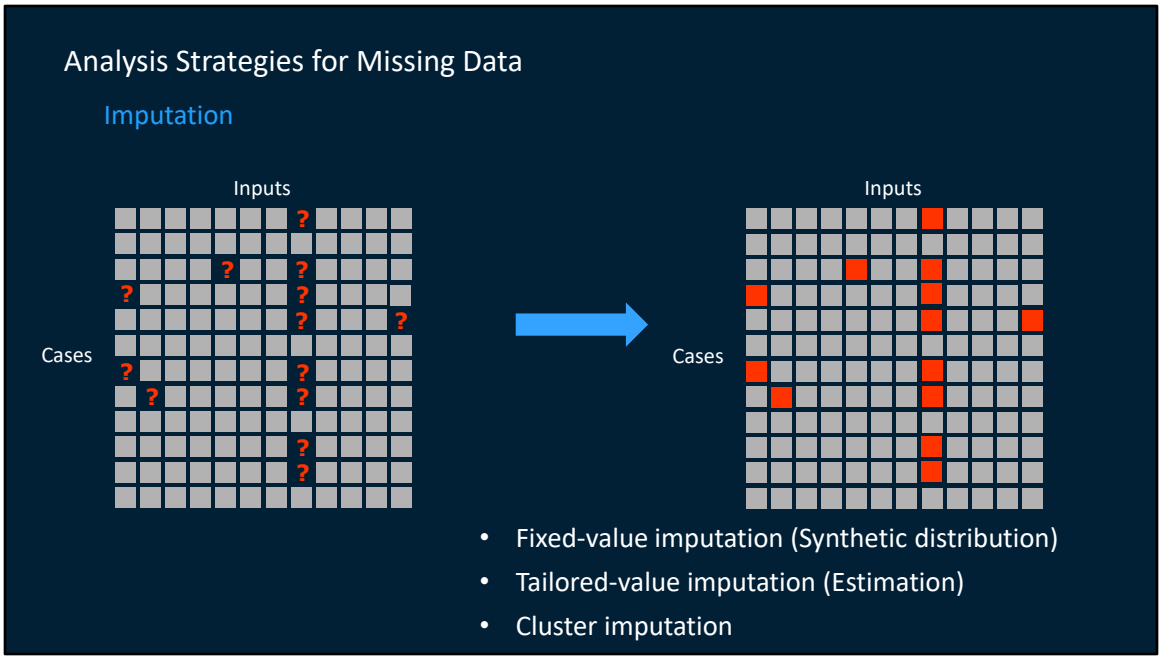


The main difficulty with this strategy is practical. Even a small amount of missing values in a high dimensional data set can cause a disastrous reduction in data. In the above example, only 13 of the 144 values (9%) are missing, but a complete-case analysis would use only four cases—only a third of the data set. If there are very few cases with missing values, this is a viable option.

Another problem relates to model deployment. How would a model built on the complete cases score a new case if it had a missing value? If there is missingness in your training data, it is very likely that your scoring data or the new data would also have, ideally, a similar type of missingness, but in limited amount. Thus, inefficient scorability.

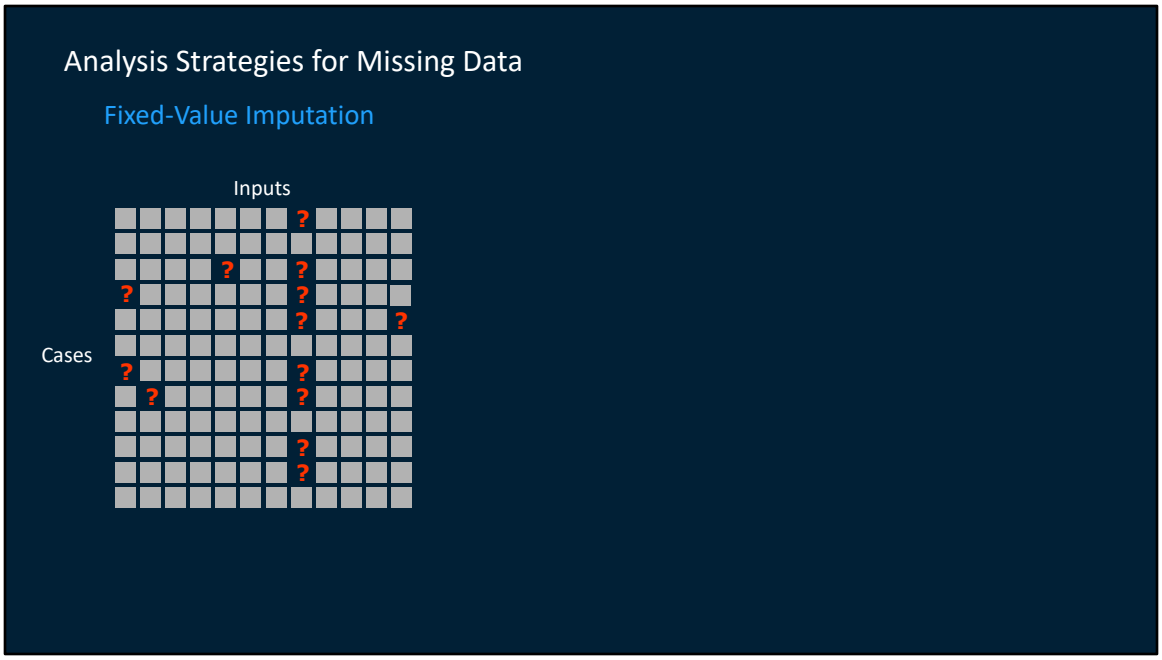
Rejecting all incomplete observations might also bias the sample, because observations that have missing values might have other things in common as well. For example, if the "missingness" is related to the inputs or to the target, then ignoring missing values can bias the results. Omitting the missing term from the parametric equation usually produces an extremely biased prediction.

Nevertheless, some other machine learning models such as **decision tree**, **forest**, and **gradient boosting** can use missing values to their advantage. For example, missing values can be treated as legitimate values while training these models. So missingness is not always awkward.



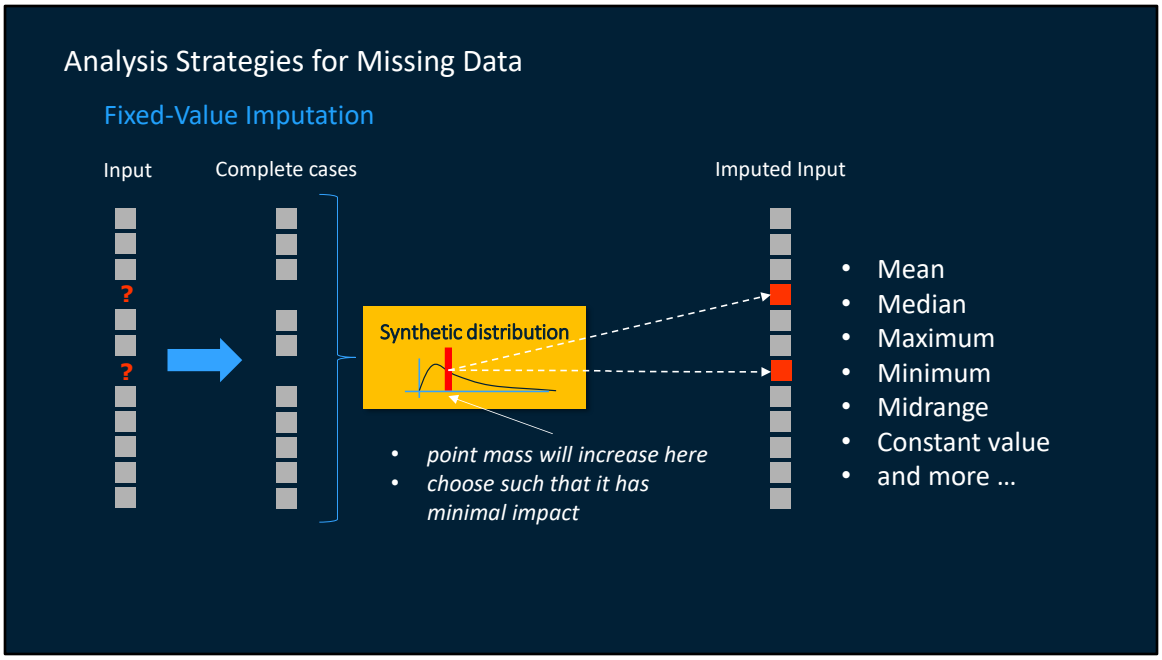
The primary remedy for missing values in most of the machine learning models is a missing value replacement strategy called *imputation*. Fill in the missing values with some **reasonable** value and then run the analysis on the full (filled-in) data. Imputation is an important aspect of data preprocessing that has the potential to make (or break) your model. The goal of imputation is to replace missing values with values that are close to what the missed value might have been. If successful, you will have more data with which the model can learn. However, you could introduce serious bias and make learning more difficult. The bottom line is that whether you should use imputation for missing values is completely dependent on your data and what the missing value represents. Be diligent!

Missing value replacement strategies fall into one of three categories: fixed-value imputation, tailored-value imputation, and cluster imputation.



The simplest types of imputation methods fill in the missing values with a fixed value. For example, you can impute the missing values in a survey with a "not answered" category. That is, use a one-size-fits-all approach to handle missing values. Fixed value imputation is a general method that works for all data types and consists of substituting the missing value with a fixed value.

As an illustration, consider the variable in the first column. Out of 12 values, two are missing.



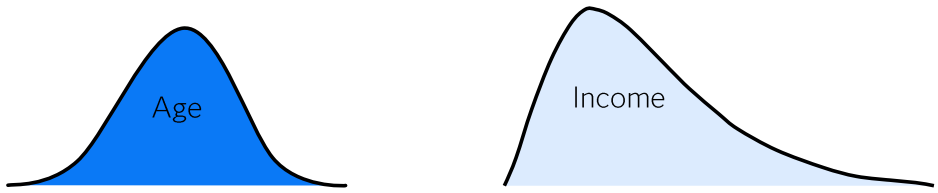
Taking one variable at a time, there are 10 complete observations. The simplest types of imputation methods fill in the missing values with the mean of the complete cases. Thus, the two missing values are replaced with the arithmetic mean computed on these 10 observations. In any case with a missing input measurement, the missing value is replaced with a fixed number. As a result, the point mass will increase at the selected fixed number in the synthetic distribution; thus, the choice of the fixed value is not arbitrary. Ideally, it should be chosen to have minimal impact on the magnitude of an input's association with the target. With many modeling methods, this can be achieved by locating the point mass at the input's mean value.

If you have determined that imputation might be beneficial for your data, there are several options to choose from. We will not be going into detail about the types of imputation available, as the type you use will be highly dependent on your data and goals. The examples listed here give you a glimpse of the possibilities.

The missing values of categorical variables could be treated as a separate category. For example, *type of residence* might be coded as *own home*, *buying home*, *rents home*, *rents apartment*, *lives with parents*, *mobile home*, and *unknown*. This method would be preferable if the missingness is itself a predictor of the target.

Question: Mean or Median Imputation

Age and **Income** are two variables that have missing values. You can use either mean or median imputation, but each only once. Variable distributions are given below. Which one you would use and why?

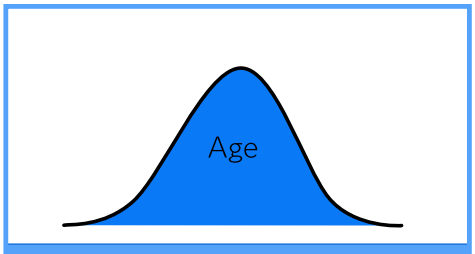


Because the **Age** variable is nearly normal, mean imputation would work. Nonetheless, for a normally distributed data, mean and median coincide.

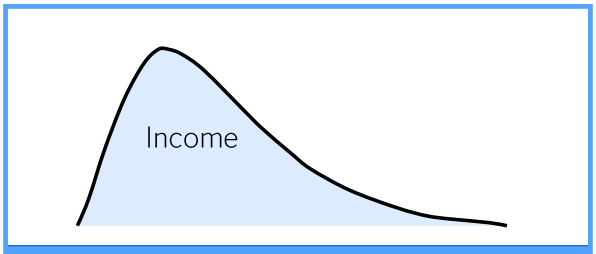
In the case of a high number of outliers in your data set, it is recommended to use the median instead of the mean. Therefore, median imputation would be used for **Income** as it is positively skewed.

Answer: Mean or Median Imputation

Age and **Income** are two variables that have missing values. You can use either mean or median imputation, but each only once. Variable distributions are given below. Which one you would use and why?



Because the **Age** variable is nearly normal, **mean imputation** would work. Nonetheless, for normally distributed data, **mean and median coincide**.



In the case of a **high number of outliers** in your data set, it is recommended to **use the median** instead of the mean. Therefore, **median imputation** would be used for **Income** as it is positively skewed.

Because the **Age** variable is nearly normal, mean imputation would work. Nonetheless, for a normally distributed data, mean and median coincide.

In the case of a high number of outliers in your data set, it is recommended to use the median instead of the mean. Therefore, median imputation would be used for **Income** as it is positively skewed.

Question: Missing Stock Exchange Price

If in the stock exchange a minimum price has been reached and the exchange process has been stopped, the missing stock exchange price can be replaced with which of the following?

a. Mean
b. Median
c. Minimum
d. Maximum

Answer: Missing Stock Exchange Price

If in the stock exchange a minimum price has been reached and the exchange process has been stopped, the missing stock exchange price can be replaced with which of the following?

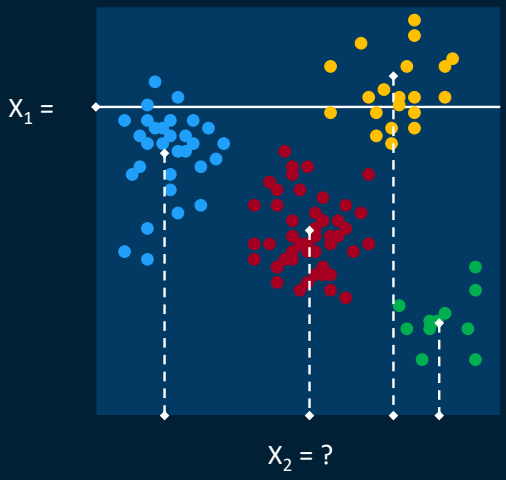
a. Mean
b. Median
c. Minimum
d. Maximum

ANSWER: c
If you know that the data must fit a given range, and if the measuring system stops recording and the signal saturates beyond one of such boundaries, you can use either minimum or maximum as the replacement value for missing values. The missing stock exchange price can be replaced with the minimum value of the law's exchange boundary.

Cluster Imputation

Cluster-Mean Imputation:

1. Cluster the cases into relatively homogenous groups.
2. Conduct mean-imputation within each group.
3. For new cases with multiple missing values, use the cluster mean that is closest in all the nonmissing dimensions.



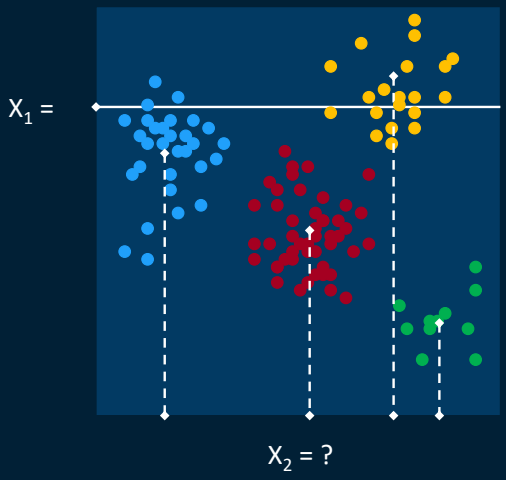
Mean imputation can be refined by using the mean within homogenous groups of the data. Cluster-mean imputation is a somewhat more practical alternative, which is a three-step process.

Step 1: Cluster the cases into relatively homogenous groups.

Step 2: Conduct mean-imputation within each group.

Step 3: For new cases with multiple missing values, use the cluster mean that is closest in all the nonmissing dimensions.

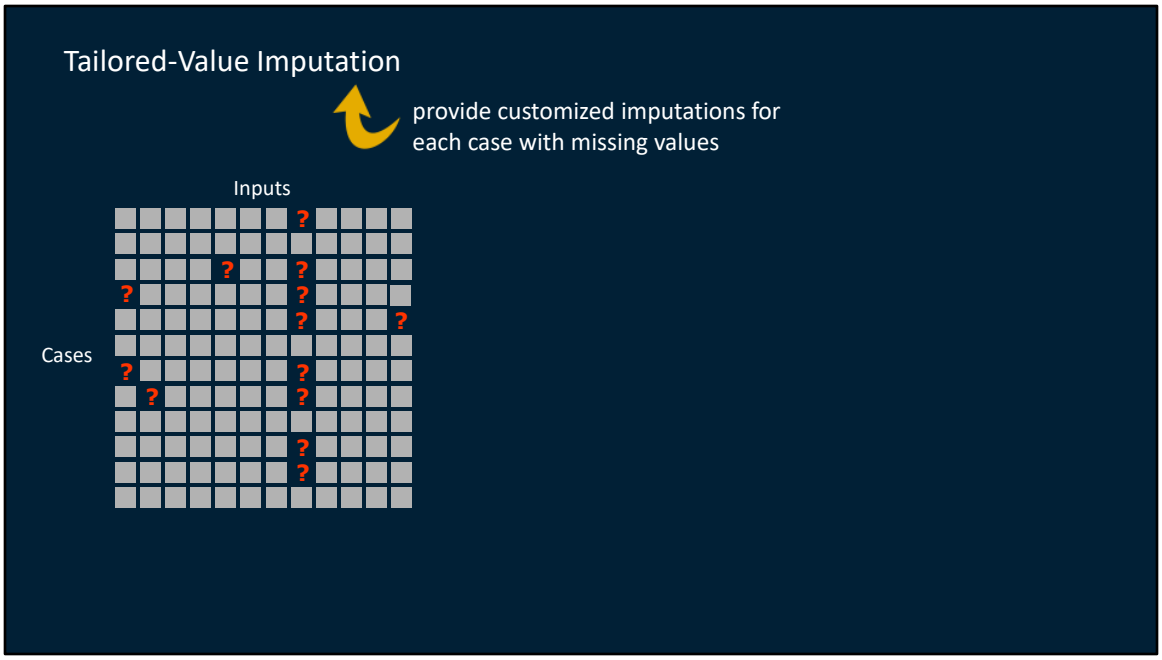
Cluster Imputation



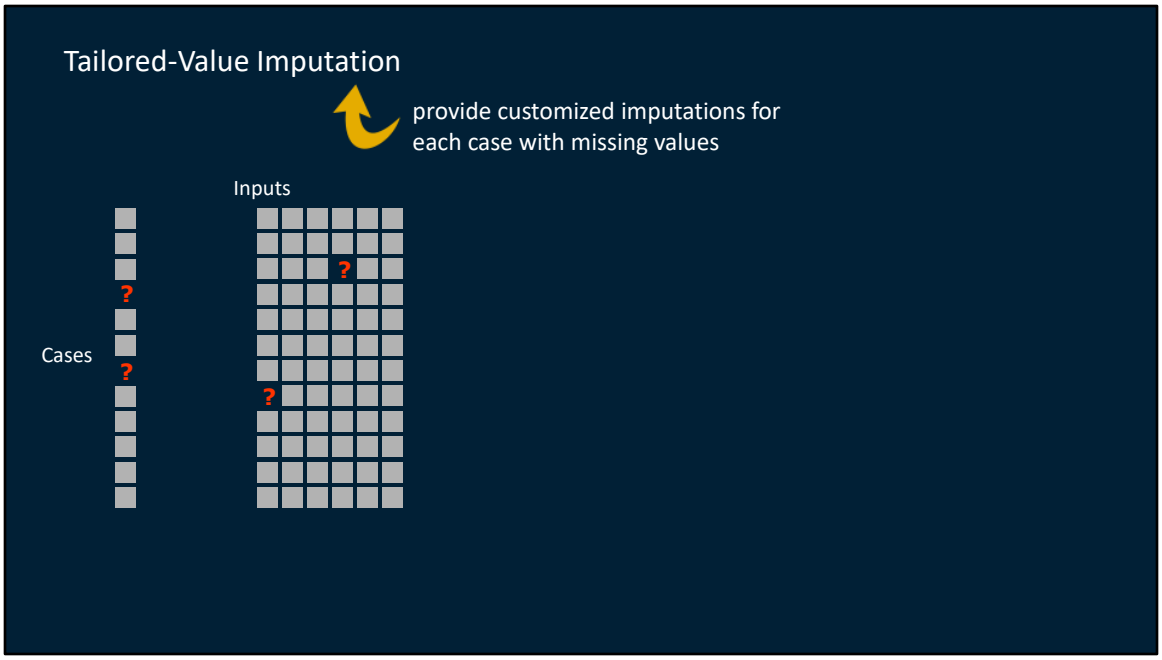
Cluster-Mean Imputation:

- Can accommodate missingness that depends on the other input variables.
- Different values are imputed for different clusters.
- Imputed values are the same or fixed within each cluster.

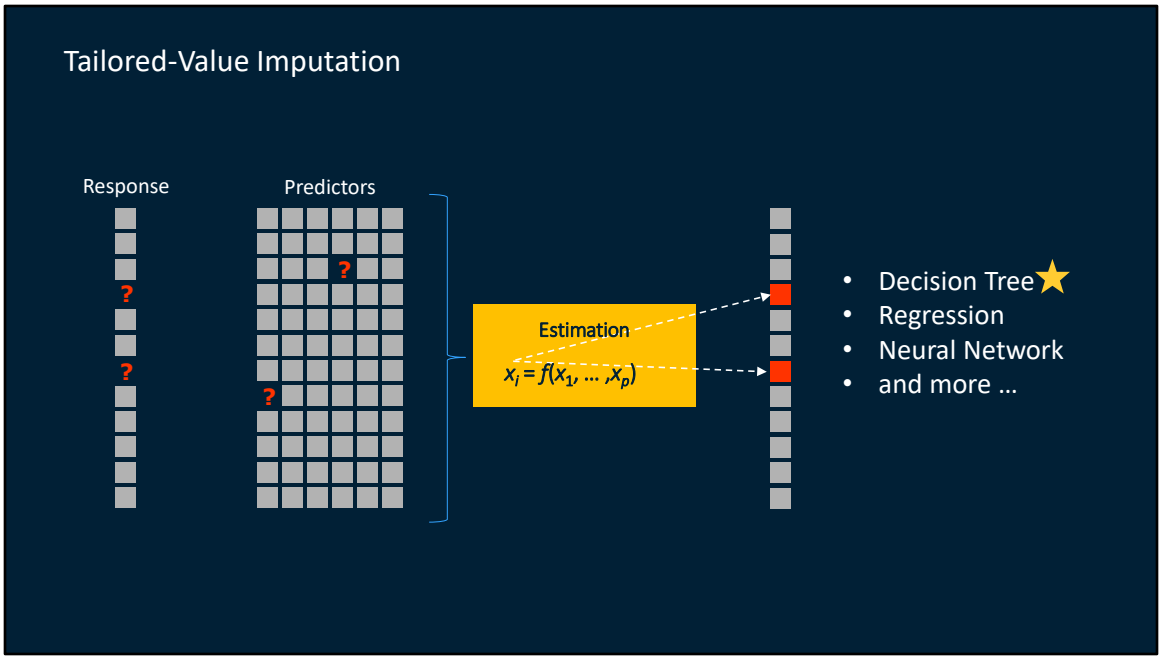
This method can accommodate missingness that depends on the other input variables. Different values are imputed for different clusters. However, the imputed values are the same or fixed within each cluster.



Many times, we might want to avoid the one-size-fits-all approach and provide customized imputations for each case with missing values. This approach is called tailored-value imputation.



Again, consider the variable in the first column. The remaining inputs will be used in this imputation method, if not all. For example, out of 11 remaining inputs, six might be used based on some criteria.



Mean imputation uses the unconditional mean of the variable. Similarly, cluster-mean imputation uses the unconditional mean of the variable within each segment. An attractive extension would be to use the mean conditional on the other inputs, that is, viewing the missing value problem as a prediction problem. When an input's value is unknown, you can use this model to predict or estimate the unknown missing value. This is referred to as tailored-value imputation. Specifically, *k* models could be built, one for each input variable using the other inputs as predictors. This would presumably give better imputations and be able to accommodate missingness that depends on the values of the other inputs. However, a complication of tailored-value imputation is that the other inputs might themselves have missing values. Consequently, the *k* imputation models also need to accommodate missing values. For decision trees, missing values are not problematic. Hence, decision trees are the preferred models for this purpose.

Question: Imputation Methods

If you need to accommodate missingness that depends on the other input variables, which of the following imputation methods you can use? (Select all that apply)

a. Fixed-value imputation
b. Tailored-value imputation
c. Cluster imputation

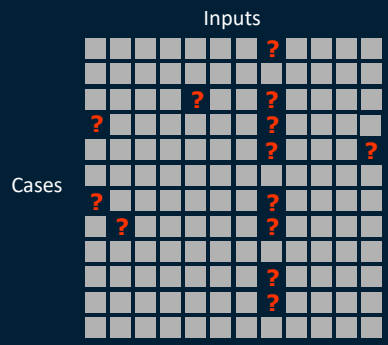
Answer: Imputation Methods

If you need to accommodate missingness that depends on the other input variables, which of the following imputation methods you can use? (Select all that apply)

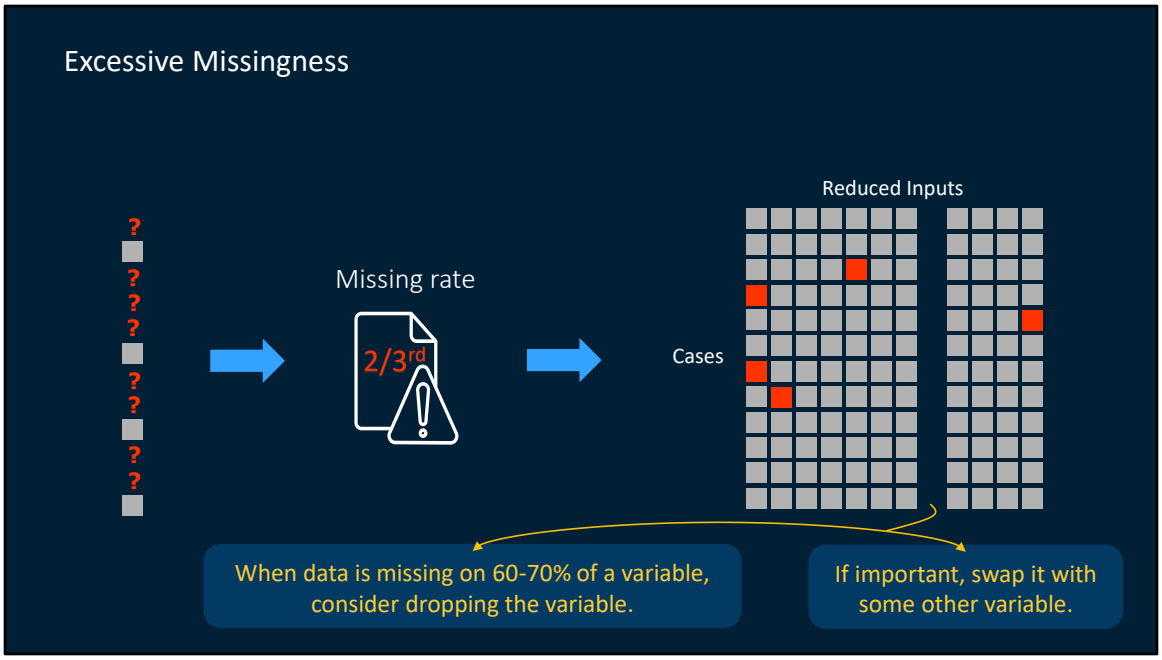
a. Fixed-value imputation
b. Tailored-value imputation
c. Cluster imputation

ANSWER: b and c
In cluster imputation, clusters are created using other input variables and then using customized imputations within each cluster. Tailored-value imputation accommodates more customization by providing estimated imputations for each case with missing values using other input variables. However, fixed-value imputation does not consider other input variables at all.

Excessive Missingness



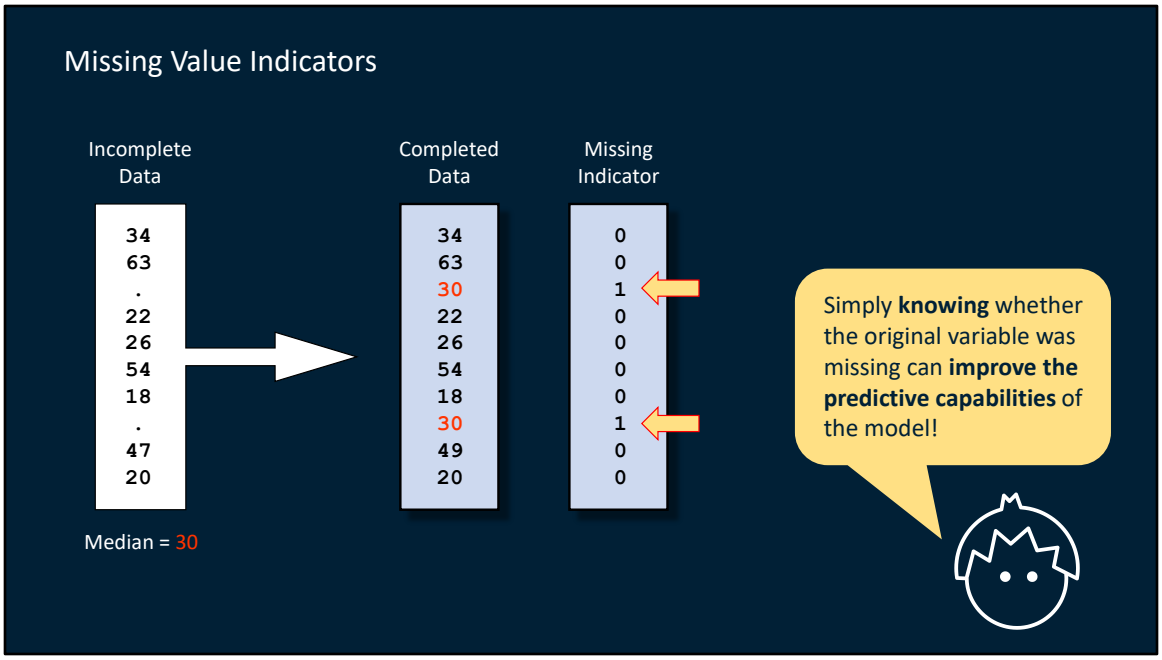
There are situations when a variable has a lot of missing values. Let's consider the variable that has the most missing values in this table.



