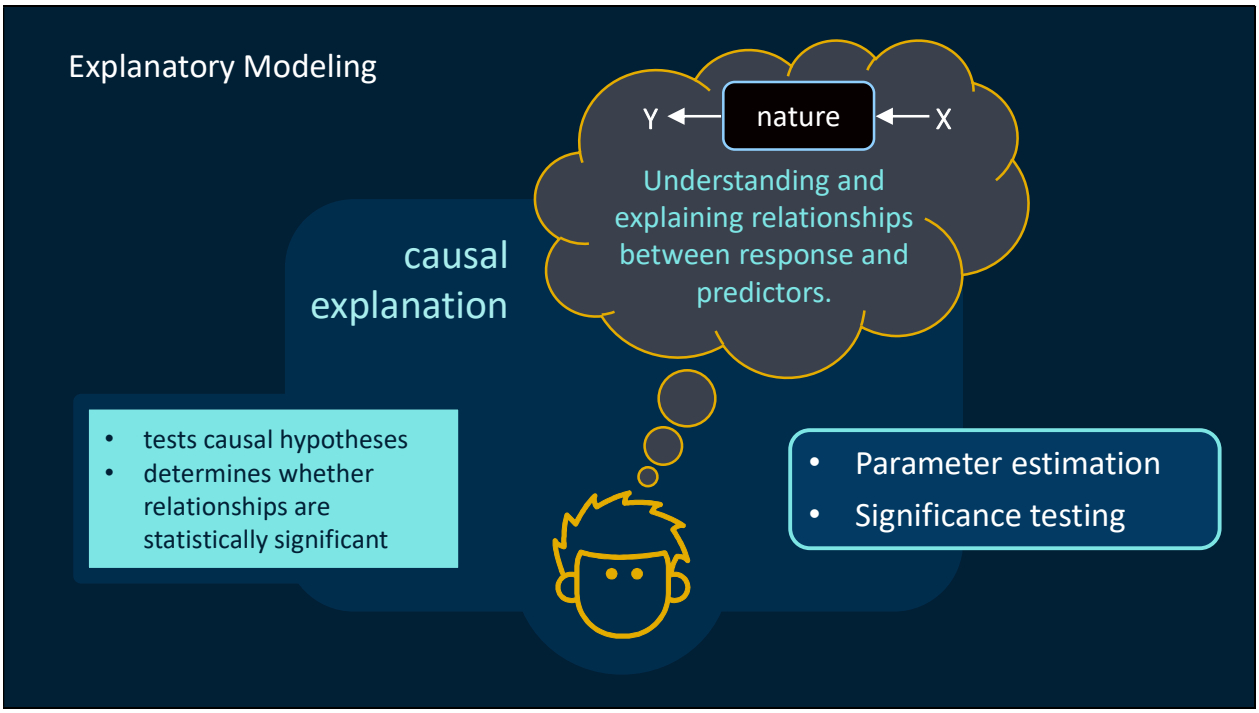
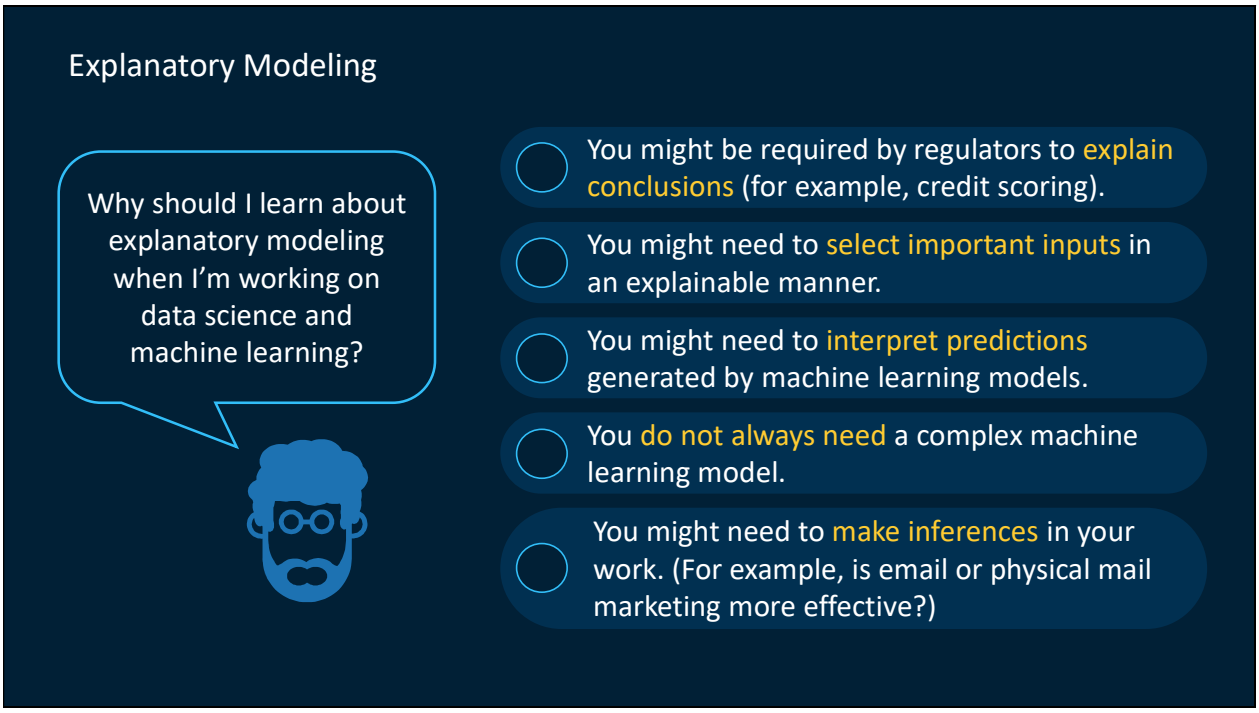
Lesson 3 Explanatory Modeling Using Linear Regression