Two-thirds of the values are missing in this variable. The variable can be dropped from the analysis due to its vulnerability of inducing high bias in your analysis. It is certainly not logical to get 67 percent missing values imputed based on 33 percent of the available values. As a general rule, when data is missing on 60 to 70 percent of a variable, you should consider dropping the variable.

If the variable is a very important predictor for the target variable, then swapping the variable with some other variable could be an option.

You must specify the threshold for rejecting inputs with missing values. You generally set this up beforehand while defining the metadata. An input is rejected if it has a higher percentage of missing values than the specified maximum percent. The default threshold of missing values in almost all SAS solutions is 50 percent.



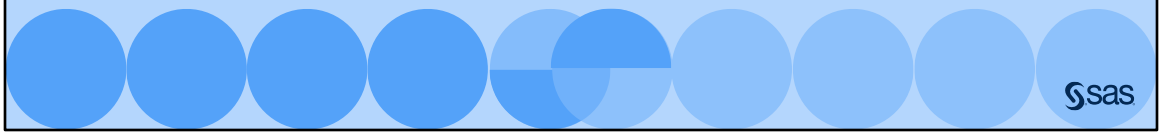
If a very large percentage of values is missing, then the variable might be better handled by omitting it from the analysis or by creating the binary missing indicator only. The missing indicator is 1 when the corresponding input is missing and 0 otherwise. Each missing indicator becomes an input to the model. Cases often contain missing values for a reason. If the reason for the missing value is in some way related to the target variable, useful predictive information is lost. It might seem strange, but sometimes just knowledge of missingness - simply knowing whether the original variable was missing or not - can improve the predictive capabilities of the model.

This strategy is somewhat unsophisticated and yet satisfies two of the most important considerations in predictive modeling: capturing the relationship of missingness with the target and efficient scorability. A new case is easily scored. First, replace the missing values with any valid imputed value like the median from the development data, and then apply the prediction model.

If a very small percentage of the values is missing (less than 0.1 percent), then the missing indicator is probably of little value.

Demo: Managing Missing Values

To explore this learning resource in greater detail, please visit the course page on learn.sas.com. There, you'll find comprehensive course materials, tutorials, and updates to support your learning journey.



Question: Missing Value Imputation

If the mean of an input variable in training data is 35 and the mean for the same variable in validation data is 39, which one would you use for missing value imputation? Select all that apply.

a. 35 for training data and 39 for validation data
b. 35 for both training and validation data
c. 39 for both training and validation data
d. 35 for scoring data

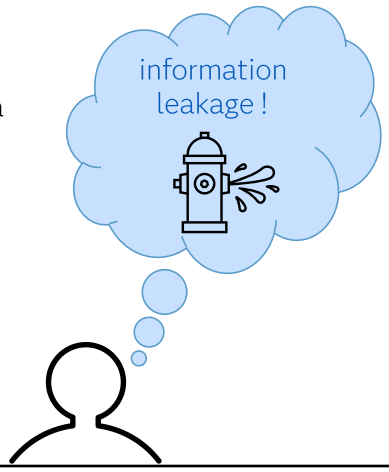
b and d are correct
The validation data is prepared exactly the same way as the training data to prevent information leakage. Validation data represents the unseen scoring data. In statistics and machine learning, leakage is the use of information in the model training process, which would not be expected to be available at prediction time, causing the predictive scores to overestimate the model's utility when run in a production environment.

Making the missing value imputation process part of the modeling process allays the modified distribution concern. Any modifications made to the training data are also made to the validation data and the remainder of the modeling population.

Answer: Missing Value Imputation

If the mean of an input variable in training data is 35 and the mean for the same variable in validation data is 39, which one would you use for missing value imputation? Select all that apply.

a. 35 for training data and 39 for validation data
b. 35 for both training and validation data
c. 39 for both training and validation data
d. 35 for scoring data



b and d are correct
The validation data is prepared exactly the same way as the training data to prevent information leakage. Validation data represents the unseen scoring data. In statistics and machine learning, leakage is the use of information in the model training process, which would not be expected to be available at prediction time, causing the predictive scores to overestimate the model's utility when run in a production environment.

Making the missing value imputation process part of the modeling process allays the modified distribution concern. Any modifications made to the training data are also made to the validation data and the remainder of the modeling population.

Question: Outliers

Outliers are anomalous data values. Are there any outliers here?

	C1	C2	C3	C4	C5	C6	C7
	cking	#cking	ADB	NSF	dirdep	SVG	bal
R1	Y	1	468.11	1	1876	Y	1208
R2	Y	1	68.75	0	0	Y	0
R3	Y	1	212.04	0	6		0
R4		.	.	0	0	Y	4301
R5	y	2	585.05	0	7218	Y	234
R6	Y	1	47.69	2	1256		238
R7		1	4687.7	0	0		0
R8	Y	.	.	1	0	Y	1208
R9		.	.	.	1598		0
R10		1	0.00	0	0		0
R11	Y	3	89981.12	0	0	Y	45662
R12	Y	2	585.05	0	7218	Y	234

Answer: Outliers

Outliers are anomalous data values. Are there any outliers here?

	C1	C2	C3	C4	C5	C6	C7
	cking	#cking	ADB	NSF	dirdep	SVG	bal
R1	Y	1	468.11	1	1876	Y	1208
R2	Y	1	68.75	0	0	Y	0
R3	Y	1	212.04	0	6		0
R4		.	.	0	0	Y	4301
R5	y	2	585.05	0	7218	Y	234
R6	Y	1	47.69	2	1256		238
R7		1	4687.7	0	0		0
R8	Y	.	.	1	0	Y	1208
R9		.	.	.	1598		0
R10		1	0.00	0	0		0
R11	Y	3	89981.12	0	0	Y	45662
R12	Y	2	585.05	0	7218	Y	234

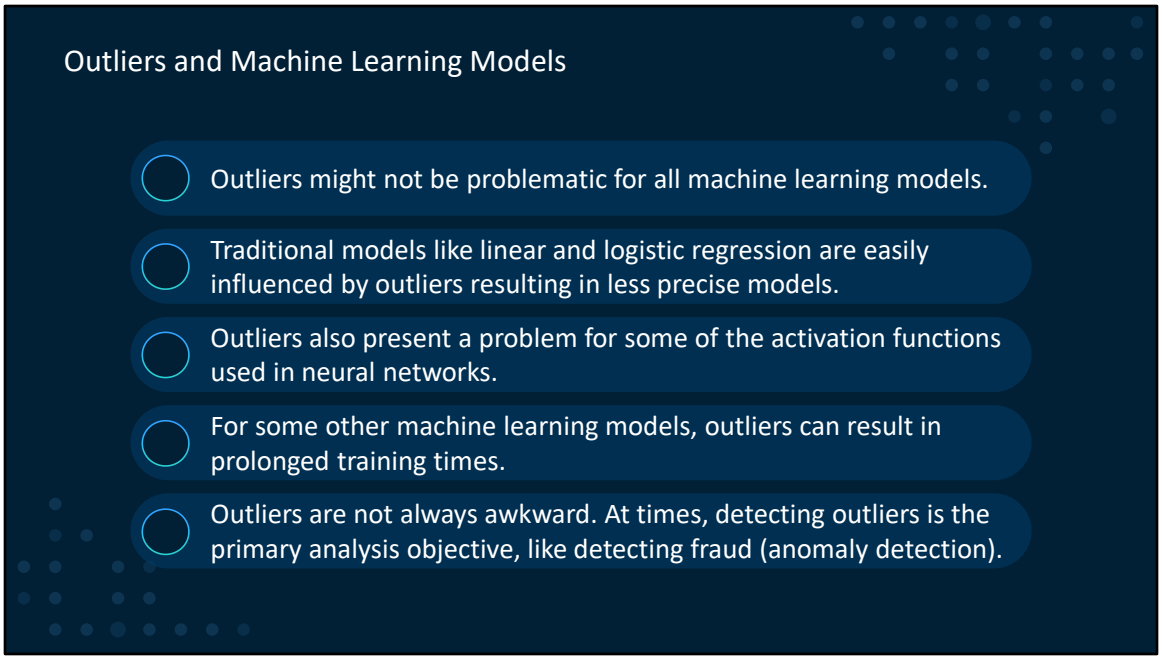
What Is an Outlier?

- An outlier is a data point that is noticeably different from the rest.

- An outlier can easily be detected using
  - visualization techniques: box plot, scatter plot…
  - mathematical functions: IQR, Z-scores…
  - consultation with a domain expert.

In statistics, an *outlier* is an observation point that is distant from other observations. That means an outlier is something that is an odd-one-out or the one that is different from the crowd. For example, if you have a variable as simple as fish size, as illustrated here, the biggest fish seems to be an outlier as it is so much bigger than all the other fish.

There is no one method to detect outliers. However, you can detect them using visualization techniques such as box plots and scatter plots, or with mathematical functions such as IQR and Z-scores. To determine whether a value is an outlier, you must consult a domain expert.



Outliers are generally concerning. However, they might not be problematic for all machine learning models. Some of the machine learning models are at risk to the statistics and distribution of the input variables. Modeling accuracy generally decreases as the percentage of outliers and their magnitude increases.

Regression models are sensitive to extreme or outlying values in the input space. Inputs with highly skewed or highly kurtotic distributions can be selected over inputs that yield better overall predictions. How do you score cases with unusual values? Regression models make their best predictions for cases near the centers of the input distributions. If an input can have (on rare occasion) extreme or *outlying* values, the regression model might not respond appropriately.

Models like neural networks are also affected by extreme or unusual values. This is because some activation functions that do not squeeze the output under certain range (squashing property) result in huge accuracy losses in the presence of outliers.

In supervised models, outliers can mislead the training process resulting in elongated training times or leading to the development of less precise models.

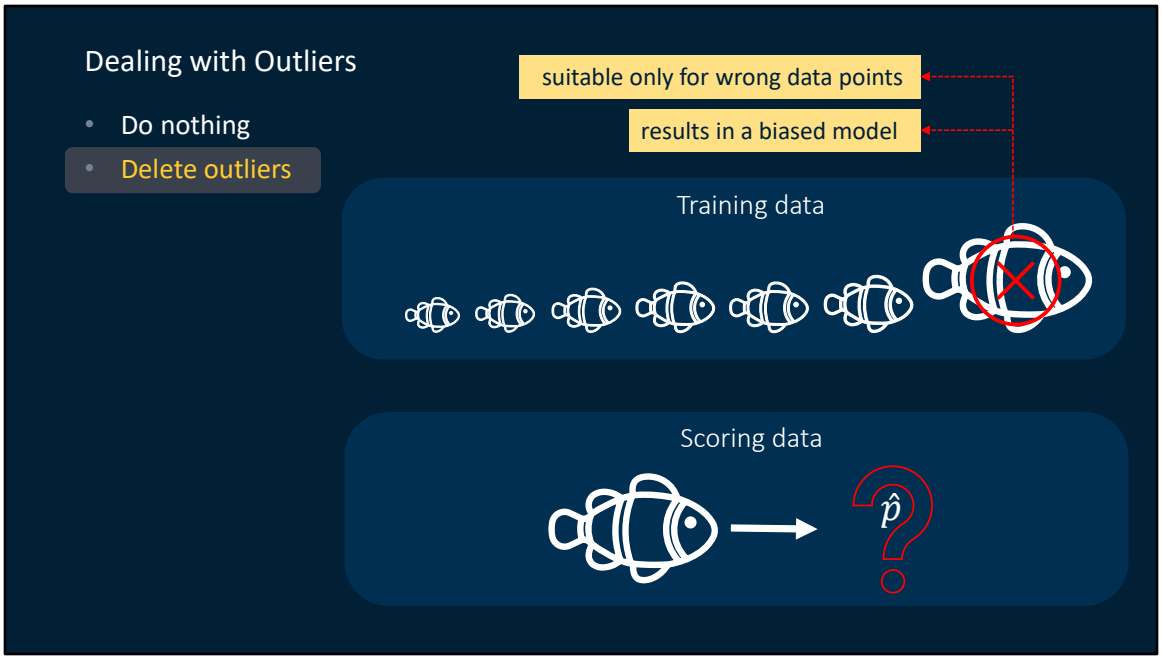
It's not that outliers are always awkward or indicative of a problem. Outliers sometimes can be helpful indicators. For example, in some applications of data analytics like fraud detection. Anomaly detection as it is sometime called becomes important because here, the exception rather than the rule might be of interest. Support vector data description (SVDD) is a common machine learning technique for detecting anomalies.

Dealing with Outliers

- Do nothing

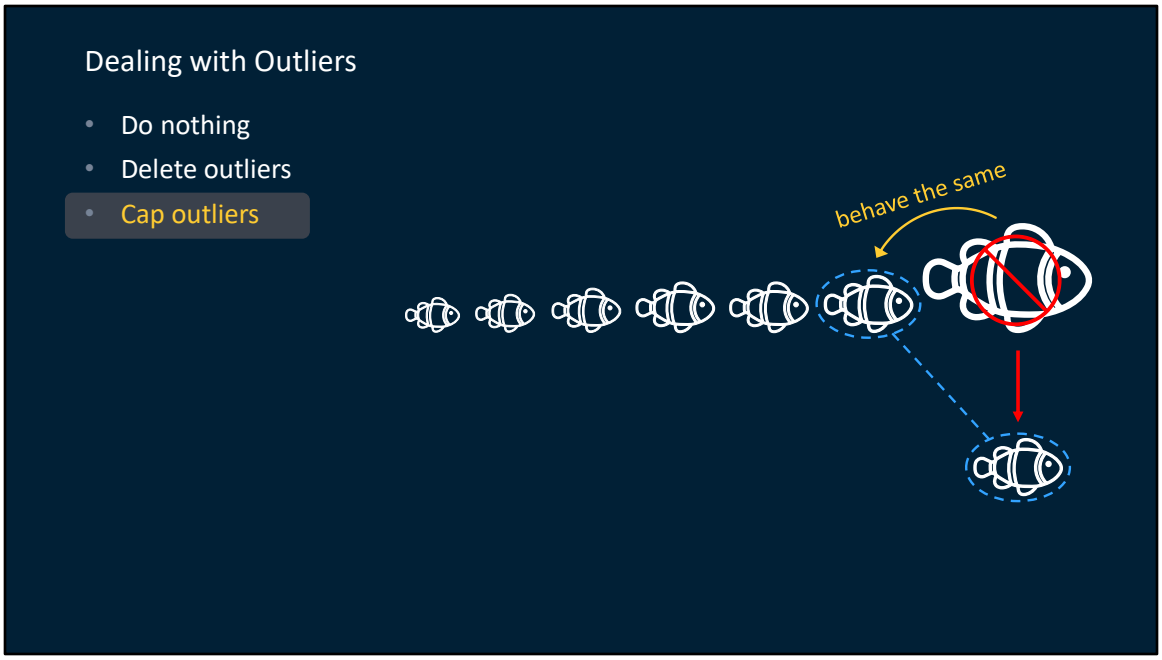
Outliers are data points that differ from the general trend of the data by more than is expected. An outlier might be an erroneous data point, or one that is atypical compared with the rest of the data.

If there's a chance that the outlier will not significantly alter the outcome, you might accept it as is.



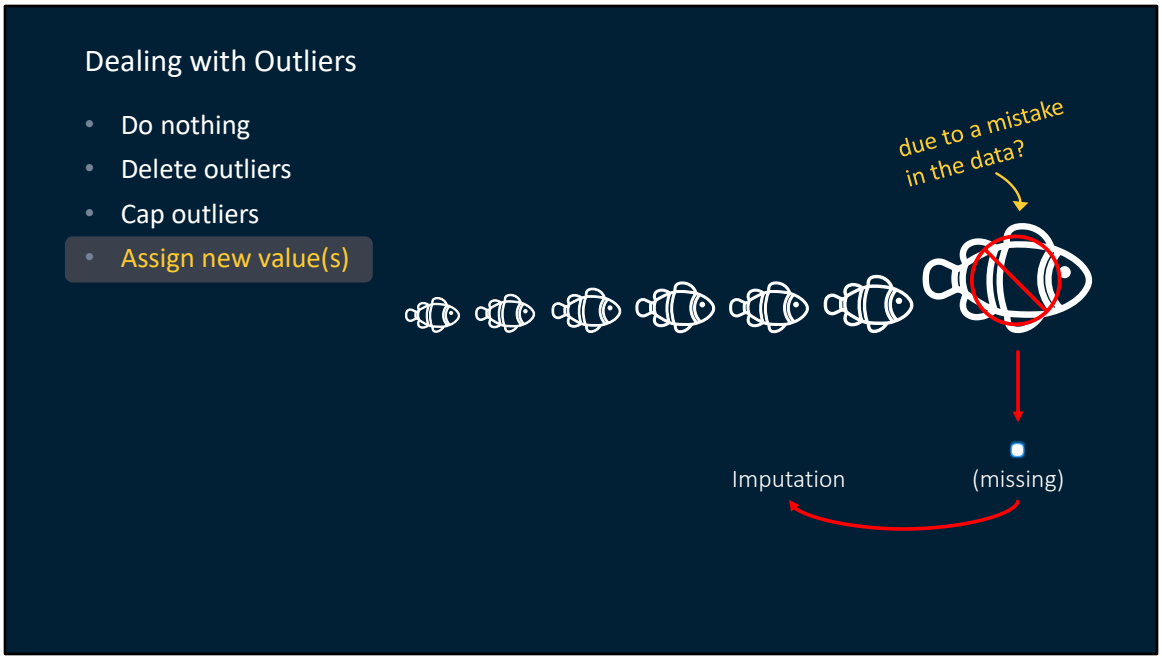
It's easy to completely remove outliers from your data set to stop them from skewing your analysis. However, this temptation is sometimes the wrong choice. You should reserve deletion only for data points that are definitely wrong. For example, you discover that the equipment in the lab was miscalibrated.

Just deleting an outlier from the training data results in a biased model. For example, the model won't be applicable for the largest fish, if it got deleted. What happens when this type of largest fish comes up in the scoring data? Certainly, the generated predicted scores might not be accurate for such cases. Remember, deleting outliers should be the last resort, not the first.

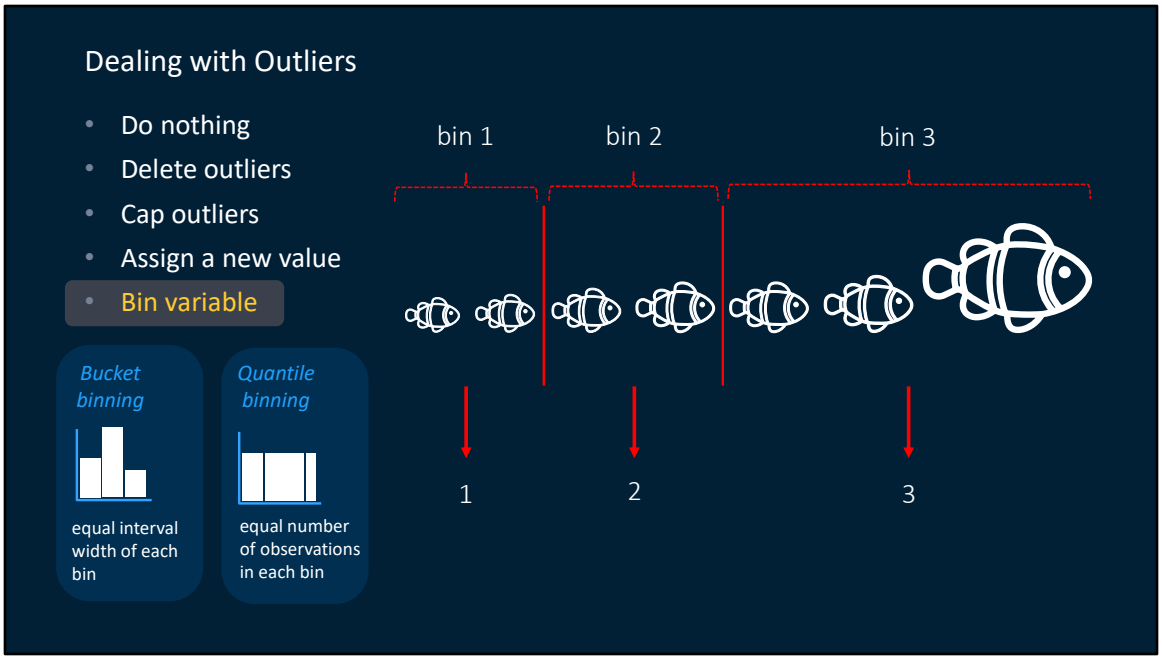


Another way to handle true outliers is to cap them. For example, if you're using age, you might find that people above a certain age behave in the same way as those with a lower age. In this case, you can cap the age value at a level that keeps that intact.

Let's see how capping outliers would look in our fish example. If the largest fish, which is an outlier, behaves like the second-largest fish, you can cap the fish size. Treat all big fish beyond a certain size just like smaller fish.



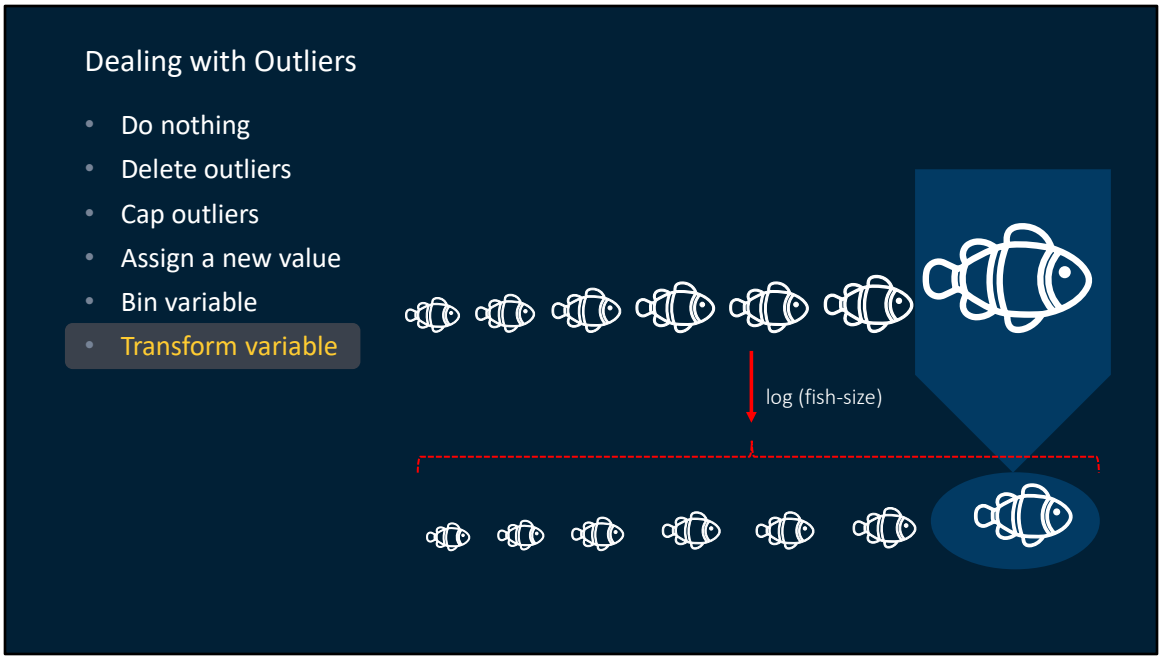
What if an outlier seems to be due to a mistake in your data? In that case, treat it like a missing value, and then try imputing a new value. Common imputation methods include using the mean of a variable or utilizing an estimation model to predict the missing value.



Binning is another effective method for reducing the influence of outliers. The original data values are divided into small intervals known as bins and then they are replaced by a general value calculated for that bin (for example, ranks).

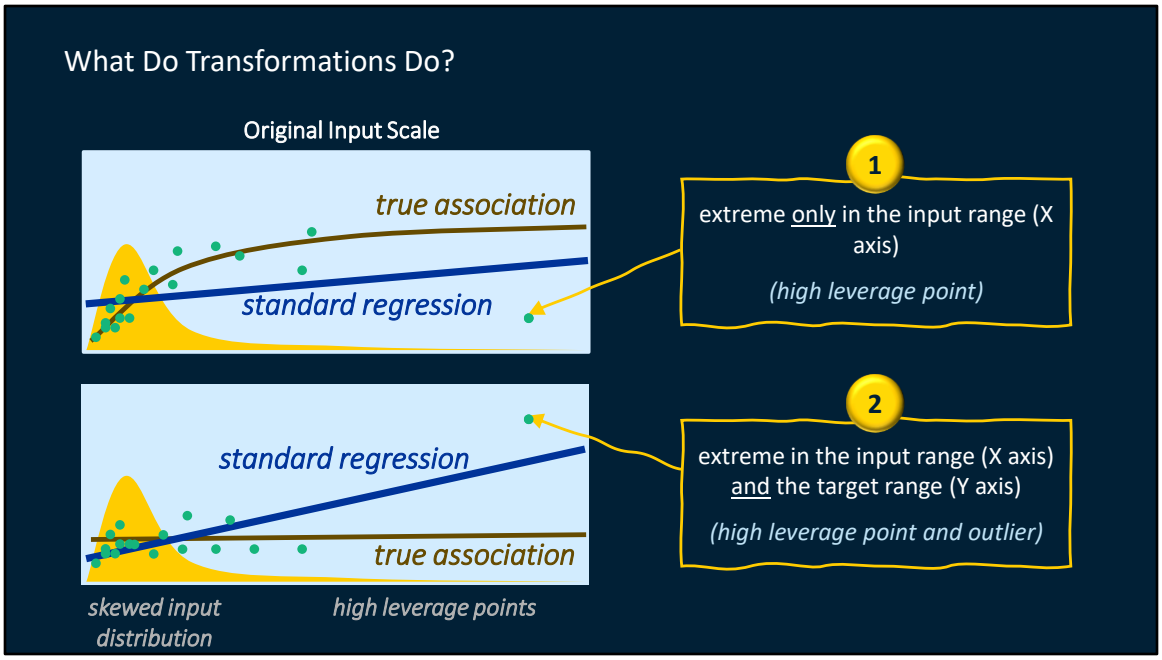
Binning is also useful for classifying missing values, generating variables for detecting nonlinear effects, and reducing categorical variables with a high number of levels to something more manageable.

Equal frequency binning (quantile binning) and equal width binning (bucket binning) are two common binning methods. Bucketing divides observations into equal-length buckets. Within a variable, each bin has the same interval. Quantile binning bins observations into approximately equal-sized bins using the pseudo-quantiles.



A different approach to true outliers could be to try creating a transformation of the data rather than using the data itself. For example, try creating a logarithmic of fish size and working with that new field instead. The influence of the largest fish is then mitigated to a great extent using an appropriate transformation.

When data are skewed or there are outliers, transformation of the data might correct the problem. But how? We'll discuss that next.



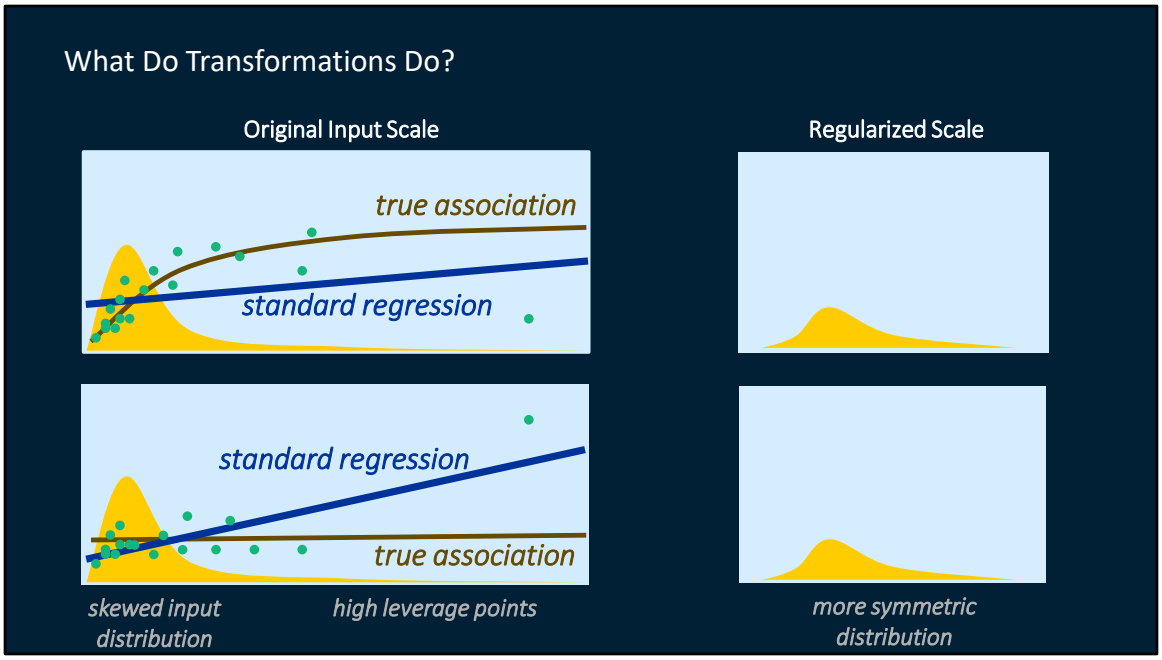
Let's consider classical regression analysis that makes no assumptions about the distribution of inputs. The only assumption is that the expected value of the target is a linear combination of input measurements. Then, why should you worry about extreme input distributions?

There are at least two compelling reasons.