3.1 Relationships between Variables



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.

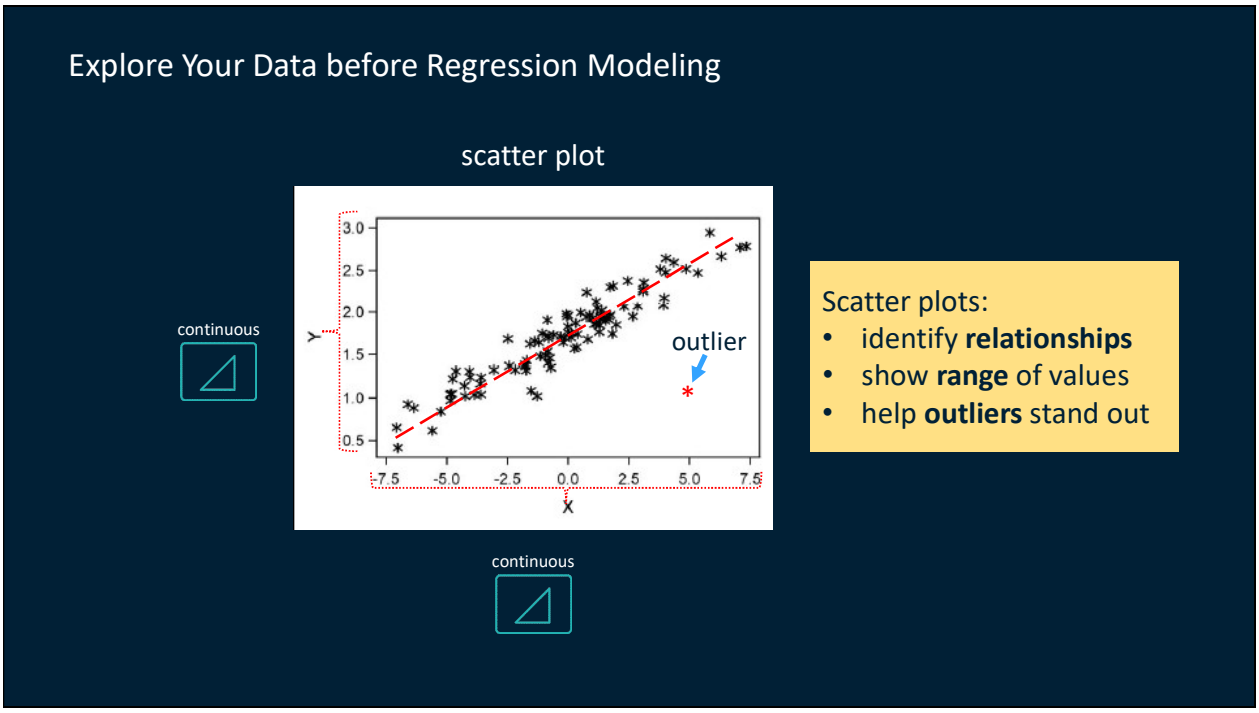
In many areas, regulators require an explanation of the model results, which is greatly facilitated by using tools and methodology of explanatory modeling. For example, in credit scoring, loan decisions need to stand up to regulators' scrutiny, and a logistic regression model is commonly used. Financial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