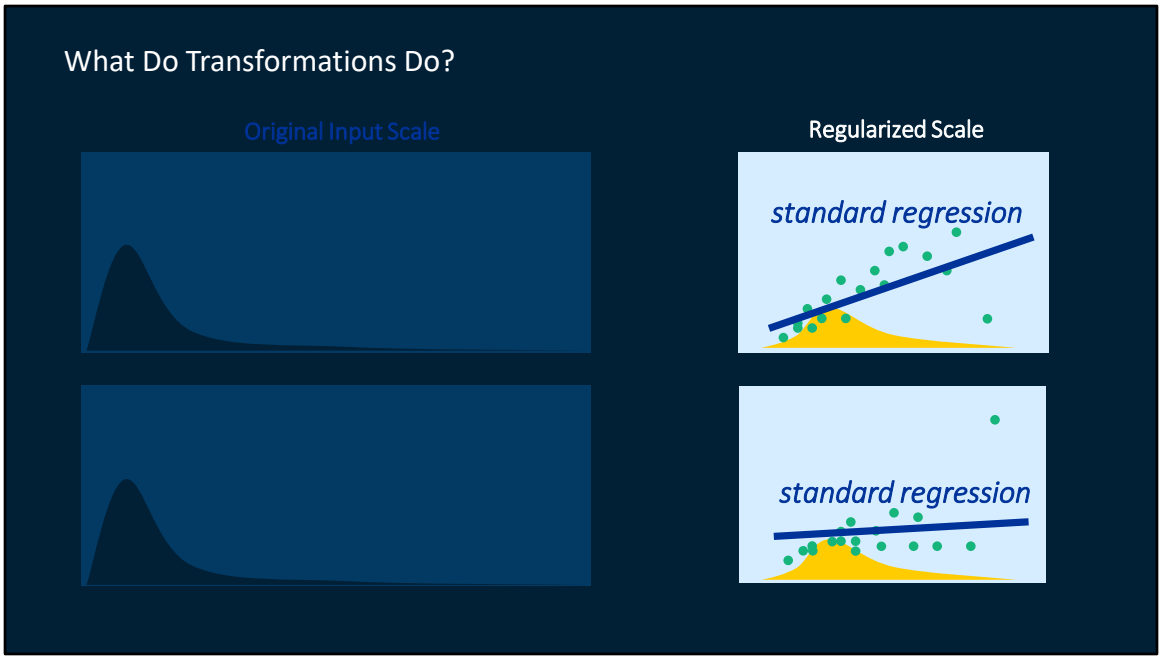
First, in most real-world applications, the relationship between expected target value and input value does not increase without bound. Rather, it typically tapers off to some horizontal asymptote. Standard regression models are unable to accommodate such a relationship. Notice that the extreme observation is a “high leverage” point (only on the input axis) and not an “outlier” (on the target axis).

Second, as a point expands from the overall mean of a distribution, the point has more influence, on model fit. Models built on inputs with extreme distributions attempt to optimize fit for the most extreme points at the cost of fit for the bulk of the data, usually near the input mean. This can result in an exaggeration or an understating of an input's association with the target. Notice that the extreme observation is an “outlier” (on the target axis) as well as a “high leverage” point (on the input axis).

Note that the first concern can be addressed by abandoning standard regression models for more flexible machine learning models. Abandoning standard regression models is often done at the cost of model interpretability, and more importantly, failure to address the second concern of leverage occurs.

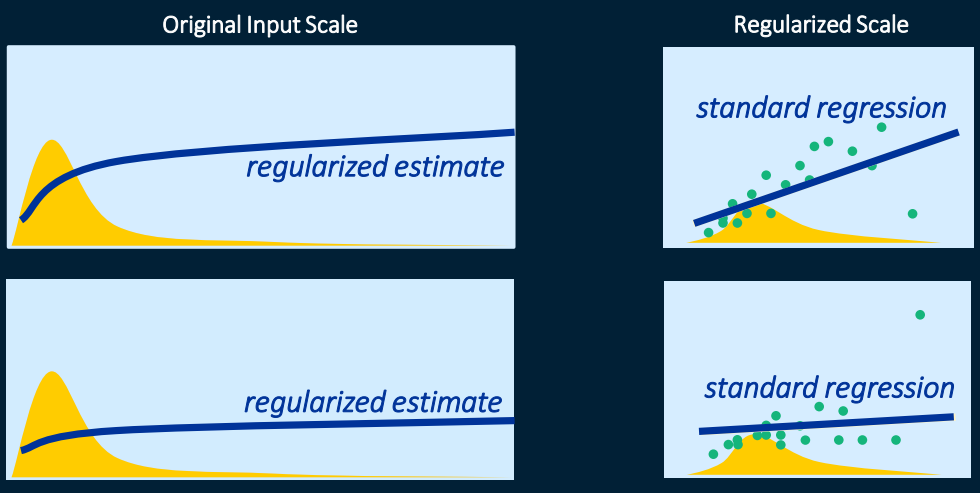


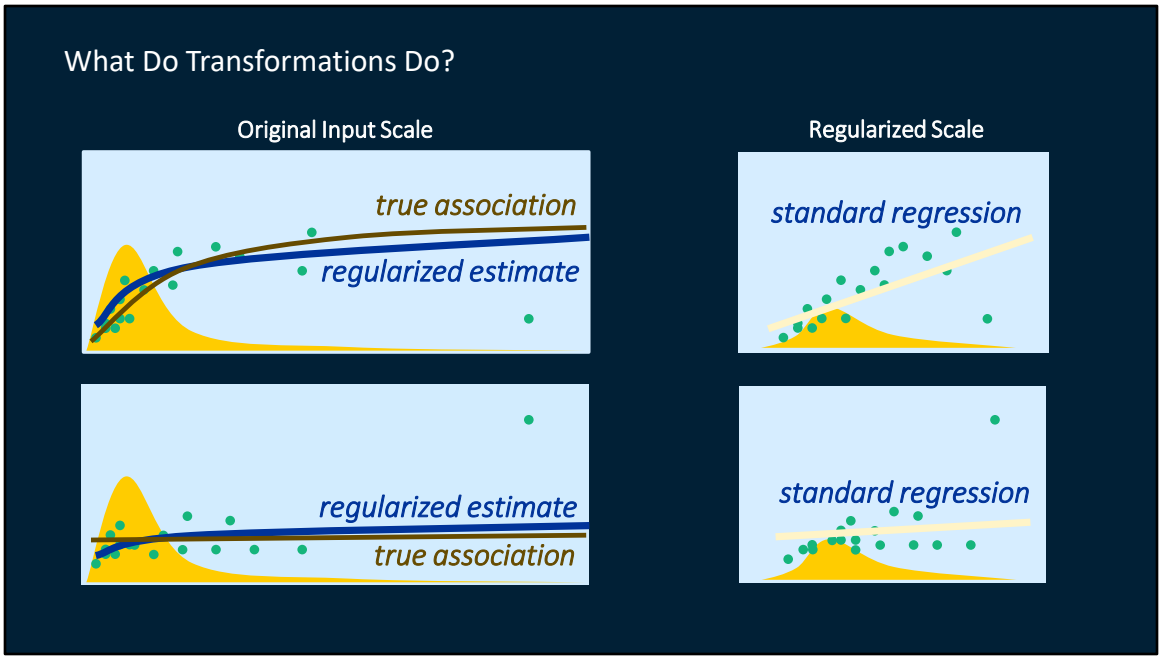
A simpler, and arguably more effective, approach transforms or regularizes offending inputs in order to eliminate extreme values.



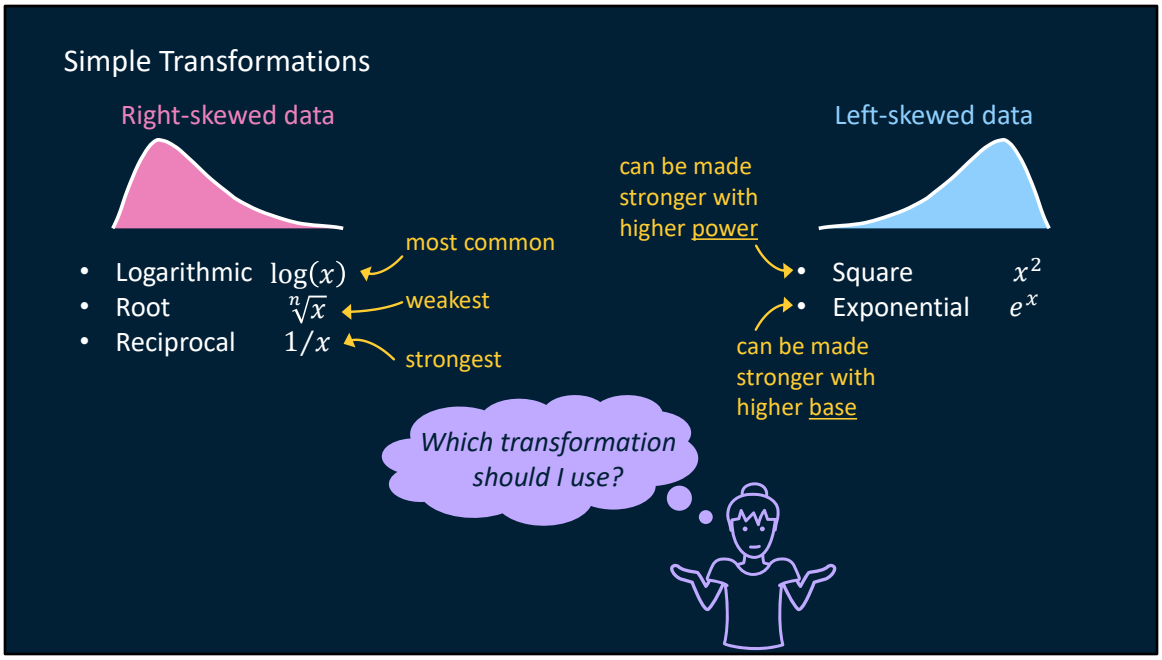
Then, a standard regression model can be accurately fit using the transformed input in place of the original input.

What Do Transformations Do?





Often this can solve both problems mentioned previously. This not only mitigates the influence of extreme cases, but also creates the desired asymptotic association between input and target on the original input scale. The benefit of this approach is improved model performance. The cost, of course, is increased difficulty in model interpretation.

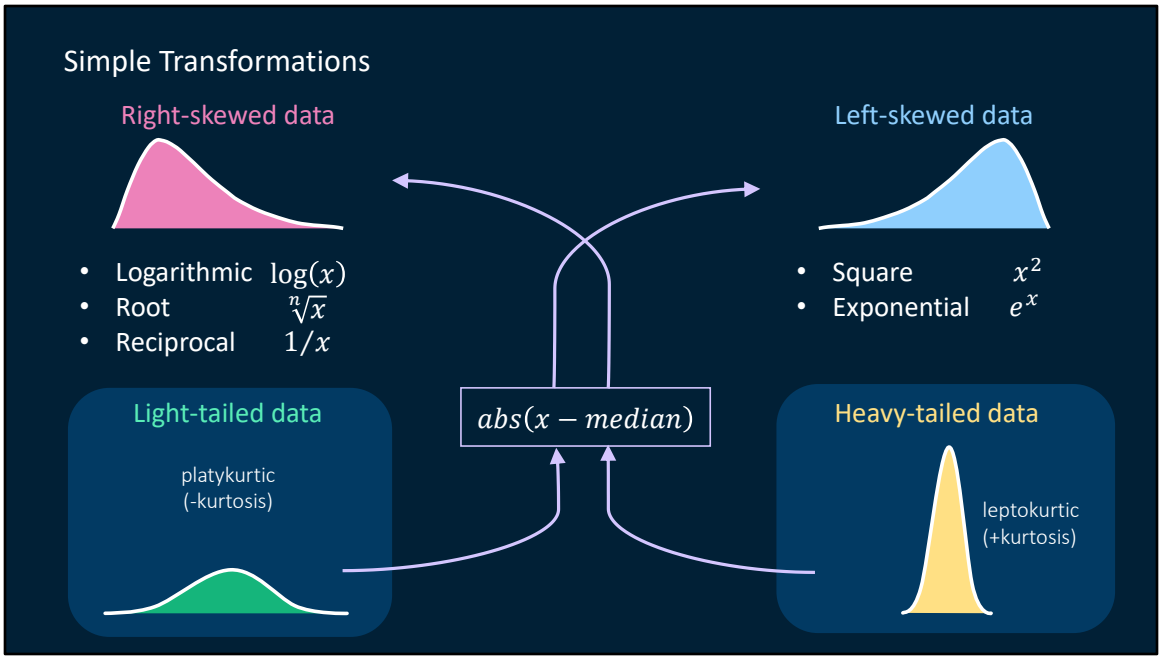


You might be wondering, *which transformation should I use?* Well, this is a million-dollar question, but like other analytical issues, this too does not have a definitive answer.

When data is right-skewed or positively skewed, logarithmic is the most commonly used transformation. The strength of this transformation can be somewhat altered by the base of the logarithm. It cannot be used on negative numbers or 0.

Another transformation is taking the root of the variable. This is the weakest transformation. If your transformation of choice is too strong, you will end up with data skewed in the other direction. The strongest transformation is taking the reciprocal of the variable. This transformation is stronger with higher exponents.

If your data is left-skewed or negatively skewed, you can use square and exponential transformations. Square transformation can be made stronger with higher power. Similarly, exponential transformation will become stronger with higher base.



Deviations of the tail from normality are usually less critical than skewness and might not need transformation after all. However, what if you still want to transform light- and heavy-tailed data? How would you do this?

First, subtract the data points from the median to set your data to a median of 0. Then use an appropriate transformation for skewed data on the absolute deviations from 0 on either side. For heavy-tailed data, use transformations for right skewed data to pull in on the median. For light-tailed data, use transformations for left-skewed data to push data away from the median.

Note that categorical inputs also warrant appropriate transformations. However, this is not discussed in this course.

Question: Variable Transformations

Which statement below is True about transformations of input variables in a regression analysis?

a. They are never a good idea.
b. They help model assumptions match the assumptions of maximum likelihood estimation.
c. They are performed to reduce the bias in model predictions.
d. They are typically done on nominal (categorical) inputs.

Answer: Variable Transformations

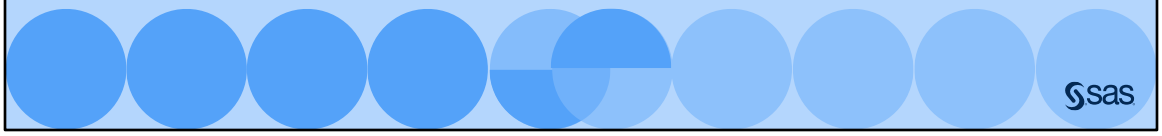
Which statement below is True about transformations of input variables in a regression analysis?

a. They are never a good idea.
b. They help model assumptions match the assumptions of maximum likelihood estimation.
c. They are performed to reduce the bias in model predictions.
d. They are typically done on nominal (categorical) inputs.

Answer: (c)
Transformations are useful when you want to improve the fit of a model to the data. For example, transformations can be used to stabilize variances, remove nonlinearity, improve additivity, and correct nonnormality in variables, thereby reducing the bias in model predictions.

Demo: Transforming Inputs

To explore this learning resource in greater detail, please visit the course page on learn.sas.com. There, you'll find comprehensive course materials, tutorials, and updates to support your learning journey.



Problems with Too Many Variables

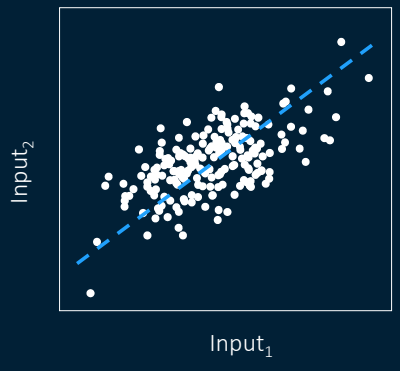
Data Difficulties and Modeling Issues

- Visualizing big data
- Processing noisy data
- **Handling high dimensional data**
- Dealing with distinct variability in inputs
- Using available data judiciously
- Adjusting model flexibility and complexity
- Understanding learning and tuning of models
- Making models explainable

There is a natural tendency to think that the more features that you have to describe the observations in your data, the more information you have about those observations. Thus, the more accurate your models should be. However, that is true only to a point because having more columns is not always helpful.

Problems with Too Many Variables

- Redundancy (correlated inputs)

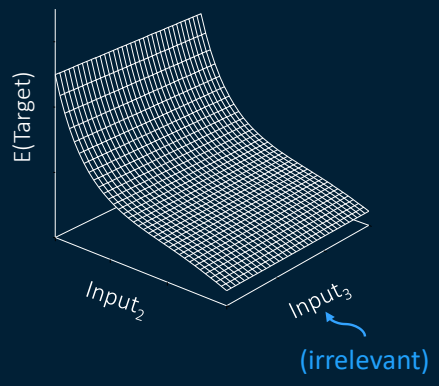


When the number of potential input variables is large, some of these inputs can be redundant. Redundant inputs complicate model construction. One consequence of input redundancy might be model instability.

Here Input1 and Input2 are correlated and thus redundant. Knowing the value of *Input*1 gives you a good idea of the value of *Input*2. Generally, there would not be any gain in having both into the model. Redundancy is discussed in a previous lesson.

Problems with Too Many Variables

- Redundancy
- Irrelevancy
  (no association with target)



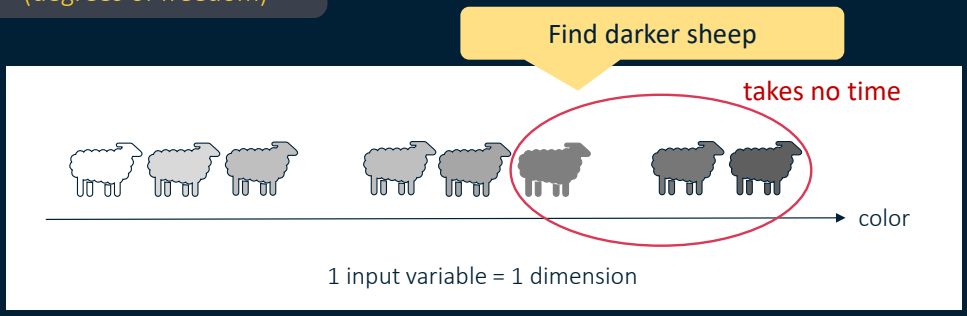
Irrelevancy creeps in when you have many input variables. An irrelevant input does not provide information about the target and therefore does not contribute to the model. Adding irrelevant features unnecessarily complicates and increases the expense of training models.

Here, predictions change with Input2 but much less with input Input3, thus, Input3 is irrelevant. There would not be any gain in adding Input3 into the model. Irrelevancy is discussed in a previous lesson.

Problems with Too Many Variables

- Redundancy
- Irrelevancy
- Curse of Dimensionality (degrees of freedom)

◯ number of features can exceed the number of observations
◯ calculations become extremely difficult

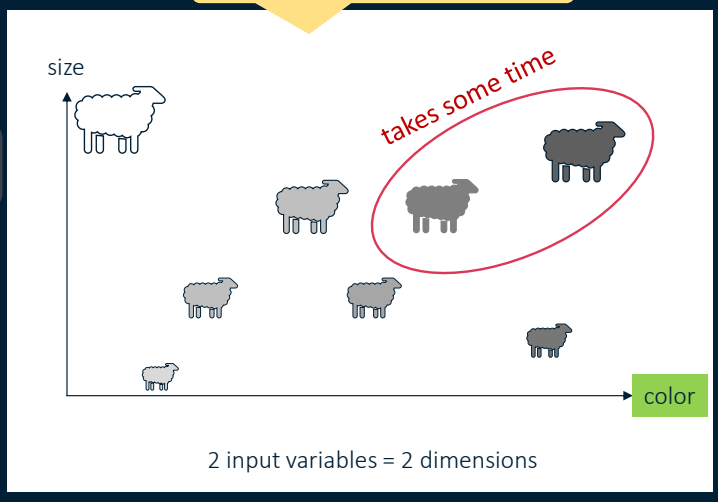


The *dimension* refers to the number of input variables (actually input degrees of freedom). We generally encounter large numbers of input variables in machine learning. The number of variables often has a greater effect on computational performance than the number of cases. Curse of dimensionality means that the number of dimensions is staggeringly high, so high that calculations become extremely difficult, and the number of features can exceed the number of observations.

Consider a simple example of finding darker sheep. The only available feature is color of sheep. It's pretty easy for you to select perhaps the three most gray sheep. Here, data density is very high. Hence, you will not face any difficulty to trace these sheep.

Problems with Too Many Variables

- Redundancy
- Irrelevancy
- Curse of Dimensionality (degrees of freedom)

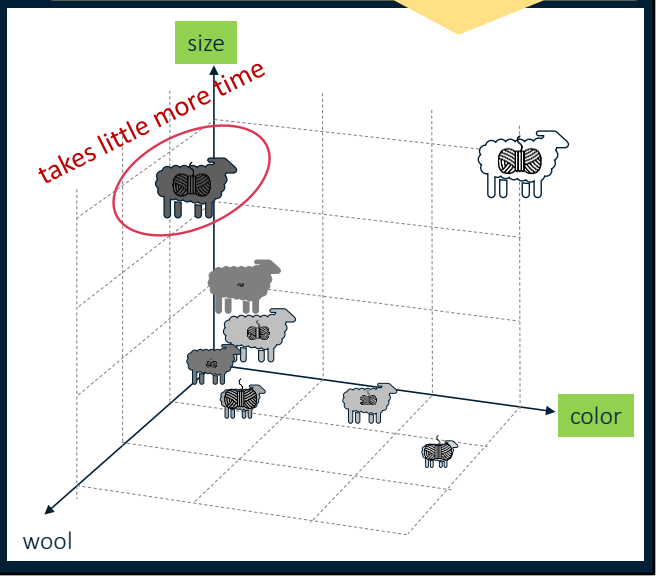


If you add another feature such as size of sheep to its color, then the complexity of a data set increases. If you need to find the darker and bigger sheep, you will take a little more time as compared to previous scenario. Perhaps you to select two sheep in this case. Thus, high dimensionality limits the ability to explore and model the relationships among the variables.

Problems with Too Many Variables

- Redundancy
- Irrelevancy
- Curse of Dimensionality (degrees of freedom)

3 input variables = 3 dimensions

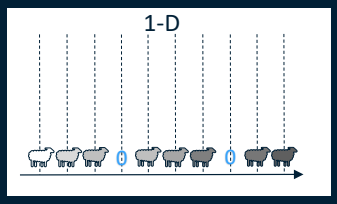


If you add one more feature, wool density to its size and color, the complexity of a data set increases rapidly with increasing dimensionality (Breiman et al. 1984) . If you need to find the darker, bigger sheep with high wool density, you will take a little more time as compared to previous scenario. Perhaps you to select one sheep in this case. This added dimensionality further limits your ability to explore and model the relationships among the variables.

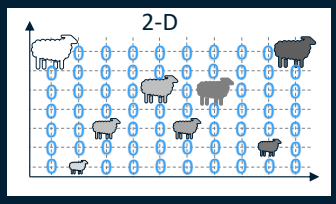
Problems with Too Many Variables

- Redundancy
- Irrelevancy
- Curse of Dimensionality
- Sparseness
  (many zeros in computation matrix)

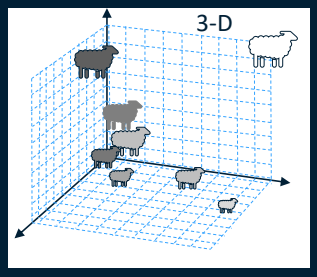
Common in applied machine learning:
→ count data
→ data encodings that map categories to counts
→ natural language processing



8/10=80%



8/(10×10)=8%



8/(10×10 ×10) =0.8%

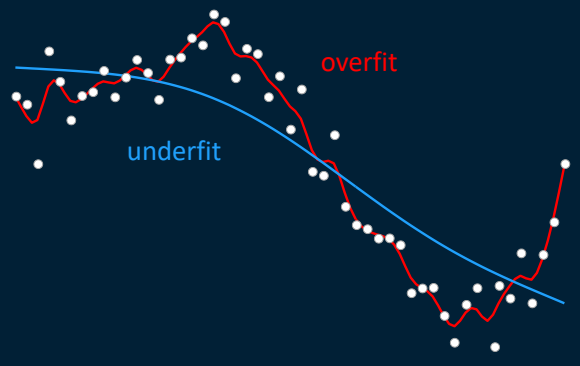
Sparseness addresses the so-called *curse of dimensionality*, which tells us that as you add features, you are introducing greater sparsity to your observations in that feature space. Sparseness refers to a matrix of numbers that includes many zeros or values that will not significantly impact a calculation.

In this simple example, where in one dimension you have the space well covered with eight data points that covers 80% of the space, when we add another dimension, those same eight data points cover a much smaller percentage of the feature space, only 8%. This sparsity grows exponentially with each feature added. The practical meaning of sparseness is that there is much space between the data points in the high-dimensional space built up by all the used input variables. The data points have a long way to the neighbors—that is, they are lonely. Many variables consist of one or two values (for example, 0 or 1). This is a source of problems for many machine learning methods as such matrices are computationally expensive.

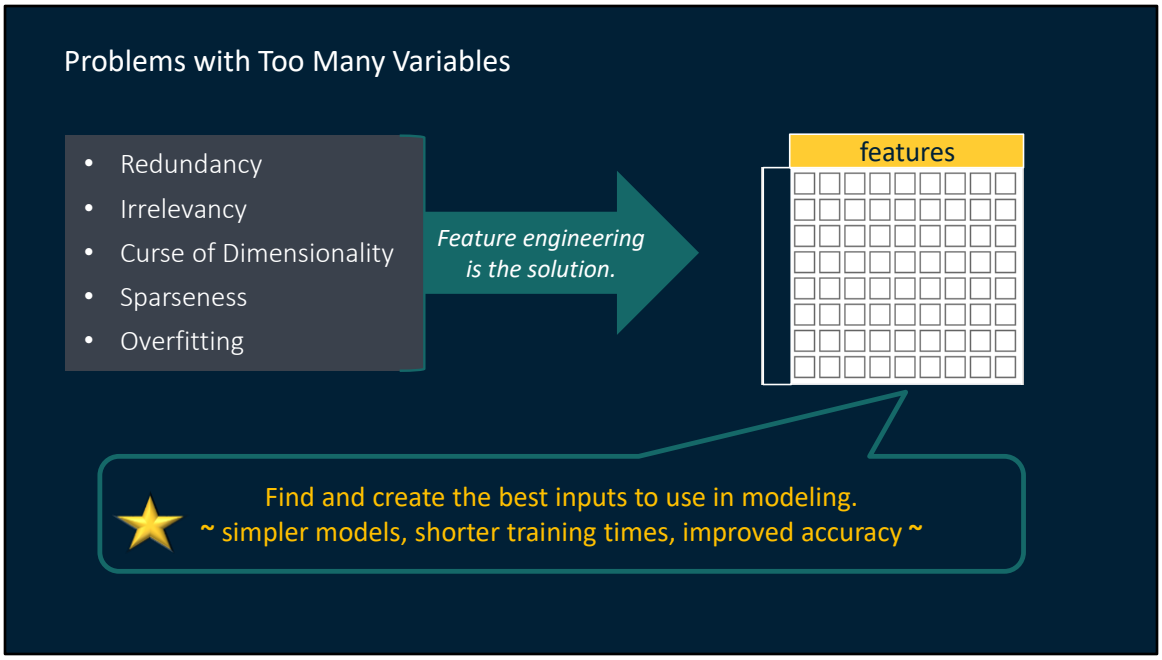
Large sparse matrices are common in applied machine learning, such as in data that contains counts, data encodings that map categories to counts, and even in natural language processing. The solution to representing and working with sparse matrices is to use an alternate data structure to represent the sparse data.

Problems with Too Many Variables

- Redundancy
- Irrelevancy
- Curse of Dimensionality
- Sparseness
- Overfitting
  (model not generalize well)



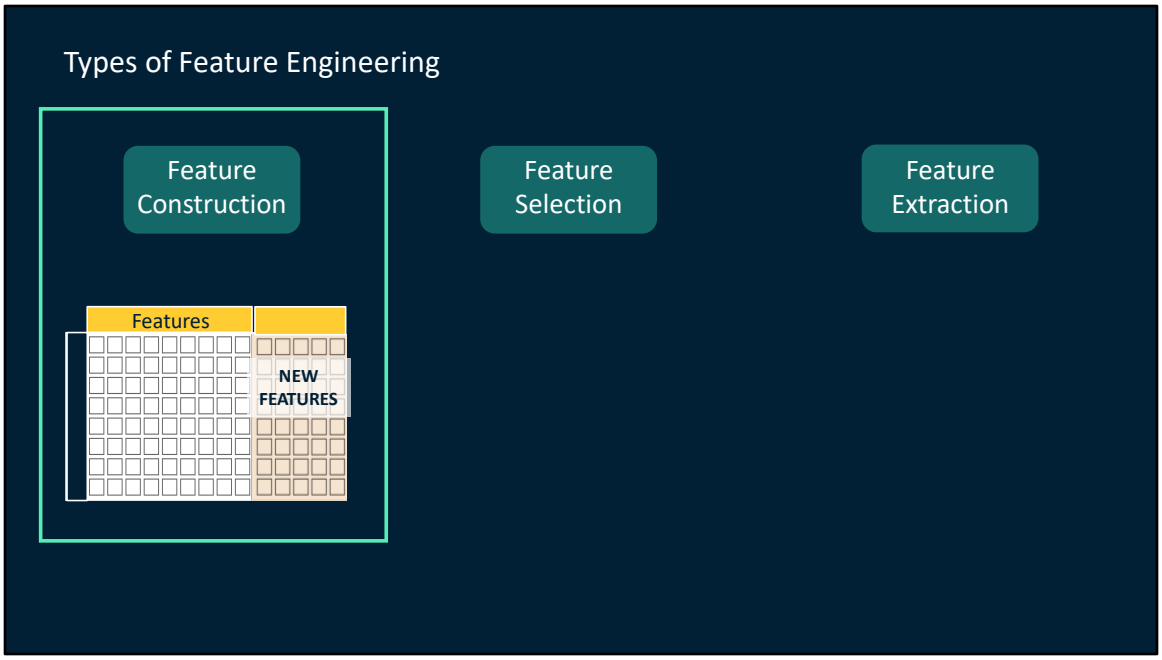
Overfitting occurs when the model fits the training data too well but cannot generalize to unseen data sets. Overfitting was discussed in the previous lesson. Overfitting will be discussed in a broader sense in the next section.



A critical part of the success of a machine learning project is producing a good set of features to train on. You need to include features to characterize the observations, but it is important to be selective about the features that you choose to incorporate in your modeling process. For example, with the curse of dimensionality, the amount of data that you would need to collect to get a meaningful predictive model would be enormous. Therefore, methods are needed to transform the data from this high-dimensional space to fewer dimensions. Feature engineering is the process of determining the best set of inputs to use in building predictive models.

In the context of the entire analytics life cycle, feature engineering is typically done as part of the data preprocessing leading into modeling. It is a big part of the 80% project time that you hear about being spent in preparation for modeling. And it is a very important step that can make a big difference in the accuracy and effectiveness of your models.

You need great features that describe the structures inherent in your data. The reduced dimensionality of the data set can result in simple, more accurate models that train in less time.



Generally, feature engineering falls into three main categories: construction, selection, and extraction of features.

In machine learning applications, the variables relevant to the analysis rarely come prefabricated. Therefore, they must be created. This is not so much an application of analytical techniques as it is a manual assessment of the existing features, ensuring a clear understanding of the problem being addressed, and then augmenting the data set by aggregating existing features into fewer, and possibly more representative, features or even decomposing individual features into multiple features that better represent the problem.

Types of Feature Engineering

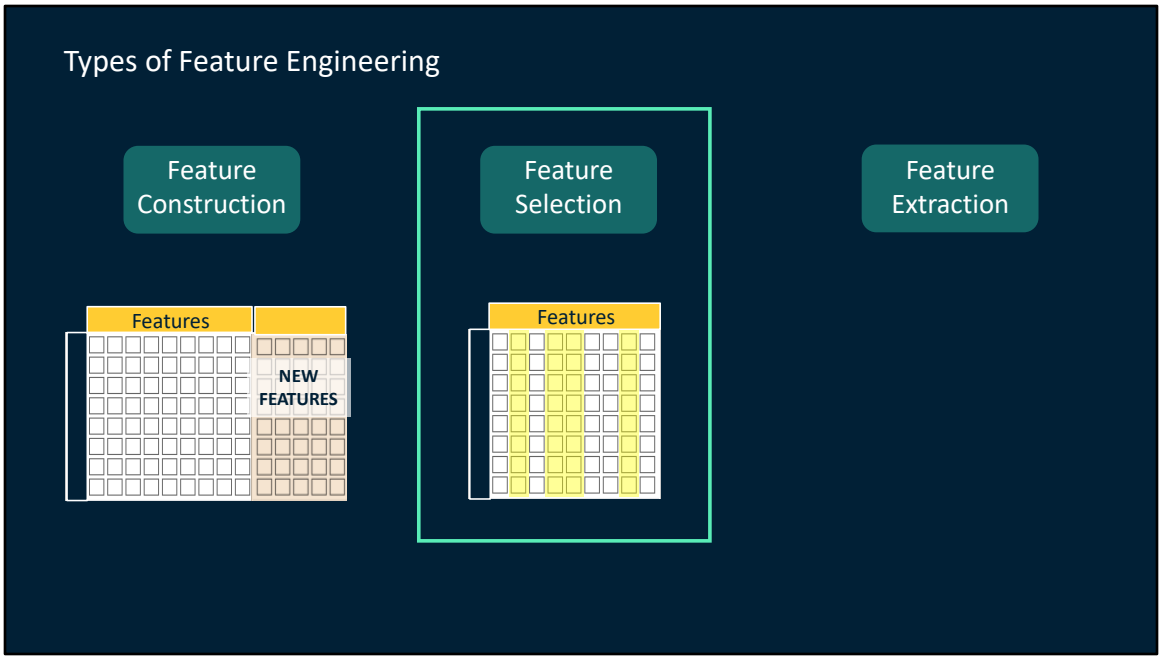
Feature Construction Example

Claim Date	Accident Date-Time
11NOV22	102322/12:38
22DEC21	012321/01:42
26APR21	042321/03:05
02JUL20	070220/06:25
08MAR22	123021/18:33
15DEC22	061222/18:12
09NOV20	110520/22:14

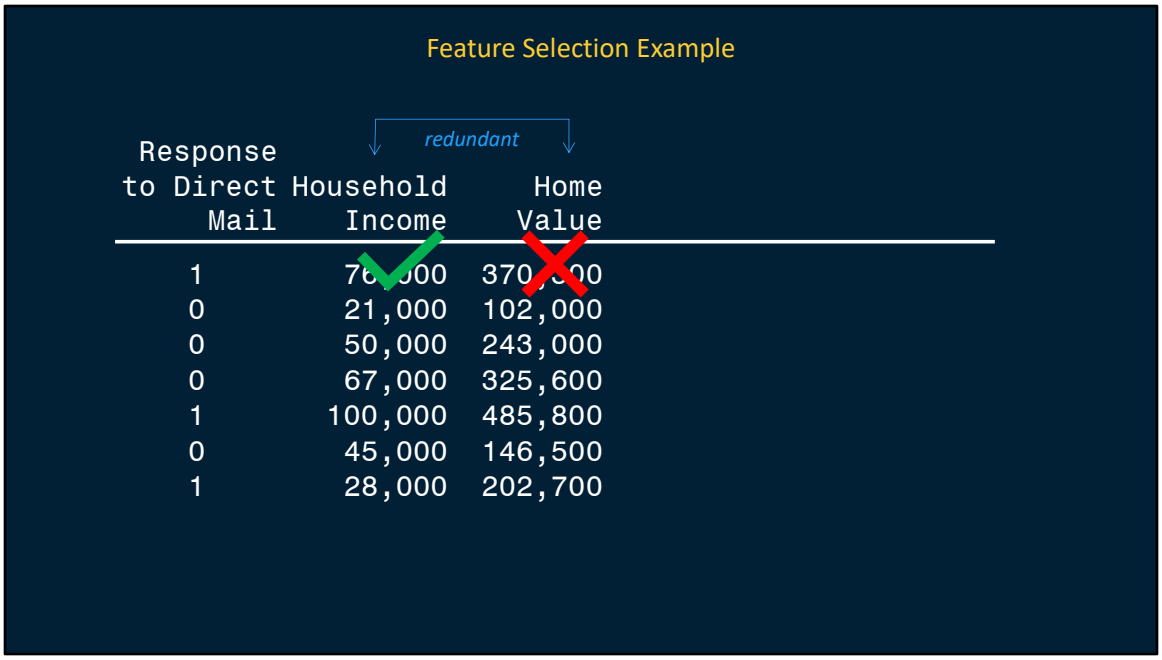
derived features

Delay	Season	Dark
19	fall	0
333	winter	1
3	spring	1
0	summer	0
69	winter	0
1816	summer	0
4	fall	1

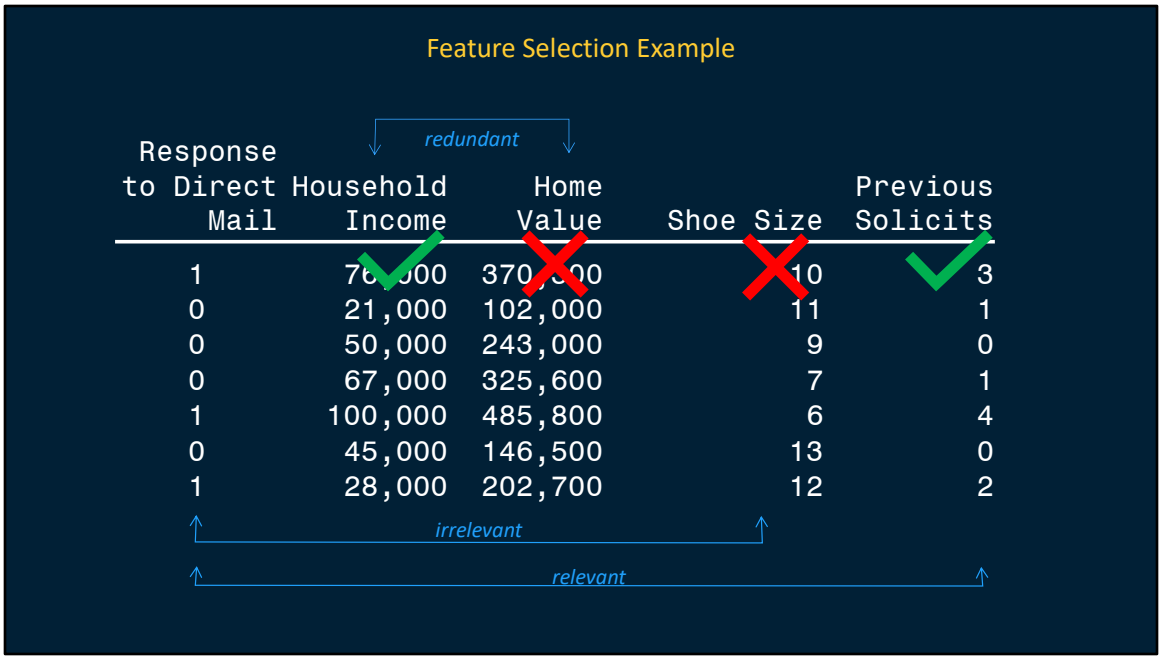
For example, the date on which an auto accident took place and an insurance claim was filed might not be useful predictors of fraud. Derived variables such **Delay**, the time between the two events, might be more useful. Similarly, **Season** and **Dark** (whether it was dark at the time of accident) would contribute more to checking the legitimacy of the insurance claim.



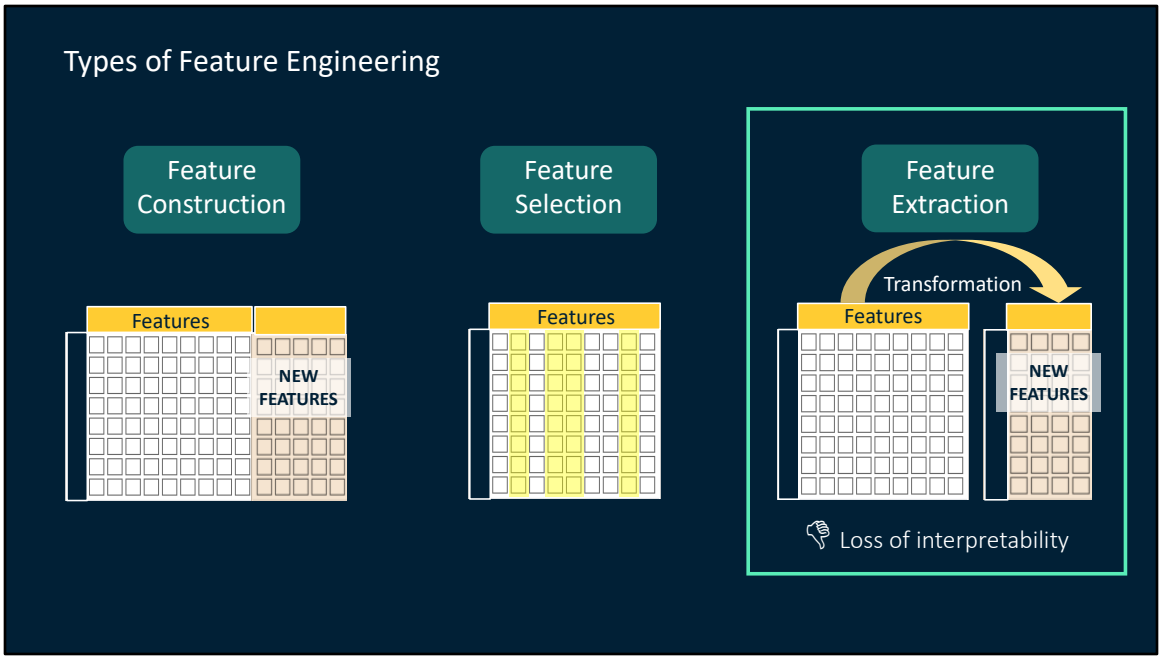
Feature engineering is essentially referring to applying techniques to reduce the feature set. Feature selection, that is, reducing the number of inputs, is the obvious way to thwart the curse of dimensionality. Unfortunately, reducing the dimension is also an easy way to disregard important information. The two principal reasons for eliminating a variable are redundancy and irrelevancy. A redundant input does not give any new information that was not already explained by other inputs. An irrelevant feature does not provide information about the target. Various criteria can be used to filter the set of features down to the most important ones.



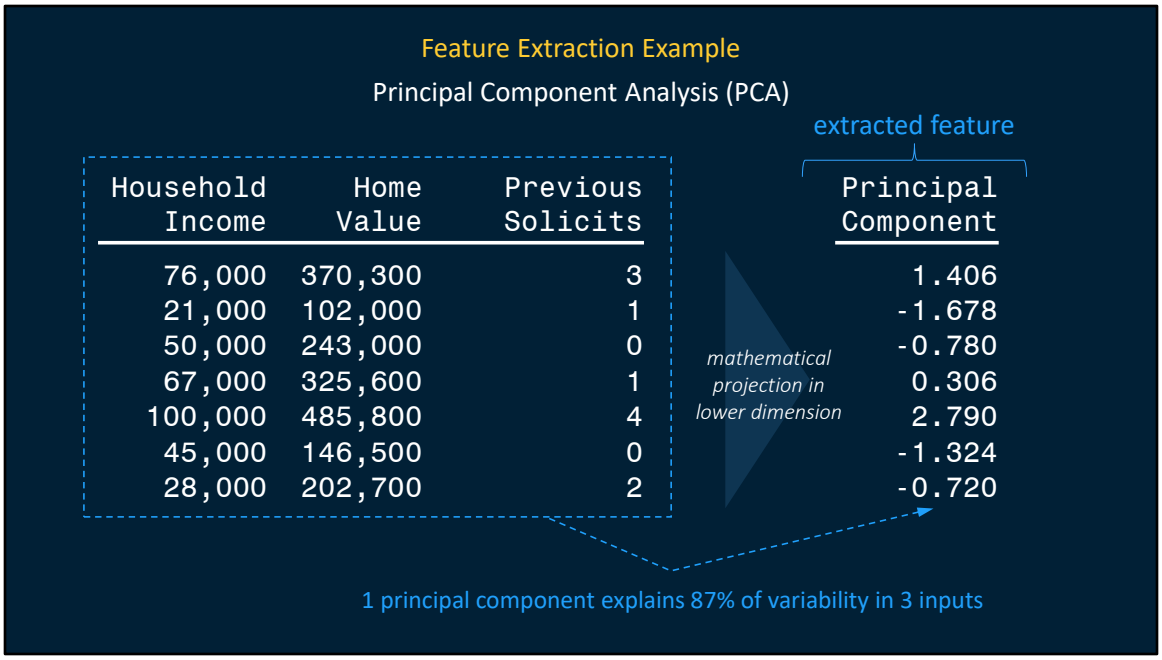
Illustrated here is a simple example of a direct marketing campaign where Response to Direct Mail is the binary target. In this example, Household Income and Home Value are obviously redundant features. Knowing Household Income gives you a good idea of Home Value. Thus, only one of them should be selected.



In addition, predictions change with the Previous Solicits feature, but not with Shoe Size. Shoe Size does not provide information about the target. So, Previous Solicits should be selected.



Feature extraction is another technique to reduce the feature set. Feature extraction transforms the data in a way such that fewer columns can be used to represent most of the information contained in the original features. Feature extraction derives new features, which leads to lower-dimensional representation of the feature information. With feature extraction techniques, the newly derived features are not interpretable in the context of the original problem. However, this is very often accepted as simply another layer of the model, representing the input to output relationships, and the ultimate benefits realized can be significant.



Principal component analysis (PCA) is a common feature extraction method in machine learning. Practically, PCA converts an original matrix of features into a new data set of fewer features. PCA creates as many new features (called the *principal components*) as the number of input variables. However, you often select only the first few principal components to represent your data. That is, it reduces the number of features by constructing a new, smaller number variables that capture a significant portion of the information found in the original features. Here, three features are represented by the first principal component, and it explains 87% of the variability of the original input space.

Singular value decomposition and *autoencoders* are some other feature extraction methods.

Question: Feature Engineering

The key difference between 'selection' and 'extraction' is that feature selection keeps a subset of the original features and feature extraction creates brand new ones.

a. True
b. False

Answer: Feature Engineering

The key difference between 'selection' and 'extraction' is that feature selection keeps a subset of the original features and feature extraction creates brand new ones.

a. True
b. False

Answer: True

Some variable reduction methods (like decision trees, regressions) use the original features as inputs into subsequent models (feature selection). Others (like principal component analysis) use combinations of the original features as inputs into subsequent models (feature extraction).

Demo: Performing Feature Selection

To explore this learning resource in greater detail, please visit the course page on learn.sas.com. There, you'll find comprehensive course materials, tutorials, and updates to support your learning journey.

Distinctly Scaled Variables

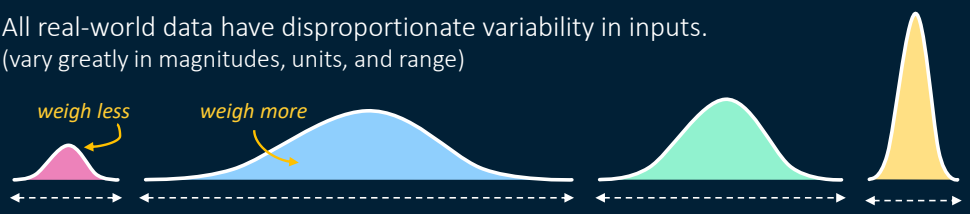
Data Difficulties and Modeling Issues

- Visualizing big data
- Processing noisy data
- Handling high dimensional data
- **Dealing with distinct variability in inputs**
- Using available data judiciously
- Adjusting model flexibility and complexity
- Understanding learning and tuning of models
- Making models explainable

Next, we'll continue our discussion of data difficulties. Let's address how to deal with distinctly scaled variables.

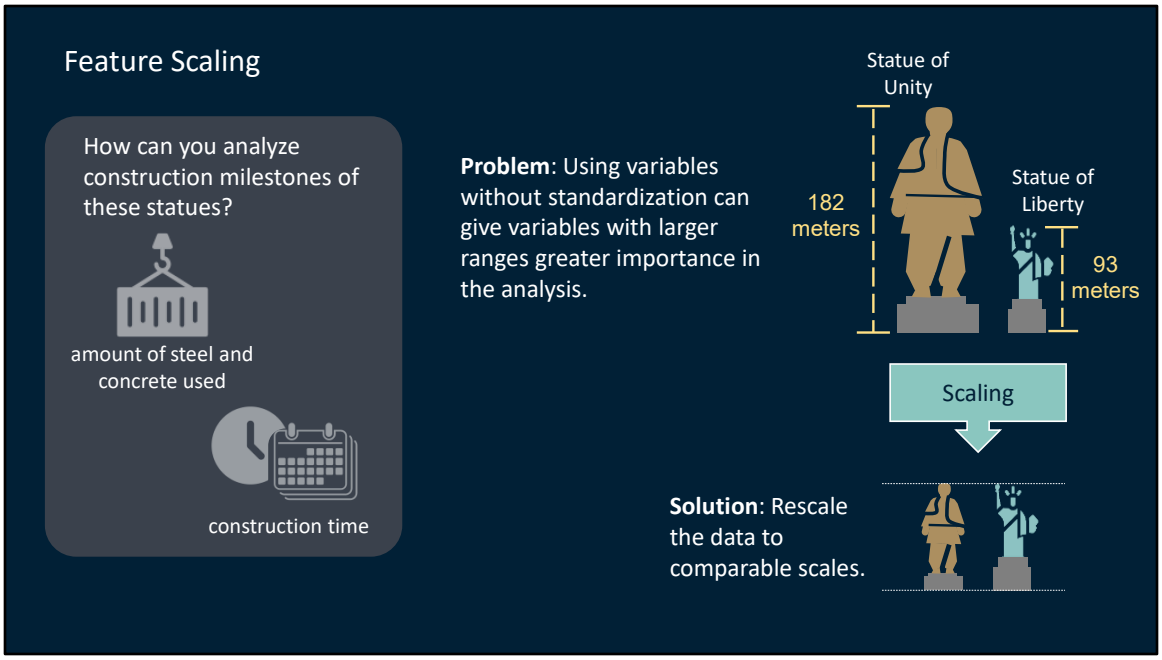
Distinctly Scaled Variables

- All real-world data have disproportionate variability in inputs.
  (vary greatly in magnitudes, units, and range)



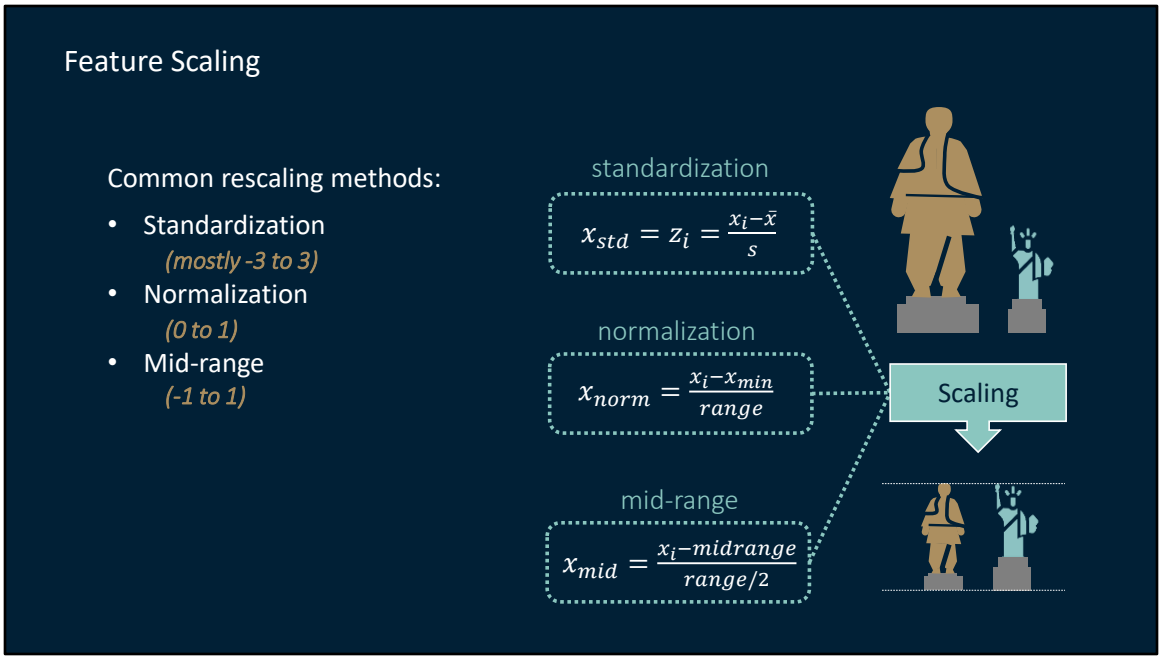
- When the features of a data set differ greatly in scale, models that are sensitive to the scale of the inputs would be affected:
  - slower model training
  - imprecise parameter estimates
  - arbitrary weight to predictors
  - model instability

Ensuring that features are within a similar scale is imperative. However, most of the time your data set will contain features that vary greatly in magnitudes, units, and range. Many machine learning algorithms are sensitive when the data is not scaled evenly. It might result in slower training times, imprecise estimates of parameters, and unstable models. The features with high magnitudes will weigh in a lot more in the distance calculations than features with low magnitudes. To suppress this effect, we need to bring all features to the same level of magnitude. This can be achieved by *scaling*.



Imagine you are tasked with analyzing construction milestones of the Statue of Unity in India with the Statue of Liberty in the United States. You just need to remember that the statues are not of the same height. The Statue of Unity is 182 meters tall whereas the Statue of Liberty is 93 meters. You cannot compare their construction-related milestones such as construction time or amount of steel and concrete used unless you make them similar in height.

Model stability and parameter estimate precision are influenced during multivariate analysis when multi-scaled variables are used. For example, in boundary detection, a variable that ranges between 0 and 100 will outweigh a variable that ranges between 0 and 1. Using variables without standardization can give variables with larger ranges greater importance in the analysis. Rescaling the data to comparable scales can prevent this problem.

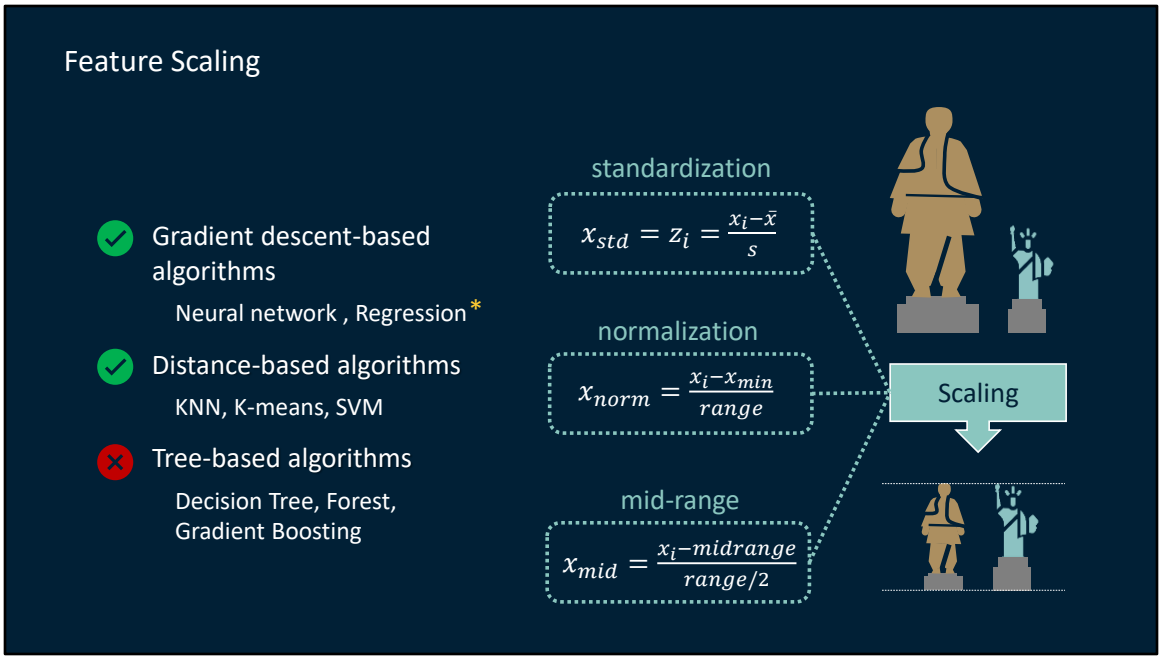


There are many rescaling methods available. Listed here are the ones commonly used in machine learning.

Standardization is the method of rescaling your data to have a mean of zero and a standard deviation of one. A standardized variable is sometimes called a Z-score or a standard score. Here's the equation to standardize your data, where x_i is the original value, *xbar* is the variable's mean, and *s* is the variable's standard deviation. The absolute value of z_i represents the distance between the raw score and the population mean in units of the standard deviation. z_i is negative when the original value is below the mean, positive when above. z_i is between negative 3 and 3 for most data.

Normalization is another rescaling method with many meanings in statistics and statistical applications. Most commonly, normalizing rescales numeric data between 0 and 1 using the equation shown here, where x_{min} is the variable's minimum value, and range is the difference between variable's maximum value and minimum value.

Normalizing, or range standardization, is the basis for another type of standardization known as *midrange*. Midrange is calculated as $(x_{max} + x_{min})/2$. Midrange scaling transforms a variable so that its range is -1 to 1 and its midrange value is 0.



Some machine learning algorithms are sensitive to feature scaling and others are virtually invariant to it.

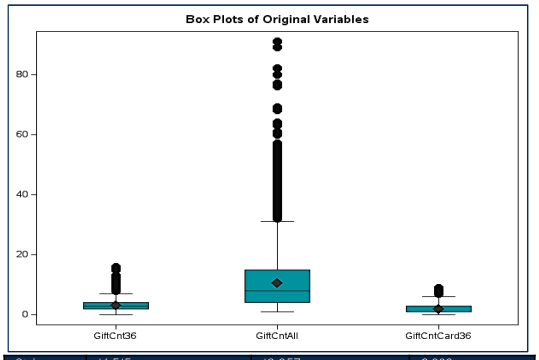
Machine learning algorithms like neural network - and, at times, linear regression, logistic regression, and so on, that use gradient descent as an optimization technique - require data to be scaled. When performing regression analysis, standardizing multi-scale variables can help reduce multicollinearity issues for models containing interaction terms. Standardizing continuous predictor variables in a neural network is extremely important. Having features on a similar scale can help the gradient descent converge more quickly toward the minima. *Note that if you want to run a linear or logistic regression on big data that cannot fit into memory, the coefficients can be optimized using gradient descent.

Distance algorithms like K-nearest neighbor, K-means, and support vector machines are most affected by the range of features. This is because behind the scenes, they are using distances between data points to determine their similarity. Therefore, we scale our data before emplying a distance-based algorithm so that all the features contribute equally to the result. Standardizing or normalizing data is also important in principal component analysis (PCA) because it projects your original data onto orthogonal directions, which maximize the variance.

However, tree-based analyses are not sensitive to outliers and do not require variable transformations. As a result, standardization of multi-scaled data is not necessary for decision tree, random forest, or gradient boosting algorithms.

Think About It: Feature Scaling Method

Suppose you are tasked with creating a neural network model on the **PVA_DONORS** data using the three variables shown below.



Std	16.515	10.057	8.993
Mean	59.151	14.876	10.508
Max	87	260	91
Min	0	0	1

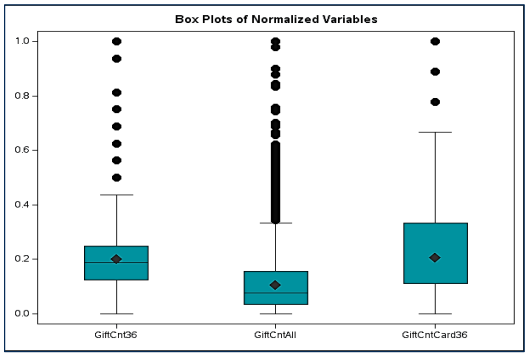
There is no obvious answer to this question. It really depends on the data and the intended model. In neural networks algorithms that require data on a 0 to 1 scale, normalization is an essential pre-processing step. Normalization is good to use when the distribution of data do not follow a Gaussian distribution. It can be useful in algorithms that do not assume any distribution of the data like *k*-nearest neighbors.

On the other hand, standardization can be helpful in cases where the data follow a Gaussian distribution. Though this does not have to be necessarily true. Standardization does not have a bounding range, so, even if there are outliers in the data, they will not be affected by standardization.