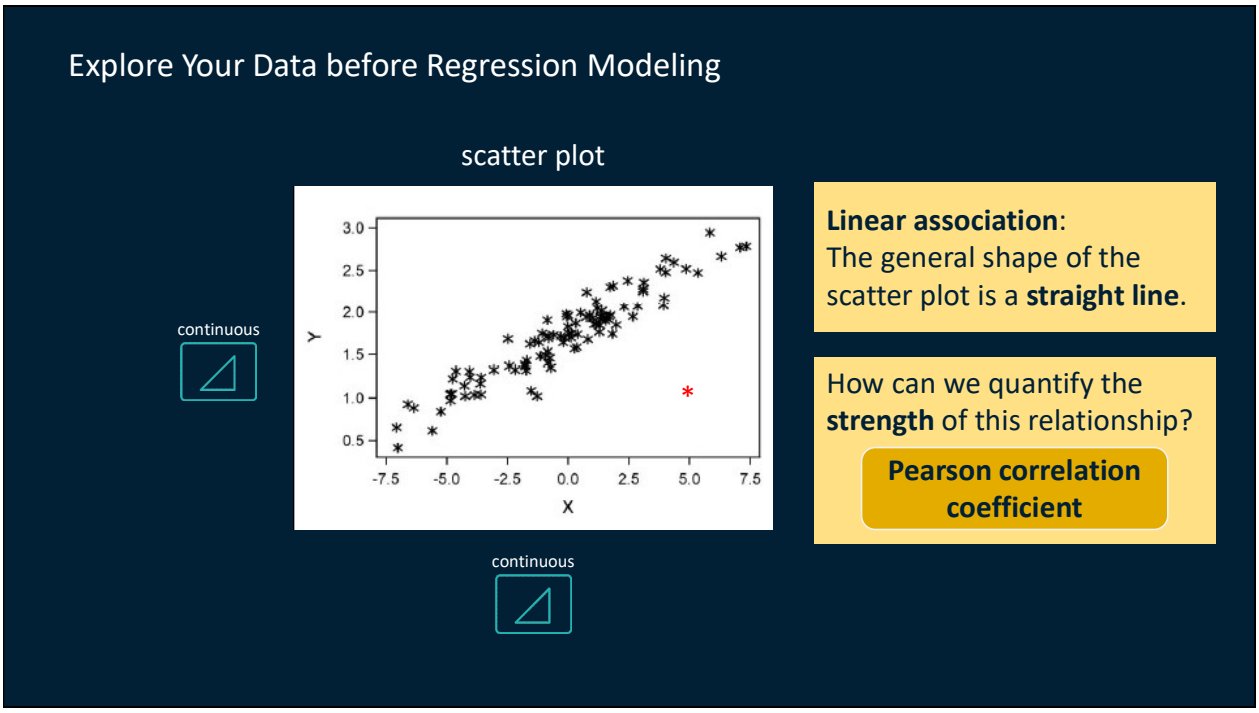
Most of the machine learning algorithms focus on predictive accuracy rather than interpretability of the model. In those situations, explanatory modeling comes to your rescue by running regression models in order to provide prediction explanations.