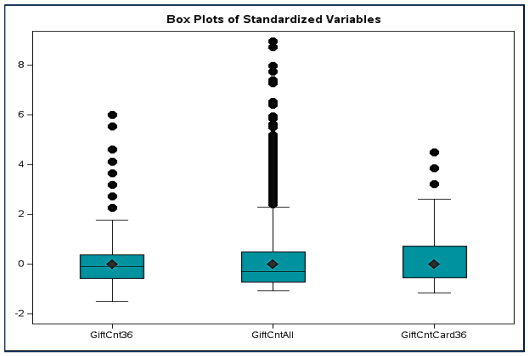
Think About It: Feature Scaling Method

Which feature scaling method would you use to correct disproportionate scales, if any?

Normalization



Standardization

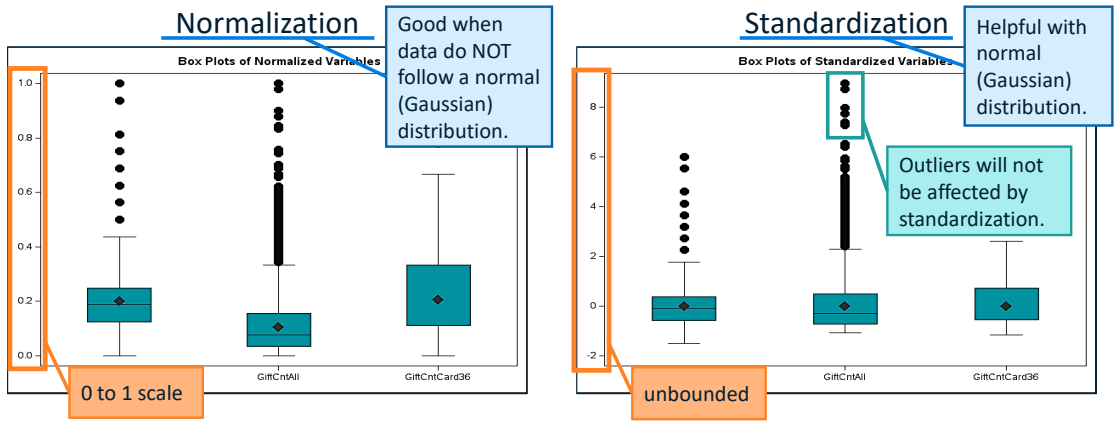


There is no obvious answer to this question. It really depends on the data and the intended model. In neural networks algorithms that require data on a 0 to 1 scale, normalization is an essential pre-processing step. Normalization is good to use when the distribution of data do not follow a Gaussian distribution. It can be useful in algorithms that do not assume any distribution of the data like *k*-nearest neighbors.

On the other hand, standardization can be helpful in cases where the data follow a Gaussian distribution. Though this does not have to be necessarily true. Standardization does not have a bounding range, so, even if there are outliers in the data, they will not be affected by standardization.

Answer: Feature Scaling Method

Which feature scaling method would you use to correct disproportionate scales, if any? ***It depends on the data!***



There is no obvious answer to this question. It really depends on the data and the intended model. In neural networks algorithms that require data on a 0 to 1 scale, normalization is an essential pre-processing step. Normalization is good to use when the distribution of data do not follow a Gaussian distribution. It can be useful in algorithms that do not assume any distribution of the data like *k*-nearest neighbors.

On the other hand, standardization can be helpful in cases where the data follow a Gaussian distribution. Though this does not have to be necessarily true. Standardization does not have a bounding range, so, even if there are outliers in the data, they will not be affected by standardization.

Question: Feature Scaling

Why do some of machine learning algorithms require you to scale a set of features in your data?

a. to make features more interpretable
b. to make your data values positive
c. because it's fun to do
d. because many algorithms are sensitive to unevenly scaled data

Answer: Feature Scaling

Why do some of machine learning algorithms require you to scale a set of features in your data?

a. to make features more interpretable
b. to make your data values positive
c. because it's fun to do
d. because many algorithms are sensitive to unevenly scaled data

Answer: d

Many machine learning algorithms are sensitive when the data are not scaled evenly. The features with high magnitudes will weigh in a lot more in the distance calculations than features with low magnitudes. To suppress this effect, we need to bring all features to the same level of magnitude. This can be achieved by *scaling*.

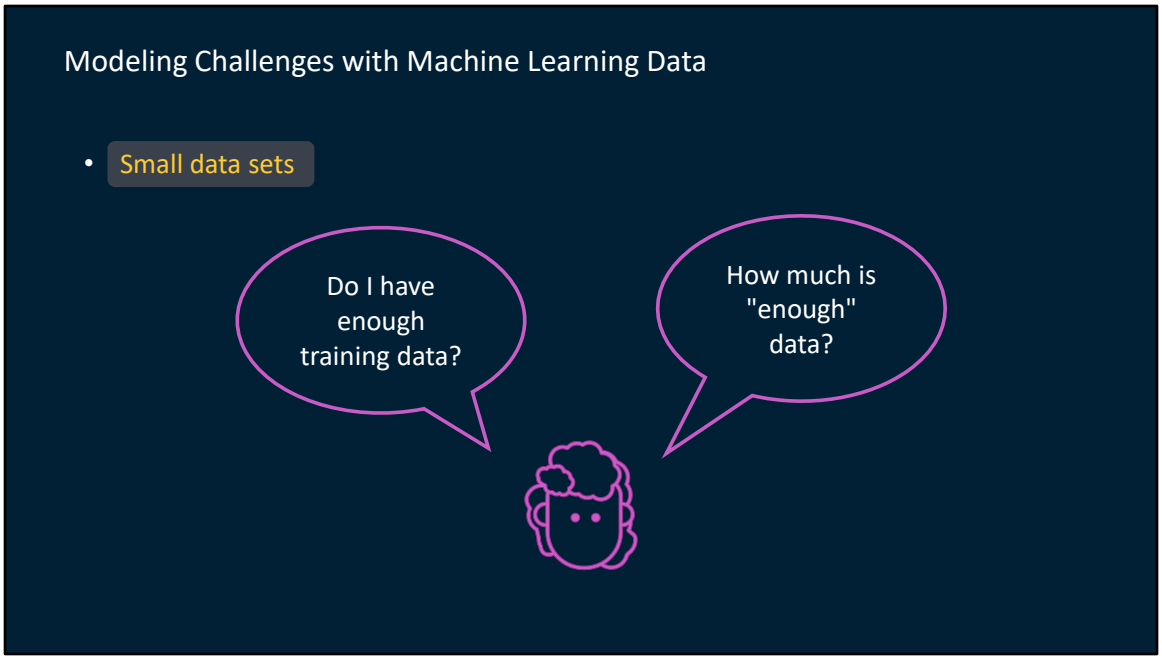
5.3 Model Evaluation, Estimation, and Post-training Tasks

Modeling Challenges with Machine Learning Data

Data Difficulties and Modeling Issues

- Visualizing big data
- Processing noisy data
- Handling high dimensional data
- Dealing with distinct variability in inputs
- **Using available data judiciously**
- Adjusting model flexibility and complexity
- Understanding learning and tuning of models
- Making models explainable

Throughout this lesson, we've been discussing data difficulties and modeling issues. Let's discuss another challenge. What if you don't have a lot of relevant data? How can you use the available data judiciously?

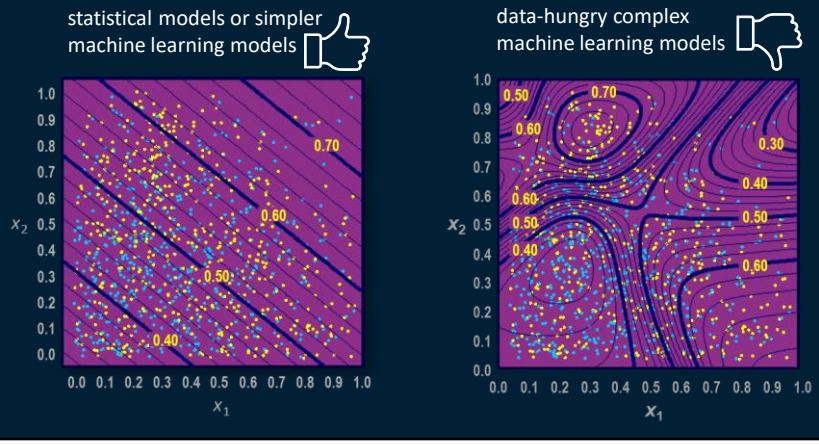


Despite having great inclination toward machine learning, you might not have enough data. Machine learning models are by and large data hungry, and their performance relies heavily on the size of training data available. In many cases, it is difficult to get training data sets that are large enough. For example, you might not have any data available for a newly launched product. Although it is difficult to evaluate the minimum amount of data required, the very nature of your project will influence the amount of data that you will need. For example, training models to identify texts, images, and videos usually require more data.

The real problem with insufficient data lies in the fact that with less data, variance increases. High variability in models means that the model will fit the training data perfectly but will stop working as soon as new data is fed into it.

Modeling Challenges with Machine Learning Data

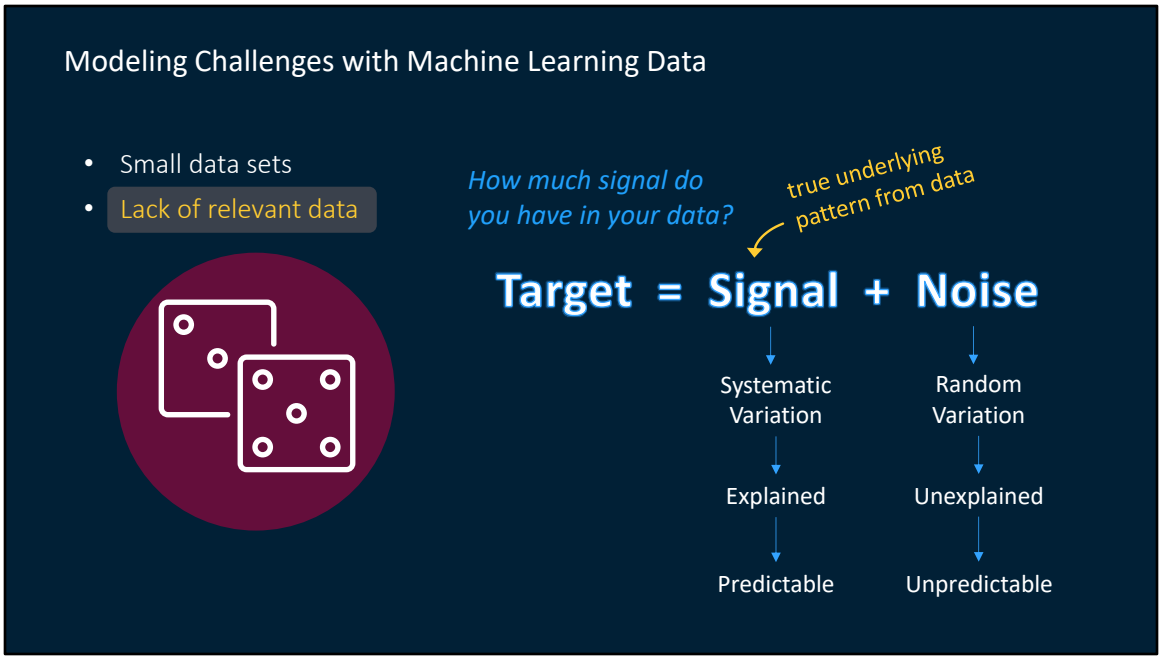
- Small data sets



Building traditional statistical models like regressions, or simpler machine learning models like short decision trees, is one of the most common approaches that can help with building predictive models from small data sets.

In general, the simpler the algorithm, the better it will learn from small data sets. From a machine learning perspective, small data requires models that have low complexity (or high bias) to avoid overfitting the model to the data.

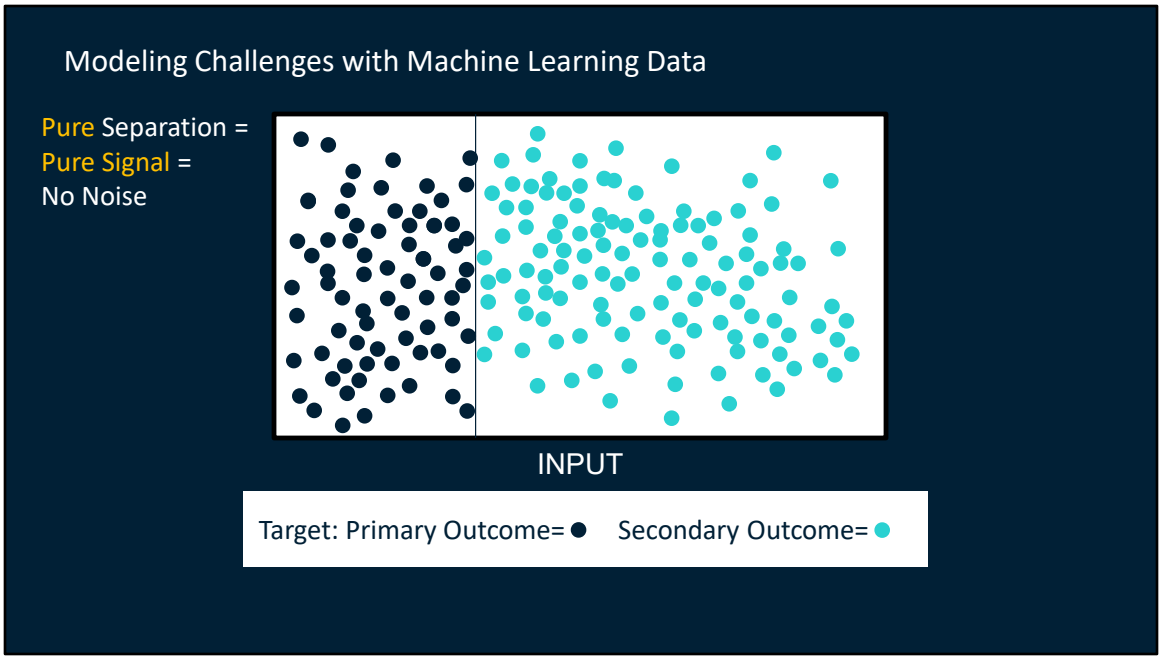
More than the size of the data, *relevance* of data is imperative.



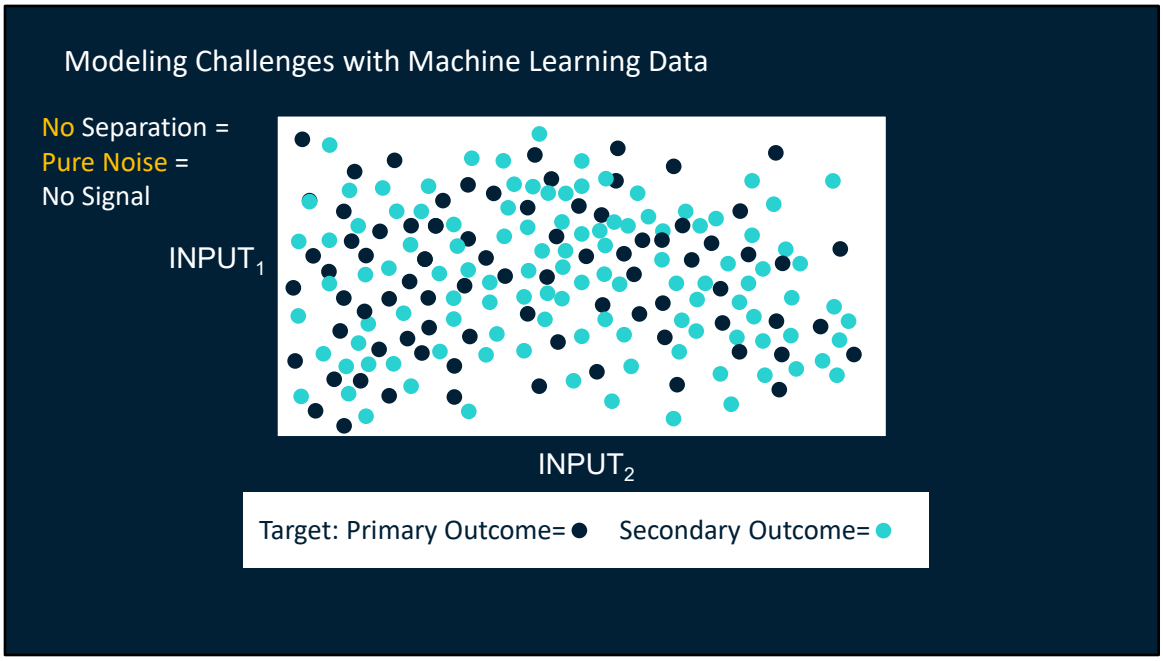
At times you have enough data, but that data might not be sufficiently relevant. Novice modelers who get caught up in the world of exotic algorithms often fall into the trap of thinking the computer can solve any problem given enough time and data. However, consider the problem of predicting the next roll of the dice in a Las Vegas casino. You could have one million data points of historic rolls of the dice, and you would have no better chance of predicting the next roll than if you had no data.

When you build a model, you assume that your data has two parts, signal and noise. You can think of the "signal" as the true underlying pattern that you want to learn from the data. Signal is the repeatable process that you hope to capture and describe. It is the information that you care about. The signal is what lets the model generalize to new situations. "Noise," on the other hand, refers to the irrelevant information or randomness in a data set. The random errors are referred to as noise. Noise is therefore inevitable and unpredictable. They are equivalent to the unexplained part of your statistical model.

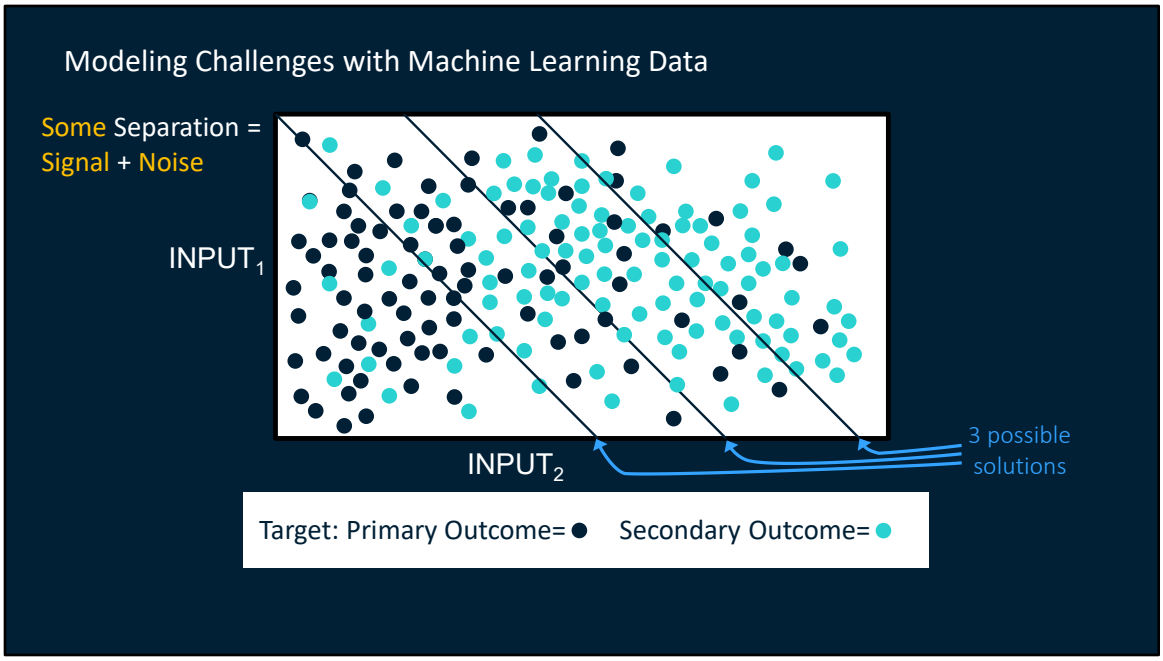
A signal-to-noise ratio is a measure of the amount of background noise with respect to the primary input signal.



This graphic illustrates the pure signal situation. In this case, the training data can be perfectly separated into primary or secondary outcomes using a linear decision boundary. You rarely expect to see this in the real world. Unfortunately, some people who are new to predictive modeling are disappointed when the methods do not perfectly categorize with this type of accuracy.



At the other extreme is the pure noise situation. In this case, the training data appears to have no patterns upon which to base a model that can separate the primary outcomes from the secondary outcomes. This situation is more common than you might like. Although pure signal is very rare, *pure noise* can actually occur in reality.

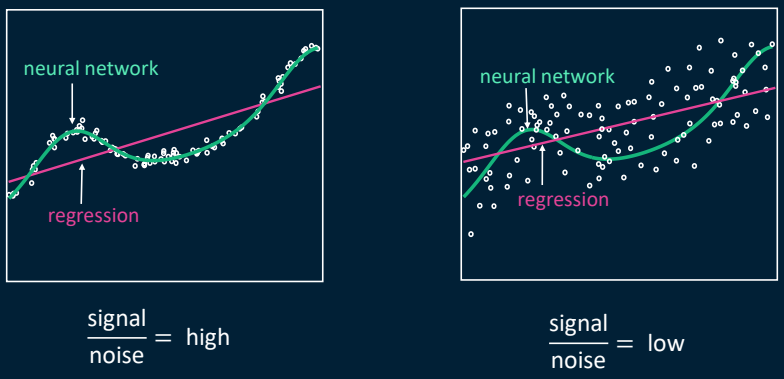


The most common situation in practice is a mixture of signal and noise. You can predict more accurately than randomly guessing. How *well* you predict depends on whether data is dominated by *systematic* variation or *random* variation. The three classifier lines in the plot are meant to represent three different possible solutions.

Modeling Challenges with Machine Learning Data

Machine Learning Models Are Not Always Superior!

Noisy data impact models.



$\frac{\text{signal}}{\text{noise}} = \text{high}$

$\frac{\text{signal}}{\text{noise}} = \text{low}$

Sometimes machine learning models like neural networks do not outperform simpler and easier to implement models like regression. The reason for this, in some situations, is found in the signal-to-noise ratio. Noisy data impacts models, including machine learning models.

To illustrate, in the left panel of the above diagram, the nonlinear model that is produced by the neural network would clearly produce a superior fit to the data, as compared to the fit produced by the linear regression model. There is a strong pattern (signal) relative to the amount of variation (noise) around the pattern. The signal-to-noise ratio is high.

The situation in the right panel is different. The signal-to-noise ratio is low. The regression and neural network models will likely offer a comparable fit to the data. So, there would be no advantage to using the more complex neural network.

Question: Signal and Noise

The true underlying pattern that you want to learn from the data is called "noise" whereas the irrelevant information or randomness in a data set is called "signal".

a. True
b. False

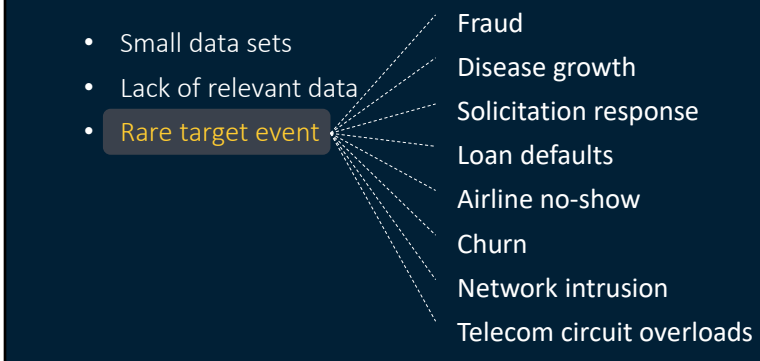
Answer: Signal and Noise

The true underlying pattern that you want to learn from the data is called "noise" whereas the irrelevant information or randomness in a data set is called "signal".

a. True
b. False

Answer: False
The true underlying pattern that you want to learn from the data is called "signal" whereas the irrelevant information or randomness in a data set is called "noise".

More about Modeling Challenges with Machine Learning Data



In supervised learning, the event of interest is often rare. Thus, in case of a binary target, the event class is in the minority. For example, a utility company might have terabytes of customer usage data and a desire to detect fraud, but it does not know which cases are fraudulent. The data is abundant, but none of it is supervised.

Another example would be health care data where the outcome of interest is progress of some condition across time, but only a tiny fraction of the patients were evaluated at more than one time point.

In direct marketing, if customer history and demographics are available but there is no information about response to a particular solicitation of interest, a test mailing is often used to obtain supervised data.

Loan defaults, airline no-show, churn, computer security threats, electricicty, and telecom circuit overloads are some other examples of rare target events.

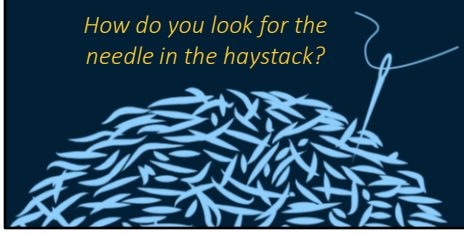
More about Modeling Challenges with Machine Learning Data

- Small data sets
- Lack of relevant data
- Rare target event

Problem

- Target search is frequently difficult
- Learns more from the majority group and can easily misclassify the minority event group
- Need a much bigger sample for building a reliable model (Harrell 1997)

How do you look for the needle in the haystack?



The event of interest or the "awkward" level of the target variable is rare, and target search is therefore frequently difficult. It's like looking for a needle in a haystack.

The machine learning algorithm learns more from the majority group and can easily misclassify the small event group.

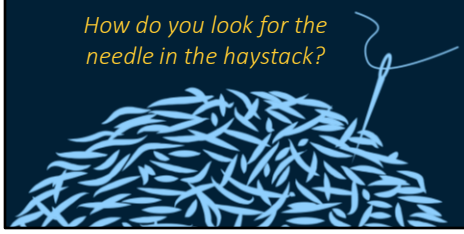
With rare events, the effective sample size for building a reliable prediction model is closer to three times the number of event cases than to the nominal size of the data set.

In such a scenario, do we have enough cases for modeling?

More about Modeling Challenges with Machine Learning Data

- Small data sets
- Lack of relevant data
- Rare target event

How do you look for the needle in the haystack?



Problem

- Target search is frequently difficult
- Learns more from the majority group and can easily misclassify the minority event group
- Need a much bigger sample for building a reliable model (Harrell 1997)

Solution

- Event-based sampling
  - undersampling
  - oversampling
- Use of artificial cases
  - SMOTE
- Unsupervised learning
- Rule induction

There are some different approaches for handling data sets with rare events. Undersampling and oversampling are two forms of event-based sampling that you are already aware of. Event-based sampling was discussed earlier in the course.

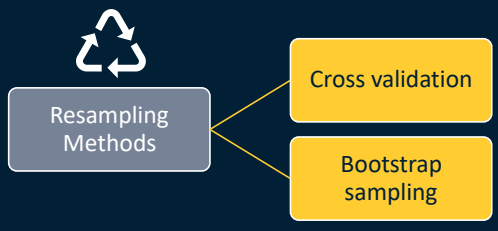
In order to have enough events, creating artificial cases is another approach. Synthetic minority over-sampling technique (SMOTE) is a technique that generate artificial cases. Simply put, SMOTE takes the rare event data points and creates new data points that lie between any two nearest data points joined by a straight line. The algorithm calculates the distance between two data points in the feature space, multiplies the distance by a random number between 0 and 1, and places the new data point at this new distance from one of the data points used for distance calculation.

When the data do not have the capacity to solve the problem, the problem needs to be reformulated. For example, there are unsupervised approaches to detecting anomalous data that might be useful for investigating possible fraud.

There are some specialized machine learning approaches available to improve the classification of rare events, for example, rule induction. Rule induction creates "if–else–then"-type prediction rules from a set of input variables and the target variable. The idea here is to initially generate prediction rules for non-events and then slowly move toward more complex rules that can predict rare events.

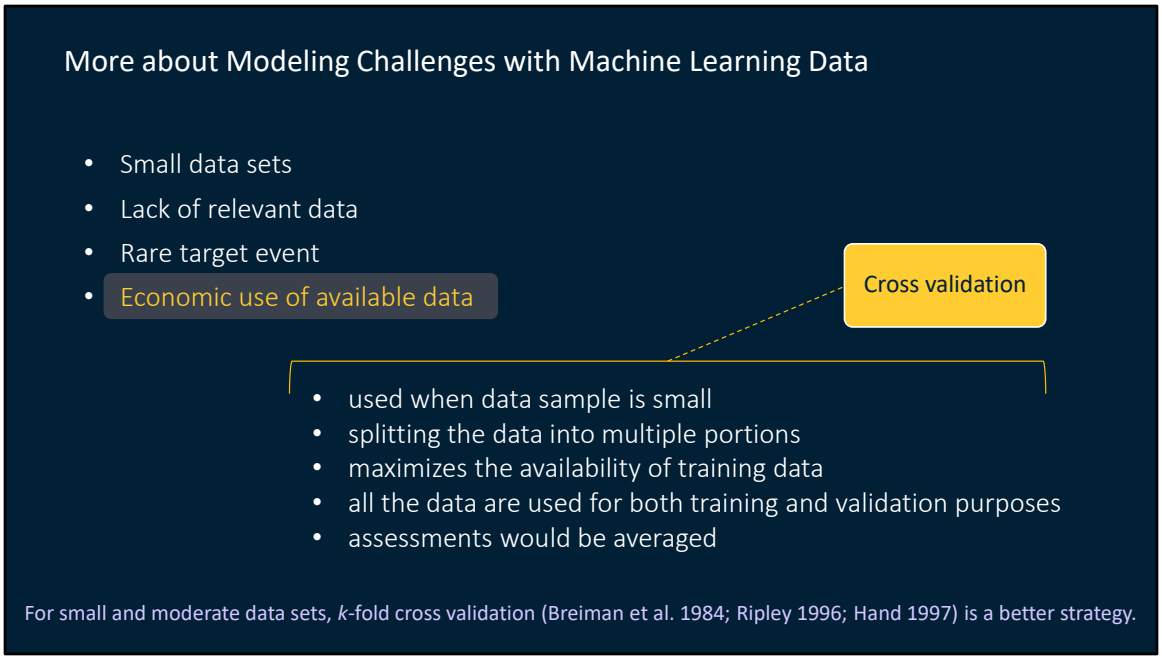
More about Modeling Challenges with Machine Learning Data

- Small data sets
- Lack of relevant data
- Rare target event
- Economic use of available data



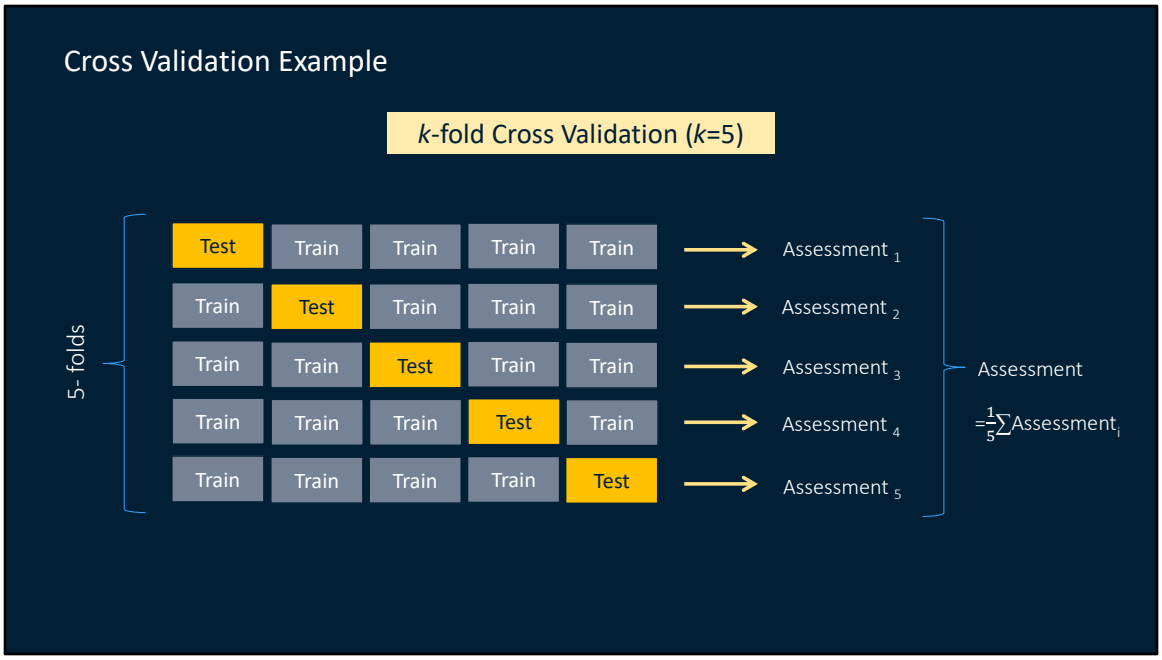
With moderate sized data sets, data splitting is inefficient. The reduced sample size can severely degrade the fit of the model. Computer-intensive resampling methods have been developed so that all the data can be used for both fitting and honest assessment. Resampling refers to methods for economically using the available data set for robustly assessing model performance and thereby improving model accuracy. Essentially a subset of samples are used to fit a model and the remaining samples are used to estimate the efficacy of the model. This process is repeated multiple times and the results are aggregated and summarized. The differences in techniques usually center around the method in which subsamples are chosen. Although the method sounds overwhelming, the math involved is relatively simple.

Resampling is a set of methods to either estimate the precision of a statistic using *cross validation* or repeat sampling from a given sample known as *bootstrapping*. Thus, the two commonly used resampling methods that you might encounter are cross validation and the bootstrap.

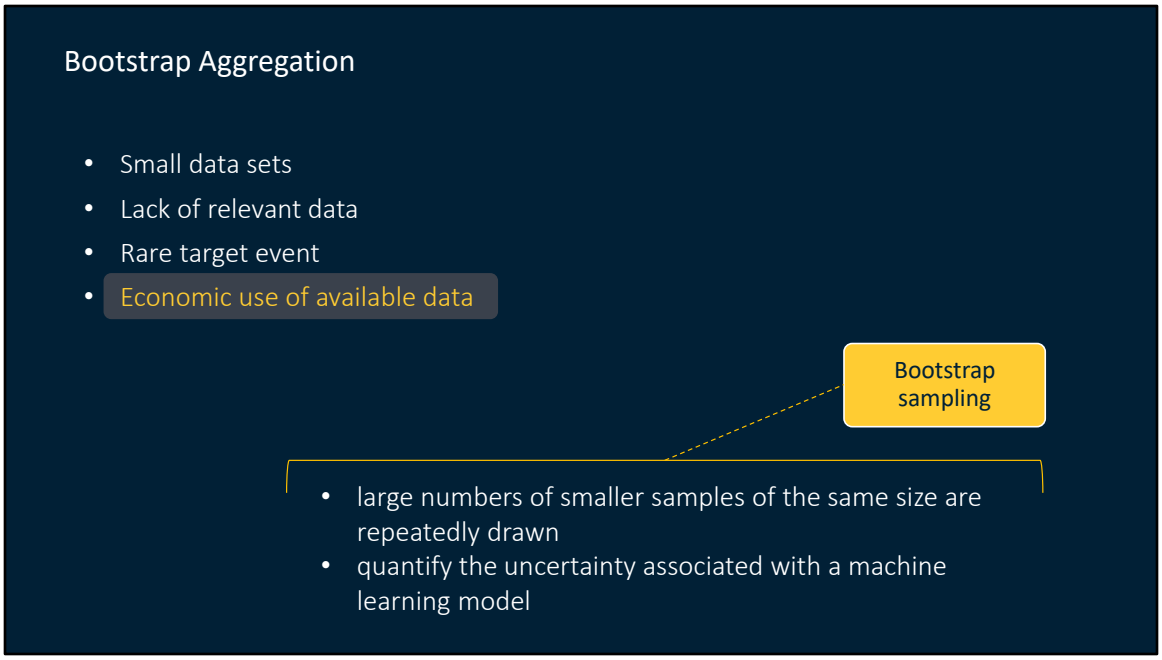


Partitioning the data into two segments (one for training and one for validation) is a simple but costly technique because a huge portion of data is held out for testing. When the test set is small, the performance measures might be unreliable because of high variability. Fortunately, there is an effective workaround for this issue. For small and moderate data sets, *k*-fold cross validation is a better strategy.

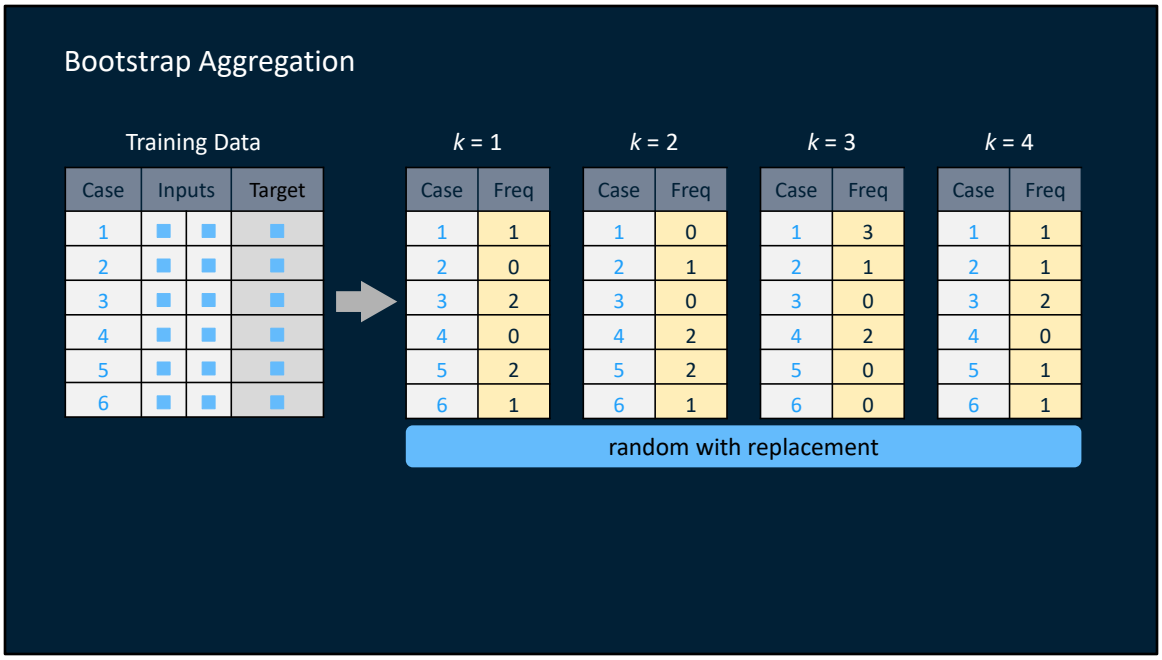
k-fold cross validation is a resampling procedure used to evaluate machine learning models on a limited data sample. Cross validation maximizes the availability of training data by splitting the data into multiple subsets, to use different subsets of data to train and test the model on. The *k*-fold cross validation involves splitting the data into *k* subsets. You take one subset from the bunch and treat it as the test set for the model. And you keep the *k*-1 subset for training the model. You repeat the procedure *k* times, rotating the validation set. *k* is selected based on the size of the data set. This ensures that our model is getting trained and tested on new data at every new step.



Here is an example of five-fold cross validation. The data would be split into five equal sets. The entire modeling process would be redone on each four-fifths of the data, using the remaining one-fifth for assessment. The five assessments would then be averaged. In this way, all the data are used for both training and validation purposes, which dramatically minimizes the potential error (such as overfitting) found by relying on a fixed split of training and validation data.

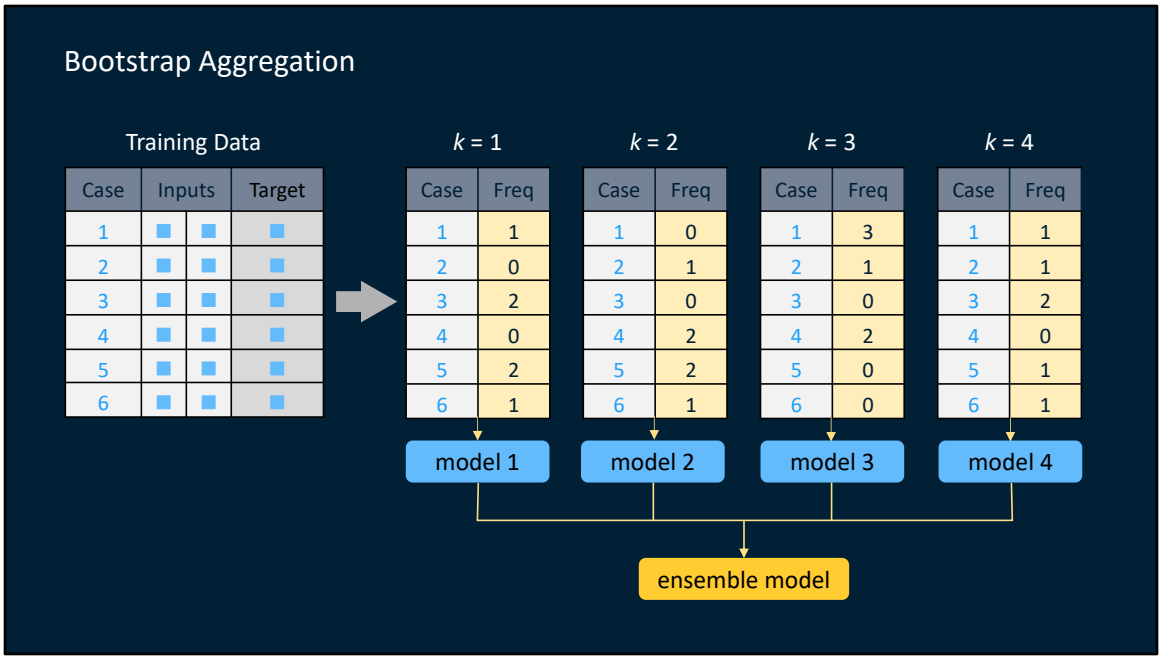


Bootstrapping is a type of resampling where large numbers of smaller samples of the same size are repeatedly drawn, with replacement, from a single original sample. It is often used as a means of quantifying the uncertainty associated with a machine learning model.



Let's look at a simple example. From a training data set that has six cases, the first step is to draw four bootstrap samples. Each sample has six cases.

A *bootstrap sample* is a random sample from a training data set. For bagging, a bootstrap sample is drawn with replacement. This means that some of the cases might be left out of the sample, and some cases might be represented more than once. For example, in the first bootstrap sample, cases 3 and 5 appear twice, cases 1 and 6 appear once, and cases 2 and 4 do not appear at all.



After a certain number of equally sized bootstrap samples of a data set are extracted, a machine learning algorithm is applied to each of these samples to get several models. Their predictions are then combined to create an ensemble model. This is commonly known as *bootstrap aggregation* or *bagging*. The idea is to average many noisy but approximately unbiased models to reduce the variance of estimated prediction function.

The ensemble approach relies on combining many relatively weak simple models to obtain a stronger ensemble prediction. This will improve accuracy, reduce bias, reduce variance, and provide robust models in the presence of new data. The commonly observed advantage of the ensemble model is that the combined model is better than the individual models that compose it. It is important to note that the ensemble model can be more accurate than the individual models only if the individual models disagree with one another.

Question: Cross Validation

Cross validation is another method of honest assessment that can be used when data size is small and you do not have a luxury of holding out data separately for validation or testing, or both.

a. True
b. False

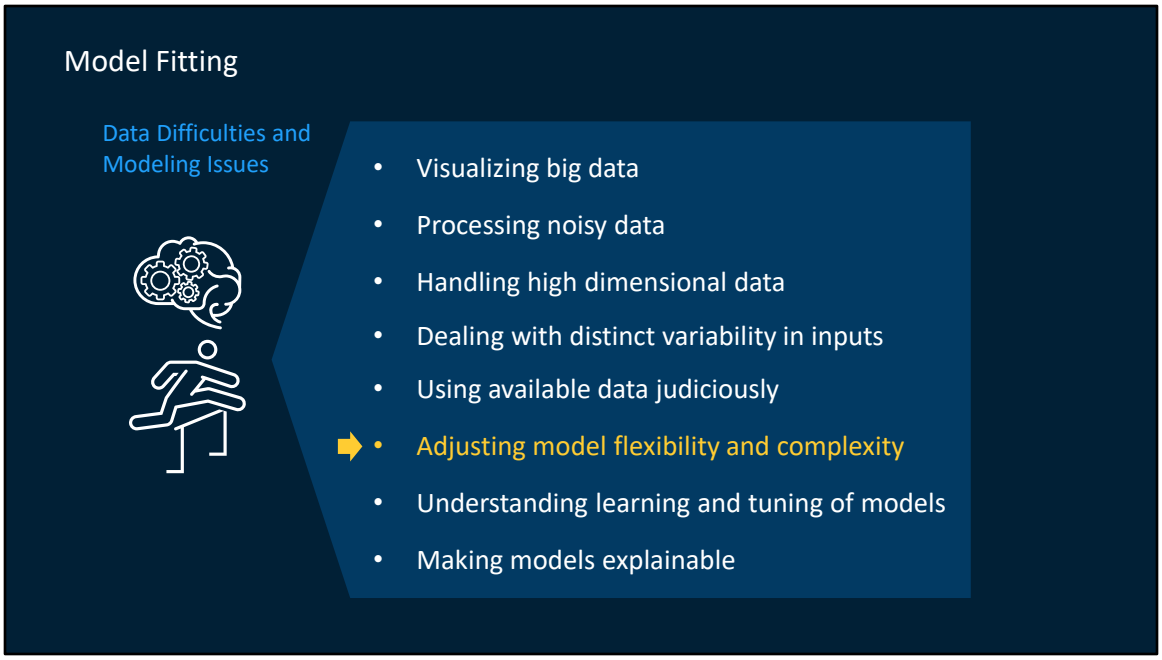
Answer: Cross Validation

Cross validation is another method of honest assessment that can be used when data size is small and you do not have a luxury of holding out data separately for validation or testing, or both.

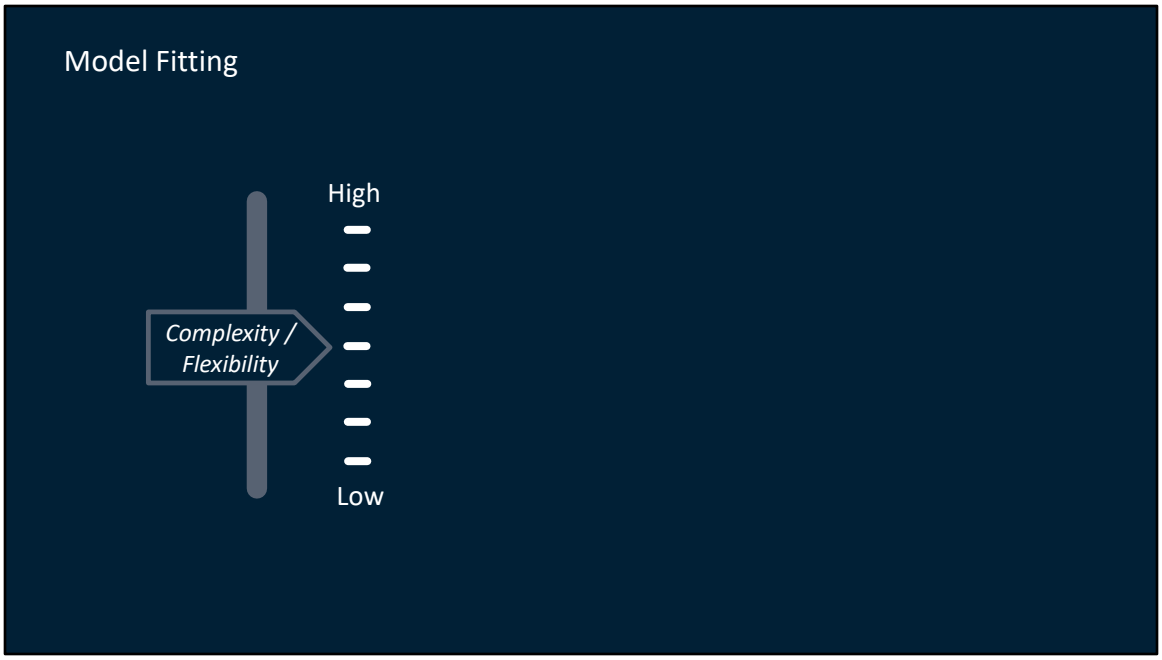
a. True
b. False

Answer: True

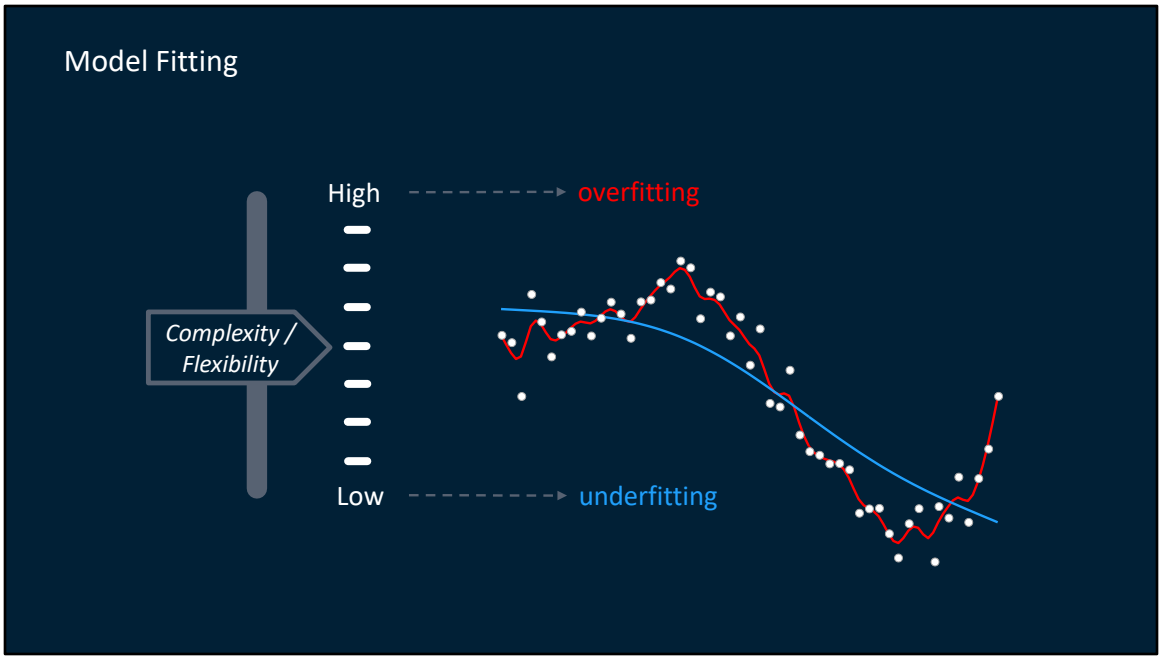
Data partitioning is a simple but costly technique because a huge portion of data is held out for validation and testing. Alternatively, k-fold cross validation, which is a resampling procedure, can be used for honest assessment to evaluate machine learning models on a limited data sample.



Next, let's talk about challenges with model fitting - how to adjust the *flexibility* and *complexity* of your model.



In statistics, model fit refers to how well you approximate a target function. This is applicable in machine learning too because supervised machine learning algorithms seek to approximate the unknown mapping function for the target given the input variables. Thus, model fitting refers to how well a machine learning model generalizes to similar data to that on which it was trained. A model that is well-fitted produces more accurate outcomes. Model fitting is closely related to the model complexity, or flexibility, and the related bias-variance trade-off.

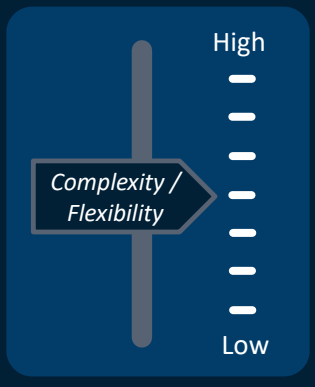


Whether we overfit or underfit can depend on the complexity or flexibility of your model. When the model is too complex, it can start to learn the "noise," or irrelevant information. When the model memorizes the noise and fits too closely to the training set, the model becomes "overfitted," (as shown by the red curve) and it is unable to generalize well to new data. If a model cannot generalize well to new data, then it will not be able to perform the classification or prediction tasks that it was intended for. A model that is overfitted matches the training data too closely. A model that is underfitted doesn't match closely enough. When the model is too simple, or less flexible, the model becomes "underfitted" (as shown in by blue curve).

Here it is important to note that not all machine learning algorithms are overfitted by having only too many inputs. This is especially a problem for models where additional input variables directly mean that additional parameters must be estimated. Examples of such models are regression and neural network models.

Model Fitting

Complexity and Flexibility in Different Models



Model	What Affects Complexity of Each Model
Regressions	# Parameters / Effects
Decision Trees	# Leaves
Gradient Boosting	# Trees
SVMs	# Support vectors
Neural Networks	# Iterations

What precisely constitutes model complexity or flexibility is an intricate matter. Many factors govern whether a model will generalize well. For example, a regression model with more parameters or effects might be considered more complex. A decision tree with more leaves and a gradient boosting model with more trees might be more complex. Similarly, a greater number of support vectors in support vector machines indicates high flexibility. Often with neural networks, we think of a model that takes more training iterations as more complex, and one subject to *early stopping* (or fewer training iterations) as less complex.

Thus, it can be difficult to compare the complexity among members of substantially different model classes. For now, a simple rule of thumb is quite useful: a model that can readily explain arbitrary facts (or noise) is what statisticians view as complex, whereas one that has only a limited expressive power (or signal), and yet still manages to explain the data well, is probably more generalizable and closer to the truth.

Question: Model Complexity

Which of the following statements are True? Select all that apply.

a. A more complex model generally leads to overfitting.
b. A more complex model generally leads to underfitting.
c. Model complexity has nothing to do with overfitting or underfitting.
d. The less complex the model is, the less flexible it is.

Answer: Model Complexity

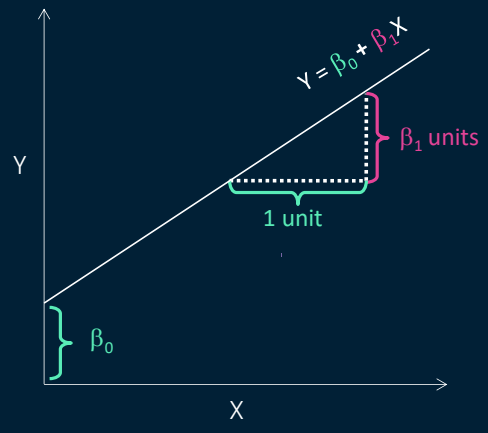
Which of the following statements are True? Select all that apply.

a. A more complex model generally leads to overfitting.
b. A more complex model generally leads to underfitting.
c. Model complexity has nothing to do with overfitting or underfitting.
d. The less complex the model is, the less flexible it is.

Correct answer: a and d

Model fitting is closely related to the model complexity, or flexibility, and the related bias-variance trade-off. When the model is too complex, it can start to learn the "noise," or irrelevant information. Whether you overfit or underfit, it can depend on the complexity or flexibility of your model.

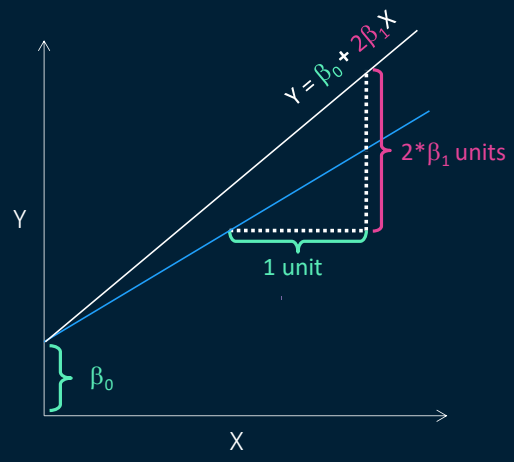
Effect of Magnitude of Coefficients



Many machine learning algorithms do not have a built-in capability of variable selection, and thus they build the models on all inputs supplied. While building a machine learning model, the model can easily be overfitted or underfitted.

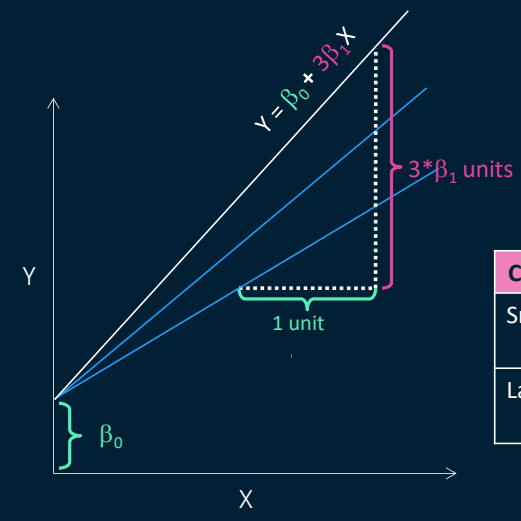
In parametric models, like regression and neural network, the magnitude of coefficients influence the model complexity. To keep things simple, let's consider the case of a simple linear regression. Y is the response variable and X is the predictor variable. In its simplest form, it attempts to fit a straight line to your data set. There are two coefficients: β_0 is the intercept parameter and β_1 is the slope parameter. The equation suggests that for every one unit increase in X, there is β_1 units increase in predicted Y.

Effect of Magnitude of Coefficients



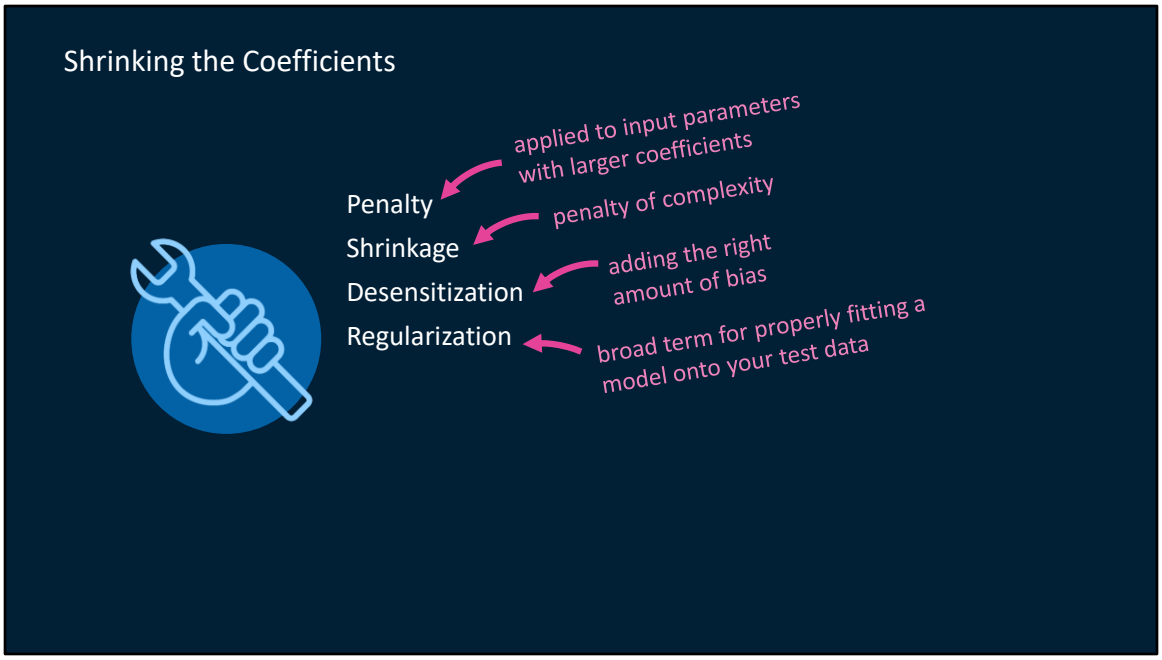
Let's see what happens when the coefficient becomes larger. For example, if the slope gets doubled, the regression line turns out to be steeper. Now the equation suggests that for every one unit increase in X, the prediction for Y increases by two times β_1 units.

Effect of Magnitude of Coefficients



Coefficients	Result
Small	predictions for Y are less sensitive to changes in X
Large	predictions for Y are more sensitive to changes in X

If the slope gets tripled, the equation suggests that for every one unit increase in X ,there is three times β_1 units increase in predicted Y. When the slope of the line is steep, the prediction for Y is very sensitive to relatively small changes in X. When the slope is small, for every one unit increase in X, the prediction for Y barely increases. In other words, when the slope of the line is small, predictions for Y are much less sensitive to changes in X.