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.

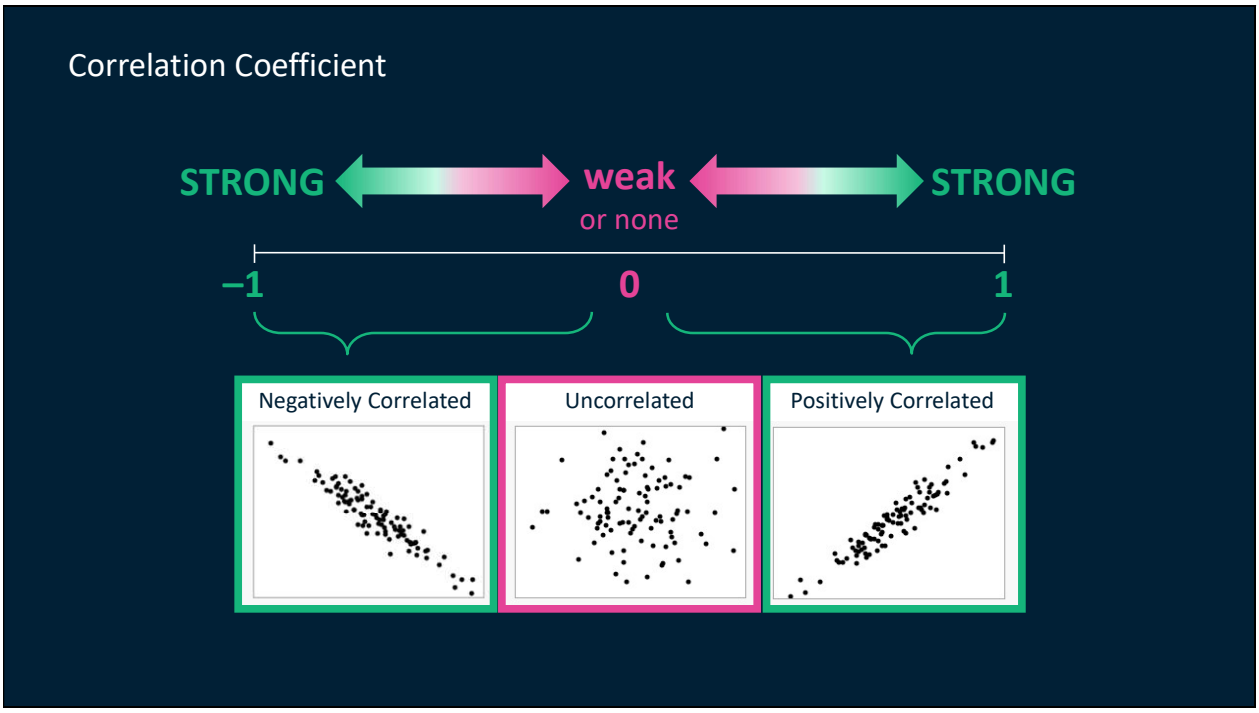
Another reason is that, despite a focus on prediction, there can be situations in which you need to make inferences in your work. Marketers might need to not only predict who is likely to buy a product, but also inform their department whether email marketing or physical mail marketing tends to get better responses.



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.



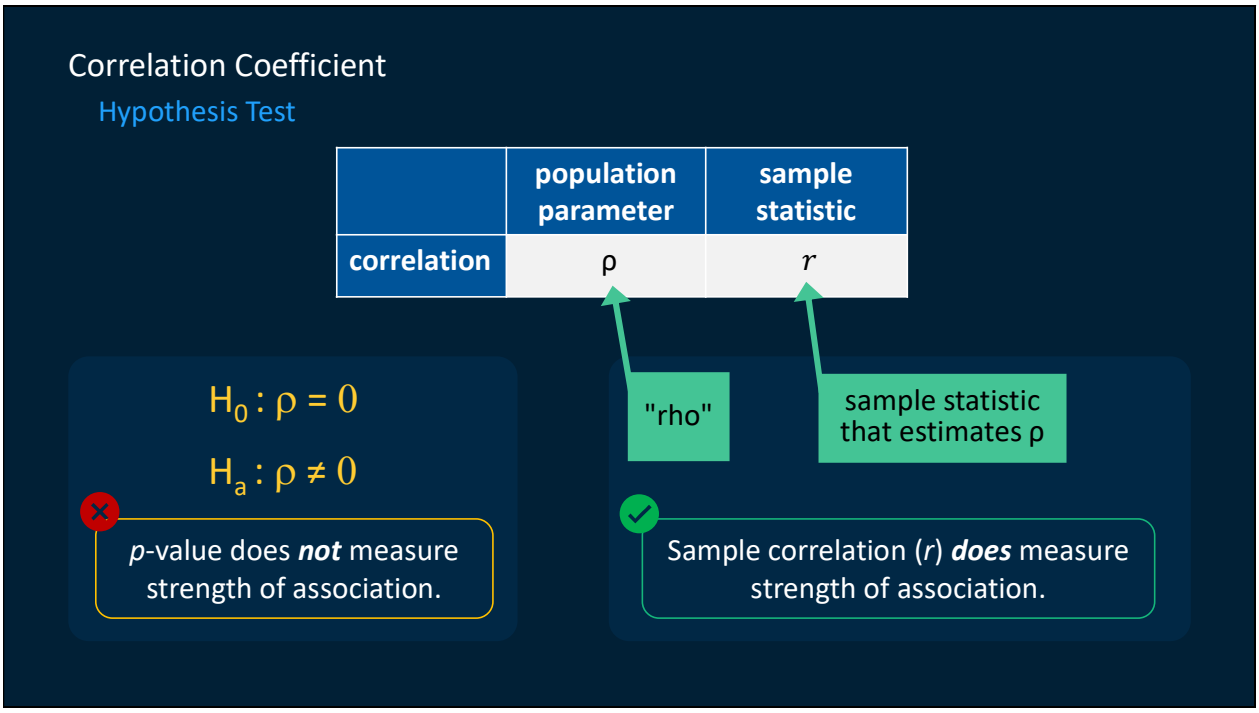
In this picture, the data points line up. We could describe the relationship pictured as a positive linear association between continuous variables. Are we seeing a strong or weak relationship? How can we quantify the strength of this relationship? A statistic that describes the strength of a linear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