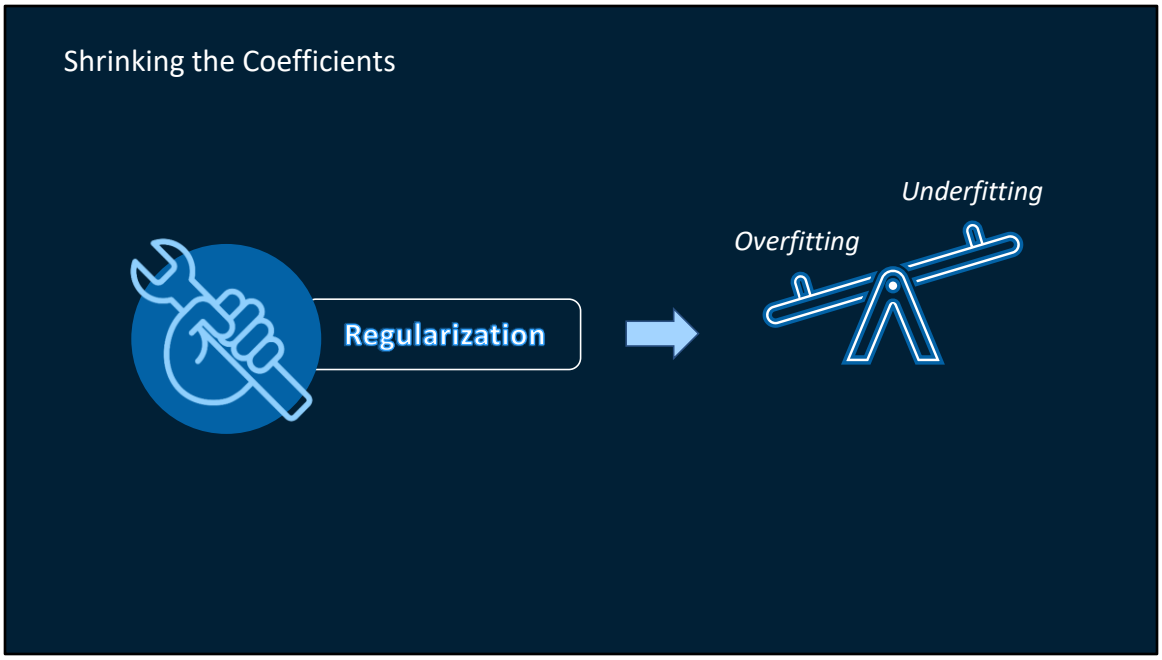
To avoid overfitting or underfitting, data scientists and statisticians often talk about penalty, shrinkage, desensitization, or regularization. All these terms refer to a broad range of techniques for artificially forcing your model to be simpler.

We can apply a *penalty* to the input parameters with the larger coefficients, which subsequently limits the amount of variance in the model.

Shrinkage can be thought of as a penalty of complexity. If we set some parameters of the model to exactly zero, then the model is effectively shrunk to have lower dimensionality and be less complex.

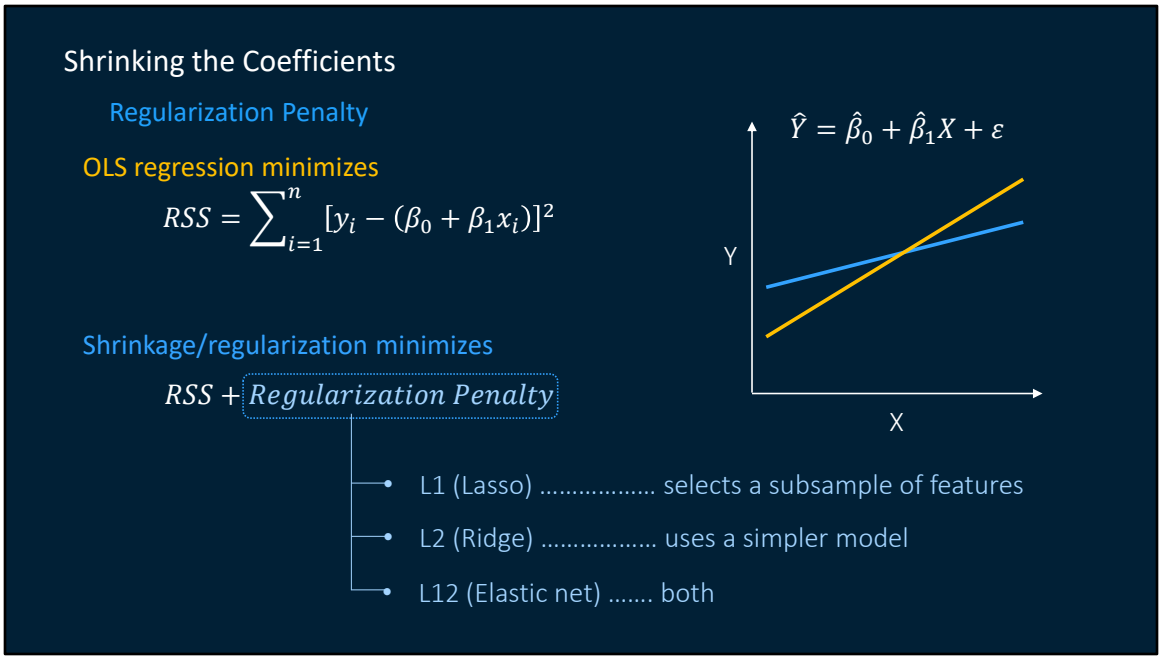
Adding the right amount of bias to a model can help make more accurate predictions by *desensitizing* it to some of the noise in the training data.

Regularization is a broader term commonly used in machine learning to properly fit a model onto your validation data. The regularization method will depend on the type of model you're using. For example, you could prune a decision tree, use dropout on a neural network, or add a penalty parameter to the cost function in regression. Many times, the regularization method is a hyperparameter as well, which means it can be tuned through cross validation.



Regularization essentially balances complexity in machine learning models. Regularization refers to techniques that are used to calibrate machine learning models to prevent overfitting or underfitting.

How does regularization penalize or shrink the coefficients?



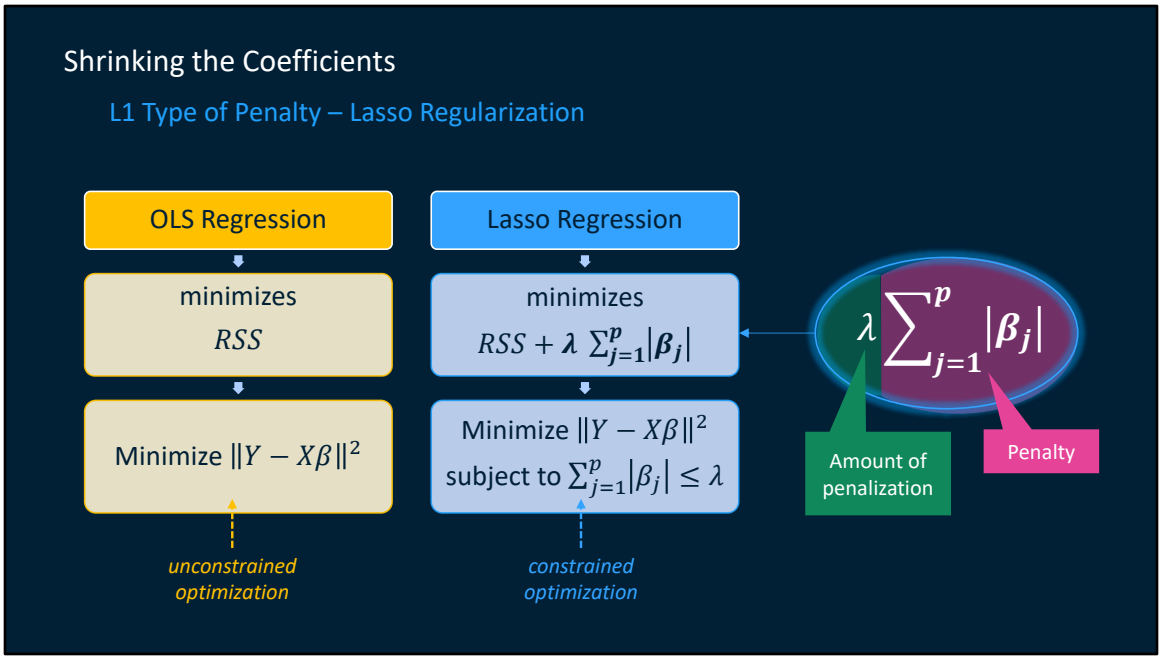
Recall a simple linear regression model that takes the form shown above the graph on the right. Here, Y-hat represents the predicted response, and β-hat represents the coefficient estimates. The values of these βs are estimated by using the ordinary least squares method that minimizes the sum of the squared residuals.

Regularization is a form of regression that constrains or regularizes the coefficient estimates to shrink them toward zero. In its most simplistic form, regularization adds a penalty to model parameters (all except intercepts), so the model generalizes the data well. Or contrary to OLS, shrinkage or regularization minimizes the penalized sum of squared residuals. Thus, the coefficients are estimated by minimizing a slightly different quantity when you regularize. A regularization penalty is added to the residual sum of squares.

The main idea behind shrinkage methods is to find a new line that doesn't fit the training data as well. So, we introduce a small amount of bias into how the new line is fit to the data. In return for that small amount of bias, we get a significant drop in variance. In other words, by starting with a slightly worse fit, regularization can provide stable predictions in the long run. In our example, we see that the regularization penalty resulted in a line (the blue line) that has a smaller slope, which means that predictions made using this line are less sensitive to the changes in the values of 'X' than the least squares line (the yellow line).

Three types of regularization penalties are often used in machine learning models:
L1 or *Lasso* regularization: L1 uses a shrinkage mechanism to zero out some parameters and selects a subsample of features.
L2 or *Ridge* regularization: L2 adds a penalty on the parameters and uses a simpler model.
L12 or *Elastic net* regularization: L12 is a tradeoff between L2 and L1 regularization and has a penalty that is a mix of L1 and L2 norm. It can perform the function of feature selection while still not discarding too many features.

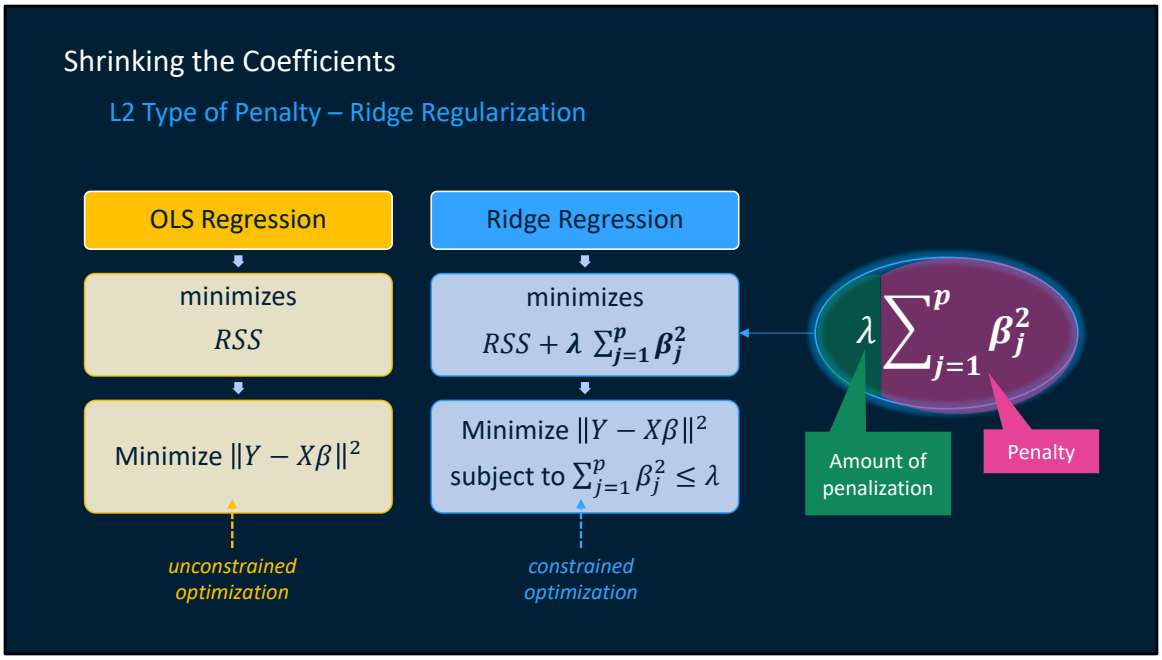
Lasso, ridge, and elastic net are all part of the linear regression family where the x (input) and y (output) are assumed to have a linear relationship.



OLS estimates the parameters of a linear function by the principle of least squares. Thus, the cost function is simply minimizing the sum of the squared residuals. The OLS coefficients βs are the solution to the unconstrained optimization problem.

Lasso is a modification of linear regression. Note that the cost function has an additional factor added to the sum of the squared residuals that is a penalty term of L1 norm emphasized in the ellipsis. The model is penalized for the sum of absolute values of the weights. Thus, the absolute values of weight will be (in general) reduced, and many will tend to be zeros. A tuning parameter lambda multiplies the penalty term, allowing control over the amount of penalization. Increasing the value of lambda increases the amount of shrinkage in the parameters.

This is equivalent to a constrained form of ordinary least squares where we are constraining the sum of the absolute values of the regression coefficients so they are smaller than lambda. If the parameter lambda is small, some of the regression coefficients are zero. Due to its strong tendency of setting some parameters to zeros, it is often used to select features when we know that some features are very sparse.



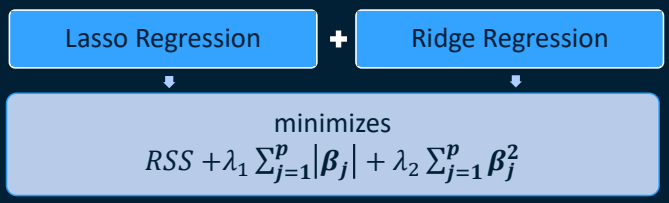
Ridge takes a step further and adds a penalty of L2 norm on the parameters βs emphasized in the ellipsis. The model is penalized for the sum of squared value of the weights. Thus, the weights not only tend to have smaller absolute values, but also tend to penalize the extremes of the weights, resulting in a group of weights that are more evenly distributed. Here also, a tuning parameter lambda controls the amount of penalization.

As with least squares, ridge regression seeks coefficient estimates that fit the data well, by making the RSS small. However, the shrinkage penalty is small when βs are close to zero, so it has the effect of shrinking the estimates of β_j toward zero but never equal to zero.

Ridge minimizes the penalized sum of squares, which is equivalent to minimization of OLS objective function subject to constraining the sum of the squared coefficients so they are smaller than lambda. Thus, the ridge coefficients βs are the solution to the constrained optimization problem.

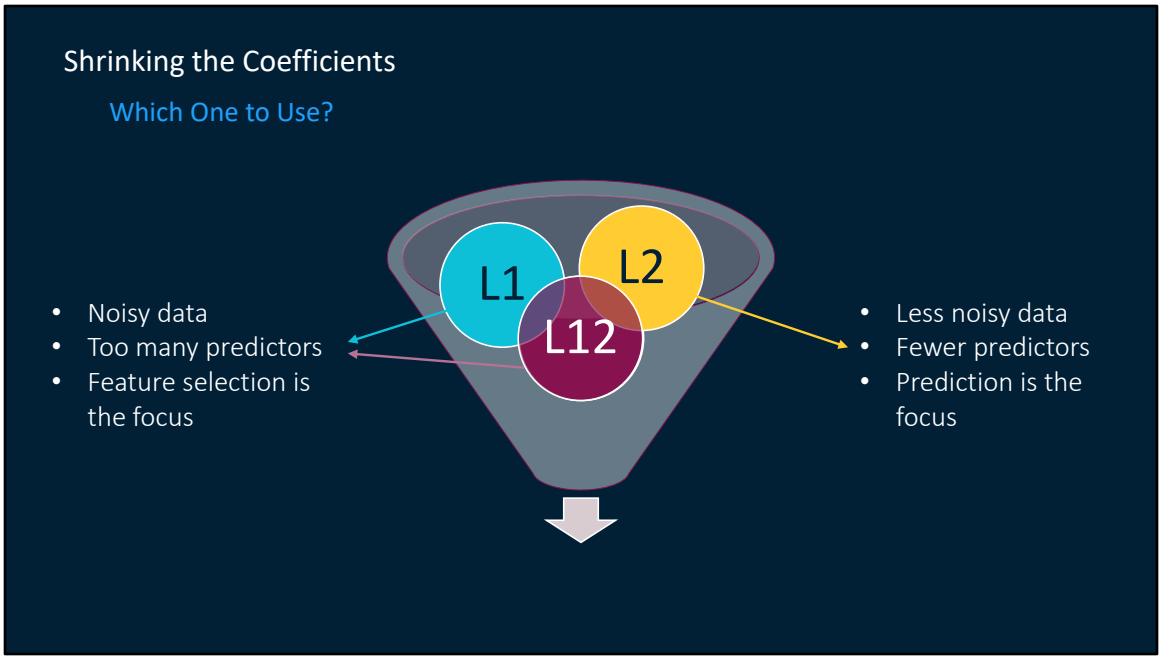
Shrinking the Coefficients

L12 Type of Penalty – Elastic Net Regularization



Selects features without discarding too many features

Elastic net is a hybrid of lasso and ridge, where both the absolute value penalization and squared penalization are included. Elastic net regularization is a tradeoff between L1 and L2 regularization and has a penalty that is a mix of L1 and L2 norm. It can perform the function of feature selection but still not imposing too much sparsity on the features (discarding too many features) by imposing a mixture of L1 and L2 regularization on parameters βs.



The selection of different penalties depends on the problem at hand.

With too many predictors, the signals are truly sparse - so L1 or L12 penalty can be used to select the hidden signals from noisy data, although it is almost impossible for L2 to fully recover the sparse signals.

L1 and L12 tend to give sparse weights (mostly zeros), because the L1 regularization cares equally about driving down big weights to small weights or driving small weights to zeros. If you have a lot of predictors (or features) and you suspect that not all of them are that important, it might be a really good idea to start with L1 and L12.

For problems with fewer predictors, the features are not sparse at all, so L2 regularization often outperforms L1 regularization.

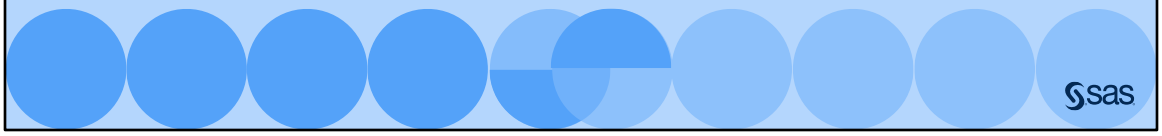
L2 tends to give small but well-distributed weights, because the L2 regularization cares more about driving big weight to small weights, instead of driving small weights to zeros. If you have only a few predictors, and you are confident that all of them are relevant for predictions, L2 is a preferred regularization method.

In the prediction applications, ridge regression with L2 is more common and often recommended. However, in the case where you have many features and you want to reduce the complexity of the model by deselecting some features, you might want to impose L1 penalty or go for more of a balanced approach like elastic net.

You will need to scale your data before using these regularized linear regression methods.

L1 versus L2

To explore this learning resource in greater detail, please visit the course page on learn.sas.com. There, you'll find comprehensive course materials, tutorials, and updates to support your learning journey.



Question: Regularization Methods

Which of the following statements is True regarding regularization methods?

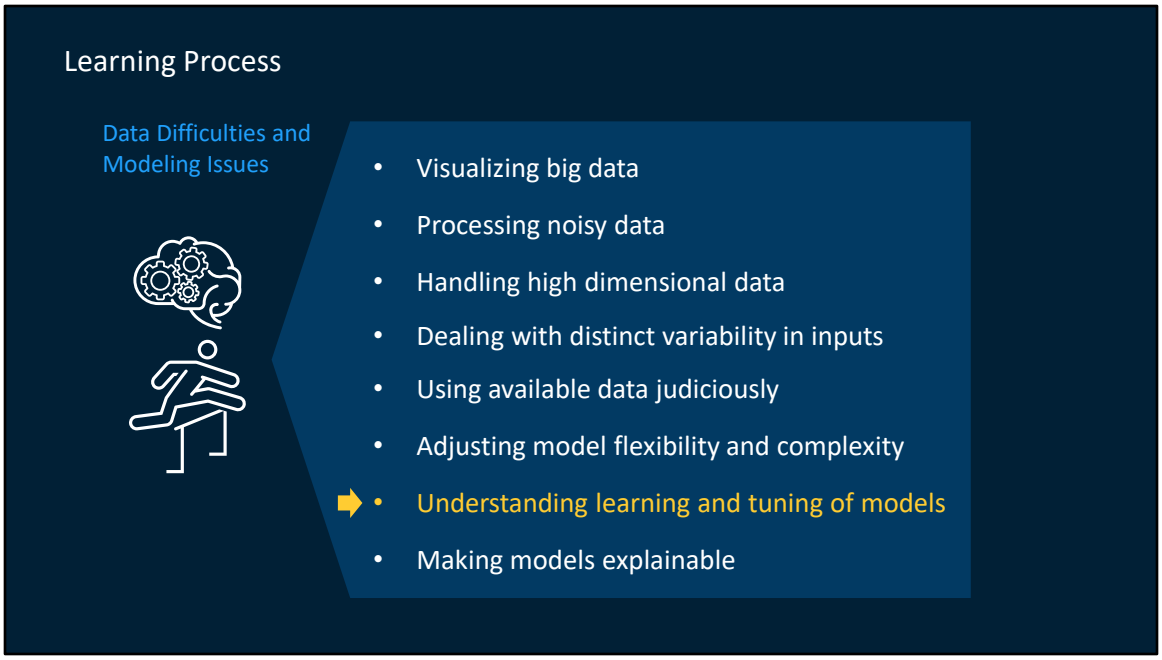
a. The sum of the absolute values of the weights is imposed as a penalty in Lasso (L1).
b. The sum of the squared values of weights is imposed as a penalty in Ridge (L2).
c. Elastic Net (L12) is a hybrid of Lasso (L1) and Ridge (L2).
d. All the above.

Answer: Regularization Methods

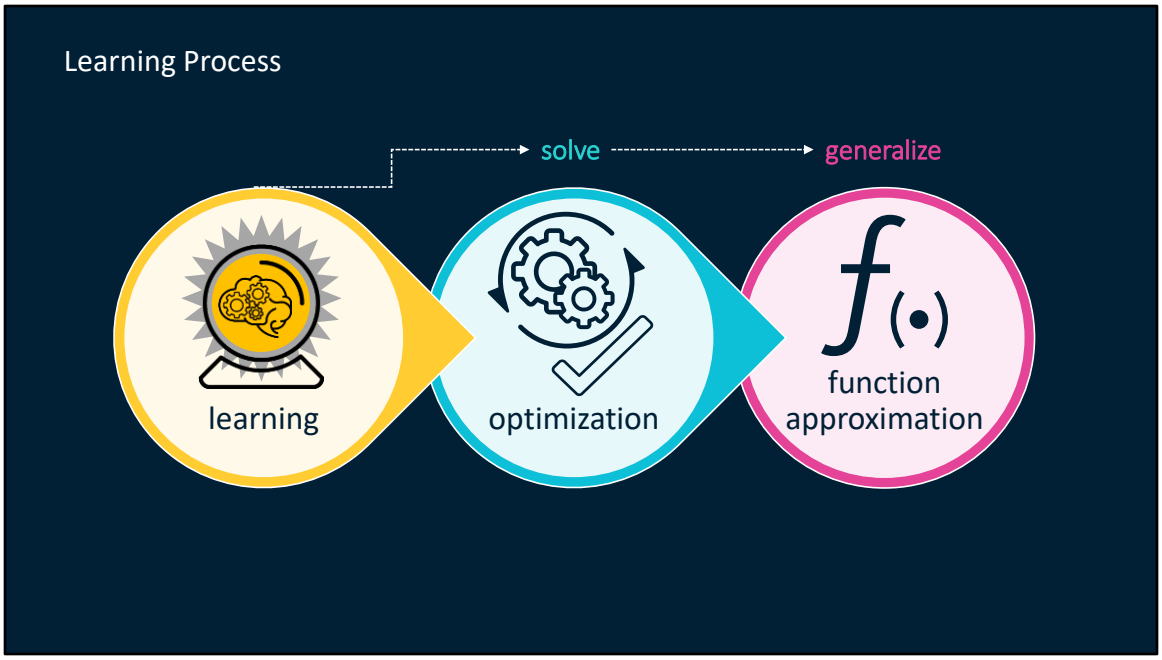
Which of the following statements is True regarding regularization methods?

a. The sum of the absolute values of the weights is imposed as a penalty in Lasso (L1).
b. The sum of the squared values of weights is imposed as a penalty in Ridge (L2).
c. Elastic Net (L12) is a hybrid of Lasso (L1) and Ridge (L2).
d. All the above.

Answer: d



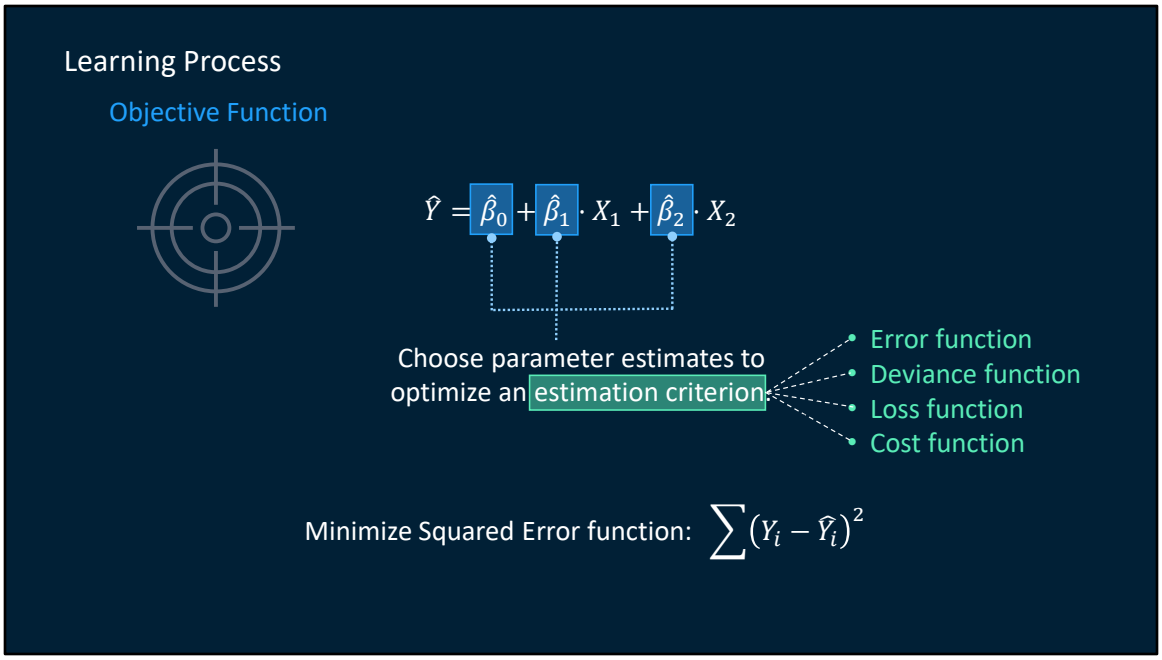
A machine learning algorithm uses a numerical optimization method that you specify to minimize the error function. It is therefore important to understand the learning process and tuning of machine learning models.



Machine learning involves using an algorithm to solve a mathematical problem and generalize from historical data in order to make predictions on new data. This is typically called learning a model.

Almost all machine learning algorithms can be formulated as an optimization problem. 'Optimization' comes from the same root as 'optimal', which means best. When you optimize something, you are 'making it best' given certain constraints. Learning refers to an optimization process where the parameters of the model update themselves using a mathematical function at every iteration of the training. This mathematical function is generally called an objective function and then the optimization problem is to find the extremum of this objective function.

Mathematically, solving a function optimization problem is essentially a problem of approximating the mapping function to map inputs to the target. A machine learning algorithm defines a parameterized mapping function (for example, a weighted sum of inputs) and an optimization algorithm is used to find the values of the parameters (for example, model coefficients) that minimize the error of the function. Function approximation generalizes from specific examples to a reusable mapping function that maps inputs to the target for making predictions on new examples.

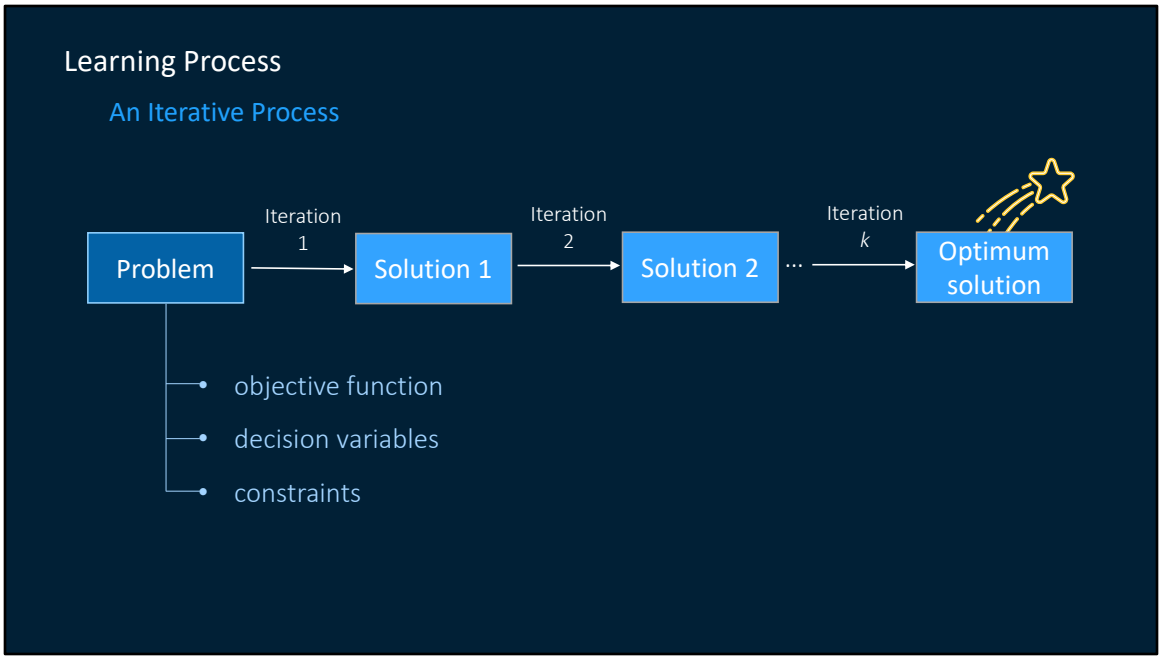


Building models and constructing reasonable objective functions are the first step in machine learning methods. With the determined objective function, appropriate numerical optimization methods are usually used to solve the optimization problem.