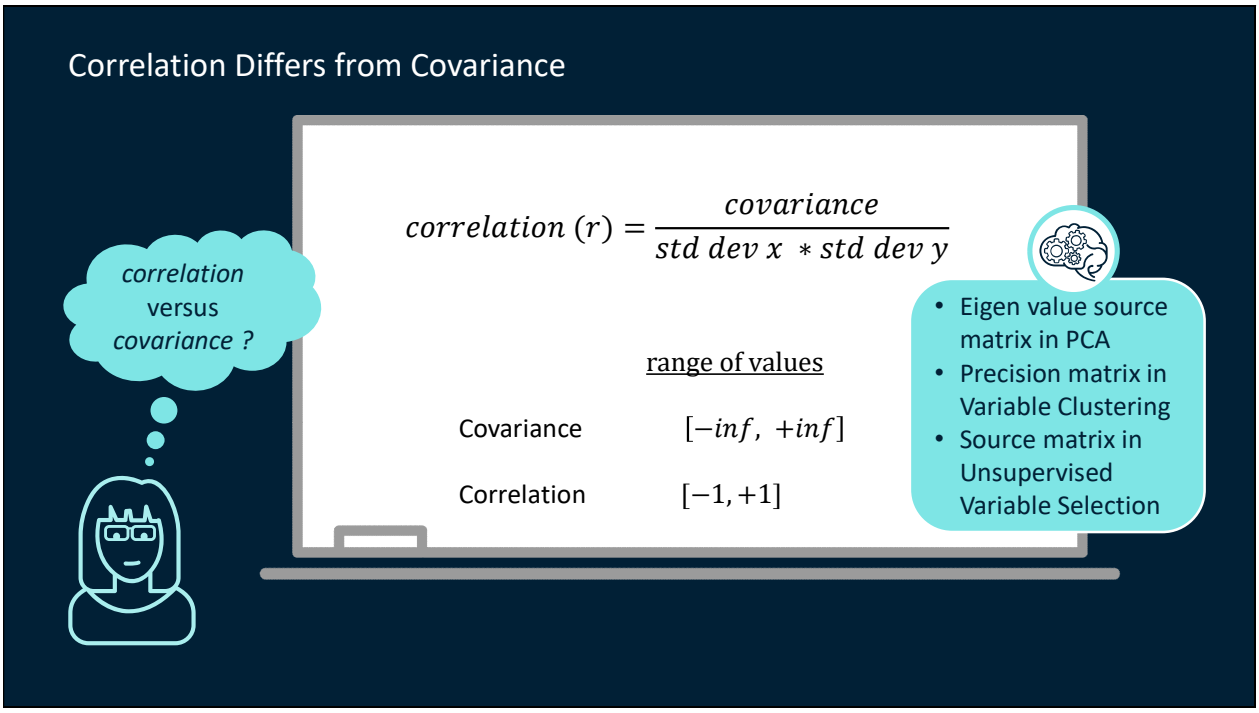
If one variable tends to increase in value as the other variable increases in value, this is a positive correlation. If one variable tends to decrease in value as the other variable increases, the correlation is negative. If no linear relationship exists between the two variables, the correlation is 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.