Recall that the primary task in a regression model is to obtain parameter estimates – or, in the neural network case, weight estimates – that result in accurate predictions. For example, in a linear regression model, *intercept* and *parameter estimates* are chosen to minimize the squared error between the predicted and observed target values.

Thus, in order to find the optimal solution, we need some way of measuring the quality of any solution. These estimates are best chosen by optimizing an estimation criterion which is known as an objective function. This function, taking data and model parameters as arguments, can be evaluated to return a number. Any given problem contains some parameters that can be changed. Our goal is to find values for these parameters which either maximize or minimize this number.

When you are minimizing the estimation criterion, you can also call it the *error function*, *deviance function*, *loss function*, or *cost function* - these terms are almost synonymous. They all converge to the same idea of minimizing the difference between what is observed in the data set and what is predicted by your model. The cost function is used more in optimization problems, and loss function is used in parameter estimation.



Optimization begins by a modeling phase, in which a given problem must be stated in terms of objective function, decision variables, and constraints. An objective function is a value that you are trying to optimize. Decision variables are inputs to your problem that your optimizer is allowed to change to try to improve the objective function value. Constraints are conditions that the decision variables must satisfy for the decision to be acceptable. This allows us to characterize the problem, and most importantly its solutions.

In general, it is not possible to directly compute a solution of an optimization problem from its formulation. Instead, almost all numerical optimization algorithms proceed in an iterative fashion. Given a current point that represents the current approximation to the solution, an optimization procedure attempts to move toward a (potentially) better solution to finally reach to an optimum solution.

So, an algorithm must have a method that will try to compute an approximate solution of the problem. By analyzing this method, it is often possible to identify how fast a method can be at getting close to a solution.

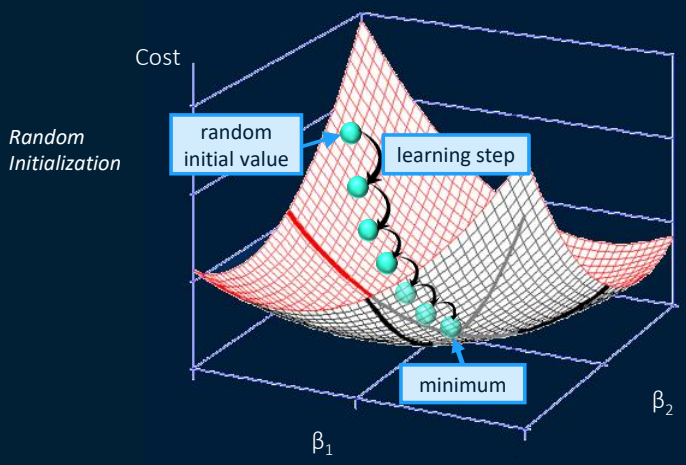
Learning Process

Gradient Descent

Gradient:
steepness of slope

Gradient Descent:
modifying parameters iteratively to minimize a cost function

- measures local gradient of error function
- goes in the direction of descending gradient

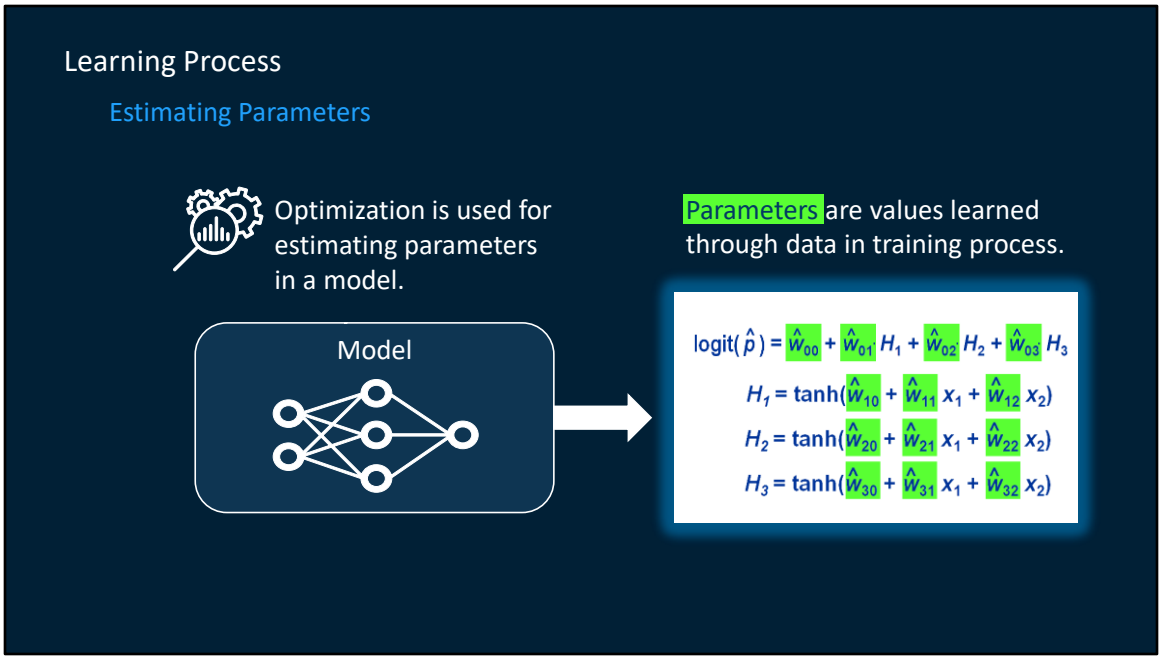


For example, gradient descent is one of the most generic optimization algorithms capable of finding optimal solutions to a wide range of problems. *Gradient* means steepness of a slope. The general idea of gradient descent is to modify parameters iteratively in order to minimize a cost function. It measures the local gradient of the error function, and it goes in the direction of descending gradient.

How big a step to take is governed by the learning rate. *Learning rate* specifies the step size for gradient descent optimization. When the gradient is zero, you have reached a minimum!

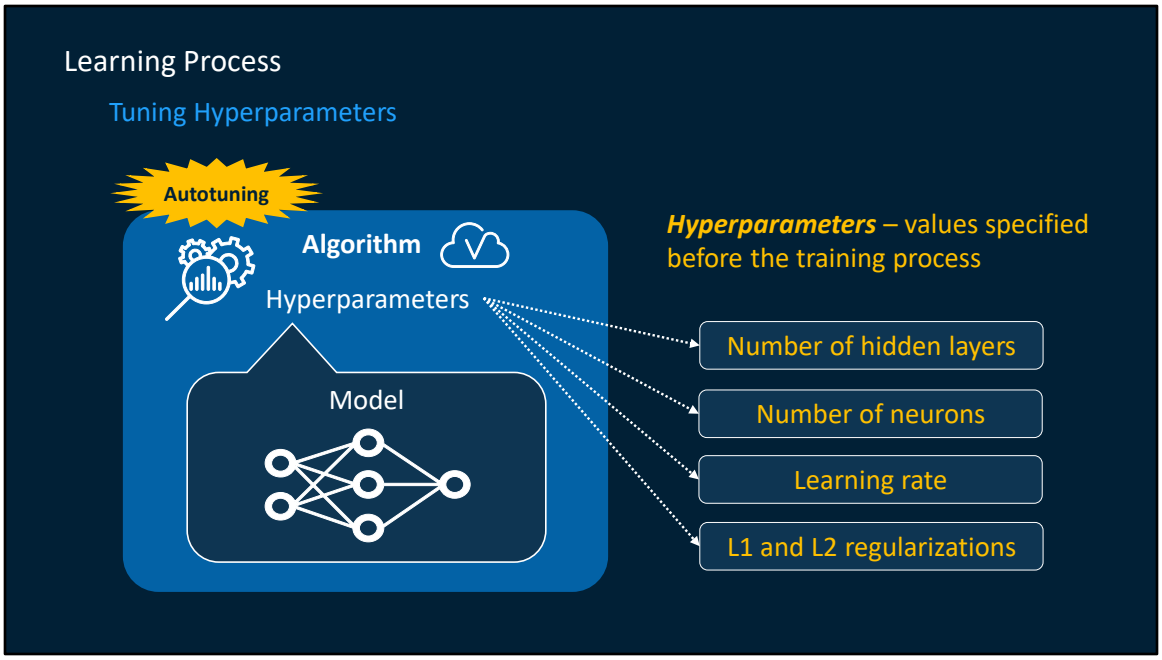
You start with random values (this is called *random initialization*), and then you improve it gradually, taking one baby step at a time, each step attempting to decrease the cost function, until the algorithm *converges* to a minimum.

As a result, each iteration has a different set of estimated parameters. For example, in case of a neural network, a different neural network model is created in every iteration.



At the core of nearly all machine learning algorithms is an optimization algorithm. For example, a global search optimization algorithm is used in a neural network model to estimate the weights and biases. These weights and biases are commonly called *parameters*. The parameters shown here are highlighted in green.

So, parameters are the values that are learned through data in the training process. They are the part of the model that is learned from historical data. Another example of parameters includes the *slope* and *intercept* in a linear regression model.



In addition, the process of working through a machine learning problem involves choosing the hyperparameters of a model.

A model hyperparameter is the parameter whose value is set before the training process. They cannot be learned by fitting the model to the data. Hyperparameters are not affected by training. In other words, training doesn't cause hyperparameters to change. Rather, hyperparameters affect the quality and speed of training. So, in a way these parameters are beyond or above (hyper-) the process of training and hence called hyperparameters.

For example, in a neural network model, you can specify the values for the number of hidden layers, the number of neurons per layer, the learning rate or L1/L2 regularizations, and so on - and these values cannot be estimated through training data. Thus, they are hyperparameters.

When building a model, you can set the value of hyperparameters for the model. Optimal values are determined using a trial-and-error process by adjusting or adapting their values for a particular machine learning task. This process of determining the values of hyperparameters is called hyperparameter tuning.

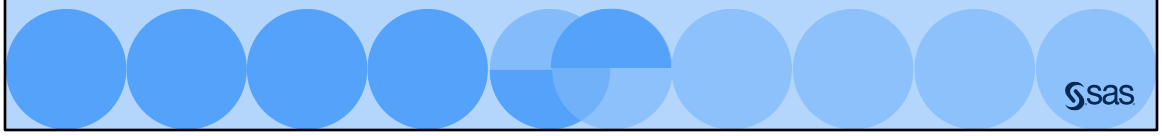
However, so many choices can be overwhelming. Then how do you select the best values for these hyperparameters, especially considering their varied effect on different data sets? SAS Viya automates the selection of hyperparameter values using an intelligent methodology. This capability can significantly improve the accuracy of the resulting model with no additional effort from the modeler. The autotuning feature uses optimization algorithms to select the optimal hyperparameters.

Parameters versus Hyperparameters

To explore this learning resource in greater detail, please visit the course page on learn.sas.com. There, you'll find comprehensive course materials, tutorials, and updates to support your learning journey.

More about Estimation Criterion

To explore this learning resource in greater detail, please visit the course page on learn.sas.com. There, you'll find comprehensive course materials, tutorials, and updates to support your learning journey.



Question: Learning Terminology

Which of the following terms would you use for the following description?

The difference between what is observed in the data set and what is predicted by your model.

a. Random initialization
b. Iterations
c. Error function
d. Mistake

Answer: Learning Terminology

Which of the following terms would you use for the following description?

The difference between what is observed in the data set and what is predicted by your model.

a. Random initialization
b. Iterations
c. Error function
d. Mistake

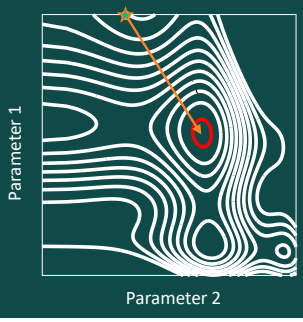
Correct answer: c

The word ‘error’ is used in specified sense in statistics. It doesn't mean the same thing as a ‘mistake’. Mistake means wrong calculation or use of inappropriate method in the collection or analysis of data. Error means the difference between the true value and the estimated value.
Random initialization can be described as starting an optimization process with random values of weights.
Iterations are the steps that the machine learning model takes while optimizing the model.

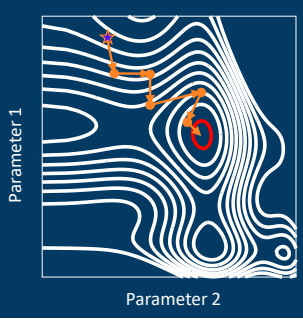
Question: Finding Parameter Values

Which of the following is true for finding the parameter values for neural networks?

a. Explicit formulas for the optimal estimates can be determined.



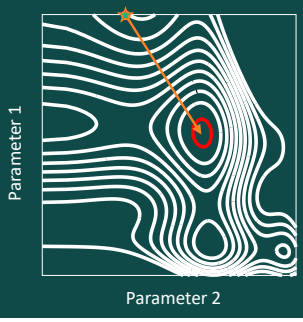
b. Iterative numerical optimization methods are used.



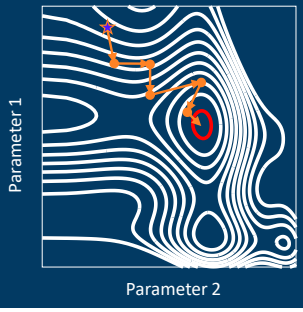
Answer: Finding Parameter Values

Which of the following is true for finding the parameter values for neural networks?

a. Explicit formulas for the optimal estimates can be determined.



b. Iterative numerical optimization methods are used.

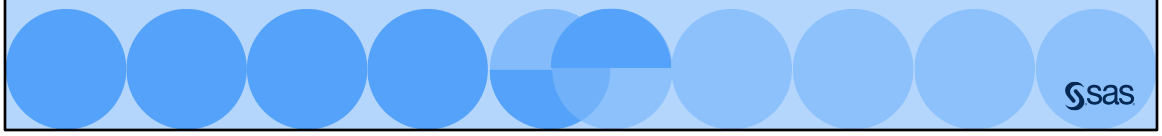


ANSWER: b

For some simple regression models, explicit formulas for the optimal estimates can be determined. Finding the parameter values for neural networks, however, is more difficult. Iterative numerical optimization methods are used. First, starting values are chosen. The starting values are equivalent to an initial guess at the parameter values. These values are updated to improve the estimates and reduce the error function. The updates continue until the estimates converge (in other words, there is no further progress).

Demo: Running a Neural Network Model and Tuning Its Hyperparameters

To explore this learning resource in greater detail, please visit the course page on learn.sas.com. There, you'll find comprehensive course materials, tutorials, and updates to support your learning journey.

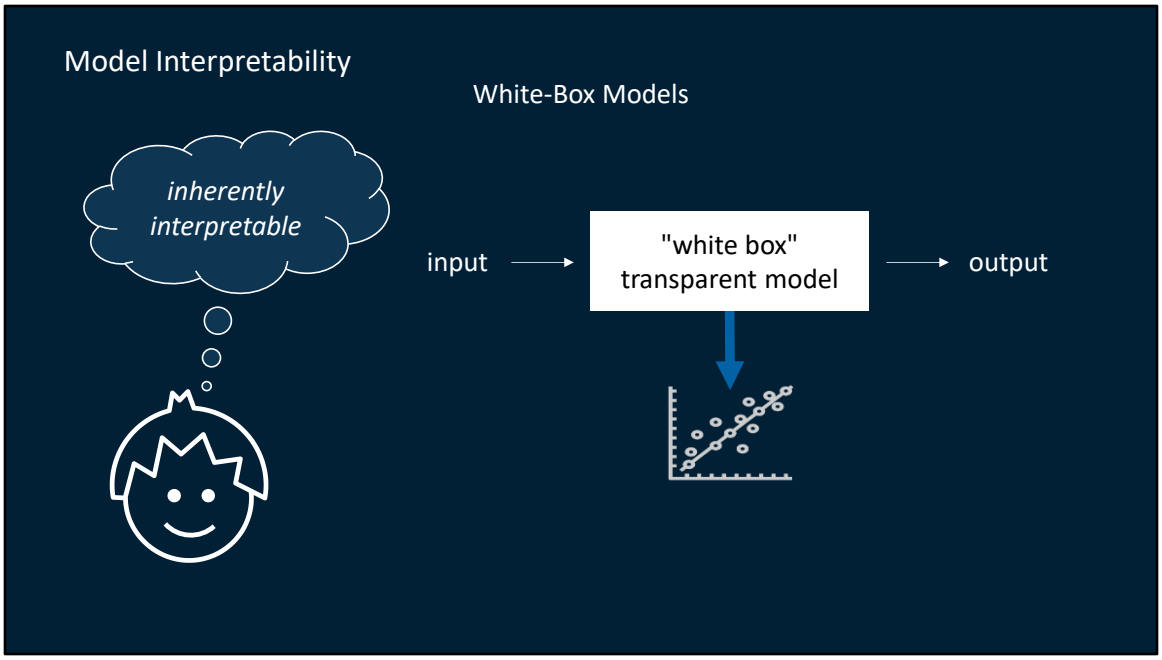


Model Interpretability

Data Difficulties and Modeling Issues

- Visualizing big data
- Processing noisy data
- Handling high dimensional data
- Dealing with distinct variability in inputs
- Using available data judiciously
- Adjusting model flexibility and complexity
- Understanding learning and tuning of models
- Making models explainable

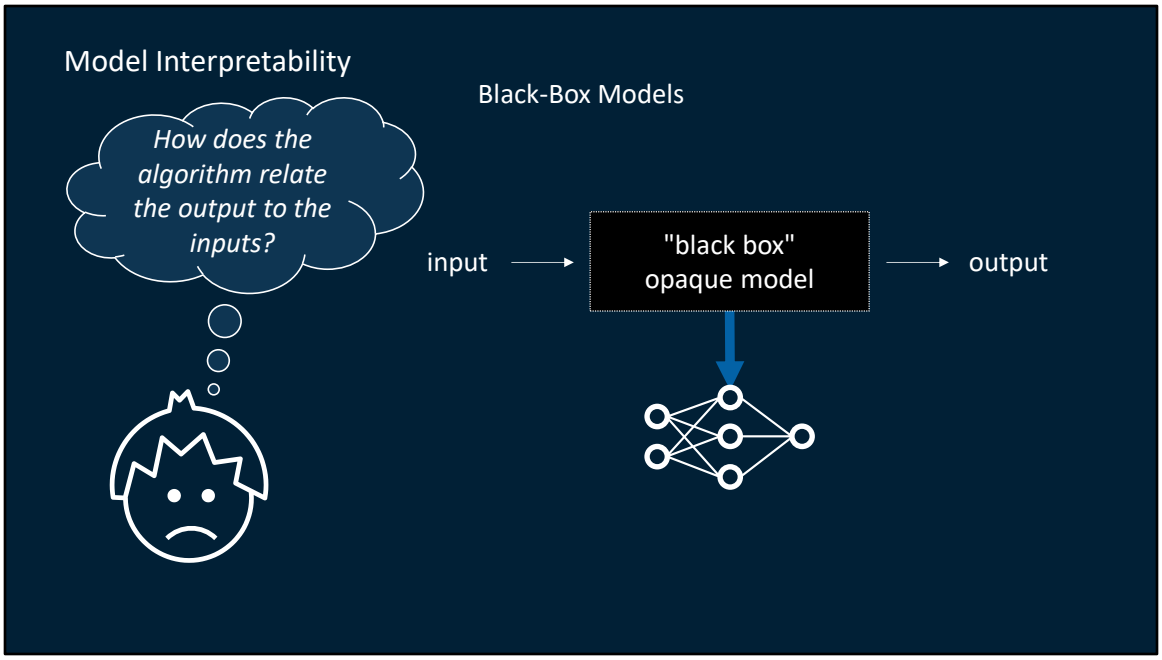
Let's discuss making models explainable, starting with white-box models.



Inherently interpretable models (also called *explainable models*) incorporate interpretability directly into the model structure, and thus are self-explanatory. One type of commonly used inherently interpretable model is the generalized linear model (GLM), which includes linear and logistic regression. The coefficient estimates of GLMs directly reflect feature contributions. Hence, these models can be explained through these coefficients.
Some machine learning models are simple and easy to understand.

For example, a decision tree is a transparent, or "white box," model because you know how changing the inputs will affect the predicted outcome and can make a justification for each prediction.

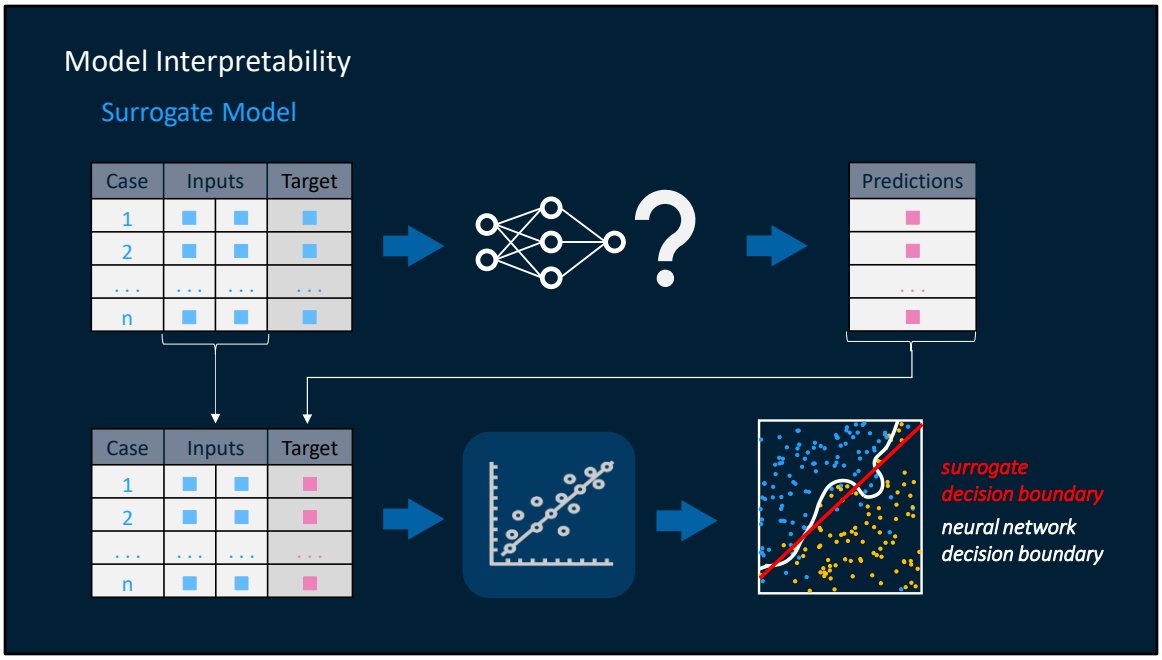
Inherently interpretable models generally achieve interpretability by forcing the models to use fewer features for prediction. Constraints on features can make complex models simpler and increase the model's comprehensibility to users. However, imposing these constraints can also decrease the predictive ability of the model when compared to an unrestricted model.



Modern machine learning algorithms can make accurate predictions by modeling the complex relationship between inputs and outputs. However, the models that are produced can be equally complex, making it difficult to understand the association between the inputs and outputs. Because of their complexity, many machine learning models are "black-box" models, producing predictions without explaining why and how they are made.

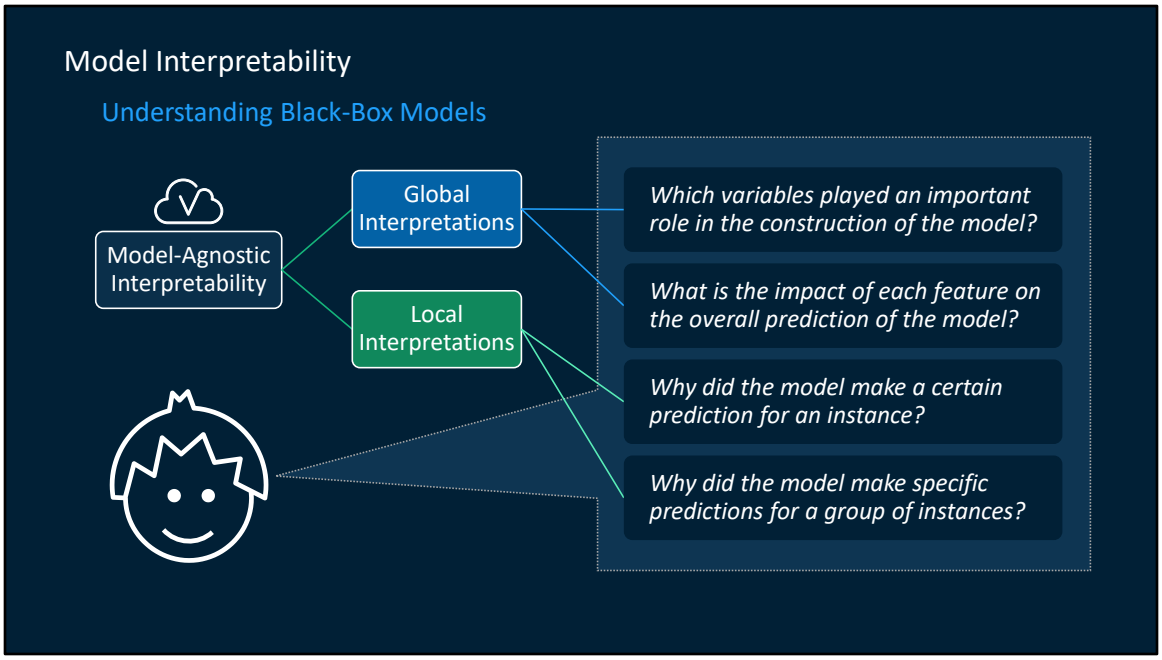
One example of a black-box machine learning model is a neural network model with one or two hidden layers. Though you can write out the equations that link every input in the model to every output, you might not be able to grasp the meaning of the connections simply by examining the equations. This has less to do with the shortcomings of the models and more to do with the shortcomings of human cognition. Often, the higher the predictive accuracy of a model, the harder it is to interpret its inner workings. So, how the algorithm creates the model is a vital problem. Algorithm transparency is about how the algorithm learns a model from the data and what type of relationships it can learn. Algorithm transparency requires only knowledge of the algorithm and not of the data or learned model.

This is where interpretability techniques come into play by providing a lens through which you can view these complex models.



Although it is true that little insight can be gained by analyzing the actual parameters of neural networks and similar flexible models, much can be gained by analyzing the resulting predictions.

A white-box surrogate model can be created using the predictions that are produced by the pretrained neural network as a target. For example, you can approximate an inscrutable model's predictions with a regression model or a decision tree. This surrogate model approximates the behavior of the black-box model, but it is much easier to explain by examining the regression coefficients. This surrogate model does not have to work very accurately (red straight line), but it must approximate the behavior of the pretrained model's decision boundary (white curvilinear line) with some loss of generality. The idea behind a surrogate model is building an easy-to-understand model that describes the decision regions. In this way, characteristics used by the neural network to make the primary decision can be understood, even if the neural network model itself cannot be. A surrogate model can provide interpretation of the model, but it should never be assumed that the surrogate explains the true nature of how the original model makes predictions.



The idea of creating a surrogate model is the bottom line behind several model-agnostic interpretability methods that are used to explain trained supervised machine learning models such as boosted trees, forests, and neural networks. These post-hoc techniques explain predictions of these models by treating the models as black boxes and then generating explanations without inspecting the internal model parameters.

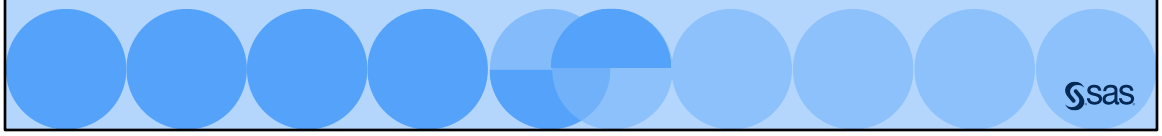
Model-agnostic interpretability techniques available in SAS Viya enable fully complex models to be interpreted either globally or locally.

Global interpretability provides explanations about the general behavior of the model over the entire population. For example, global interpretability might explain which variables played an important role in the construction of the model or describe the impact of each feature on the overall prediction of the model. Variable importance plots and partial dependence plots are global interpretability techniques.

In contrast, local interpretations promote understanding of a single data point or of a small region of the distribution, such as a cluster of input records and their corresponding predictions, or decile of predictions and their corresponding input rows. Because small sections of the conditional distribution are more likely to be linear, local explanations can be more accurate than global explanations. You can zoom in on a single instance and examine what the model predicts for this input and explain why. The global methods can be applied by taking the group of instances, treating them as if the group were the complete data set, and using the global methods with this subset.

Why Interpretability Matters

To explore this learning resource in greater detail, please visit the course page on learn.sas.com. There, you'll find comprehensive course materials, tutorials, and updates to support your learning journey.



Question: Interpreting Black-Box Models

Machine learning models are mostly black-boxes. However, SAS Viya provides several model interpretability plots that help interpret machine learning model results.

a. True

b. False

Answer: Interpreting Black-Box Models

Machine learning models are mostly black-boxes. However, SAS Viya provides several model interpretability plots that help interpret machine learning model results.

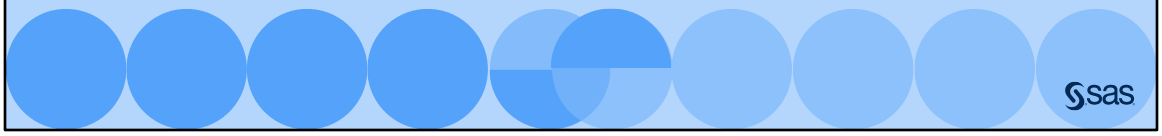
a. True
b. False

Answer: True

Model-agnostic interpretability techniques available in SAS Viya enable fully complex models to be interpreted either globally or locally.

Lesson Summary

To explore this learning resource in greater detail, please visit the course page on learn.sas.com. There, you'll find comprehensive course materials, tutorials, and updates to support your learning journey.



Where Should I Go From Here?

You are ready to take any of the following SAS courses:

Machine Learning Using SAS Viya

Interactive Machine Learning in SAS Viya

Supervised Machine Learning Procedures Using SAS Viya in SAS Studio

SAS Visual Statistics in SAS Viya: Interactive Model Building