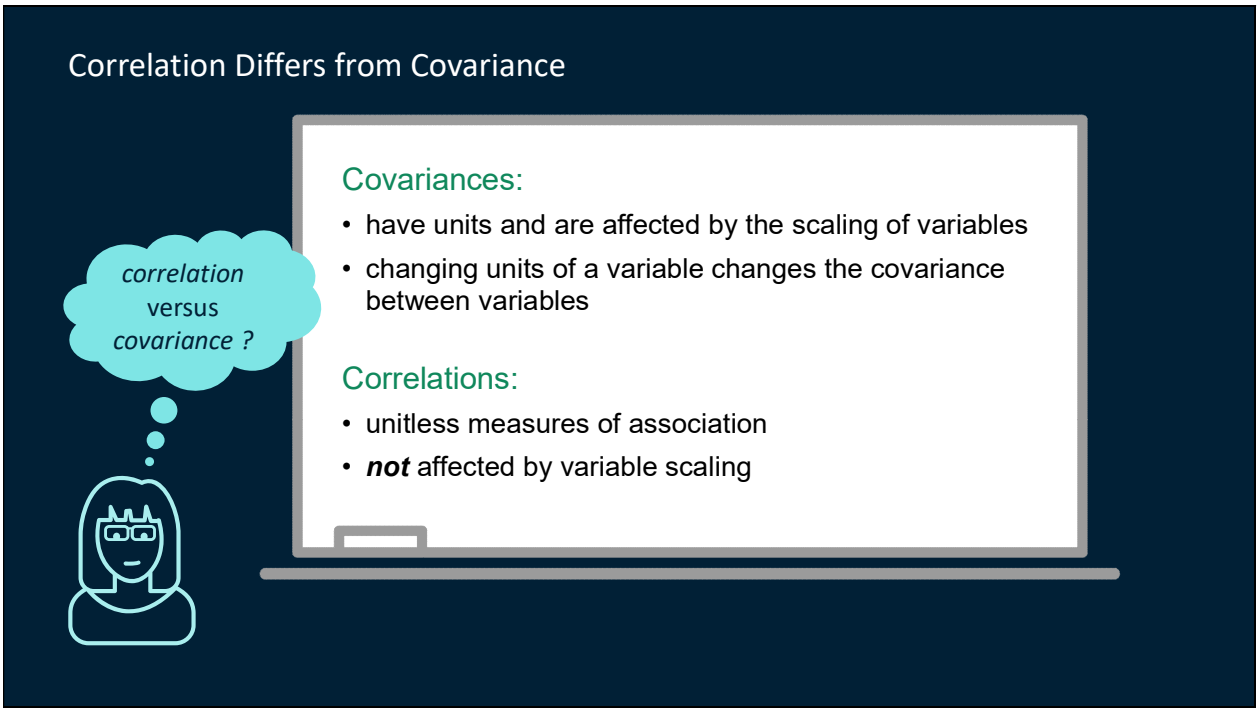
The population parameter that represents a correlation is ρ, and the correlation coefficient *r* is the sample statistic that estimates ρ. The null hypothesis for a test of a correlation coefficient is that ρ equals 0, and the alternative hypothesis is that ρ is not equal to 0. Rejecting the null hypothesis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

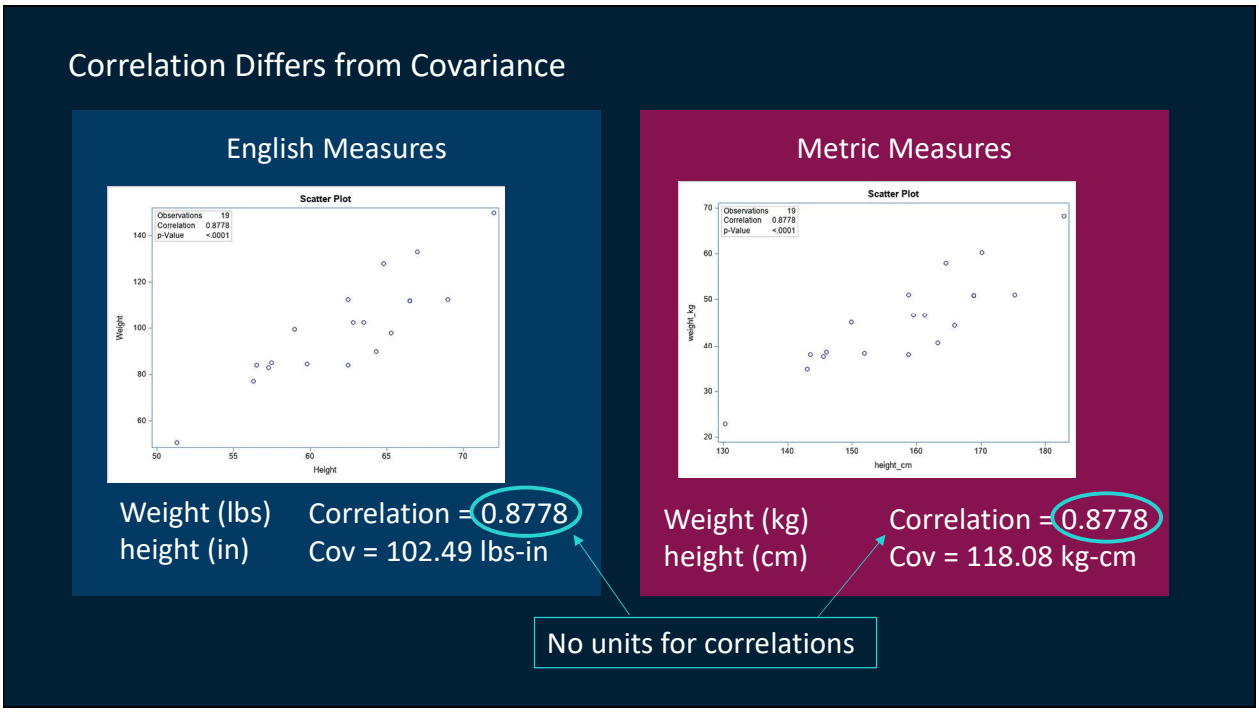


Students sometimes confuse correlation with covariance because both describe an association between continuous variables that can take on positive or negative values. **At times you need to make a choice between a correlation matrix and a covariance matrix.** In machine learning, for example, you need 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 positive infinity to negative infinity, the correlation will have values between positive and negative 1.

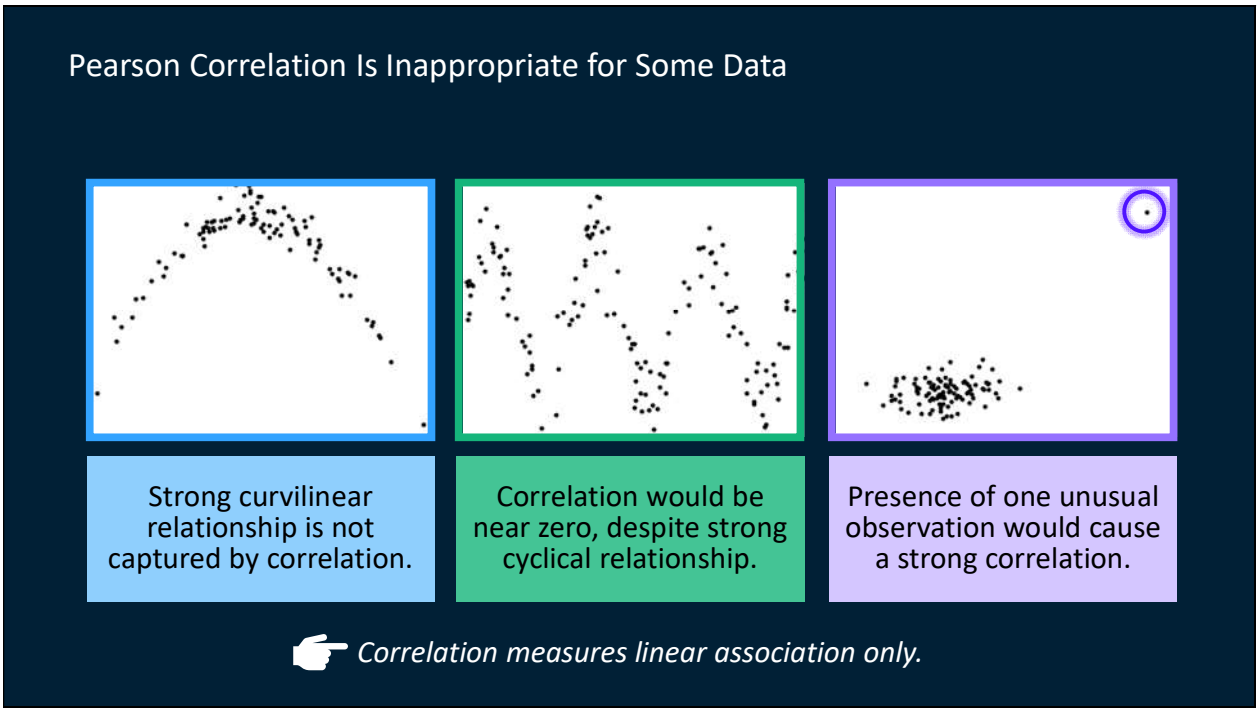


Another difference between covariance and correlation is that covariances have units and are affected by the scaling of variables. Changing the units of a variable changes the covariance between variables. Correlations are unitless measures of association and are not affected by variable scaling.



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.

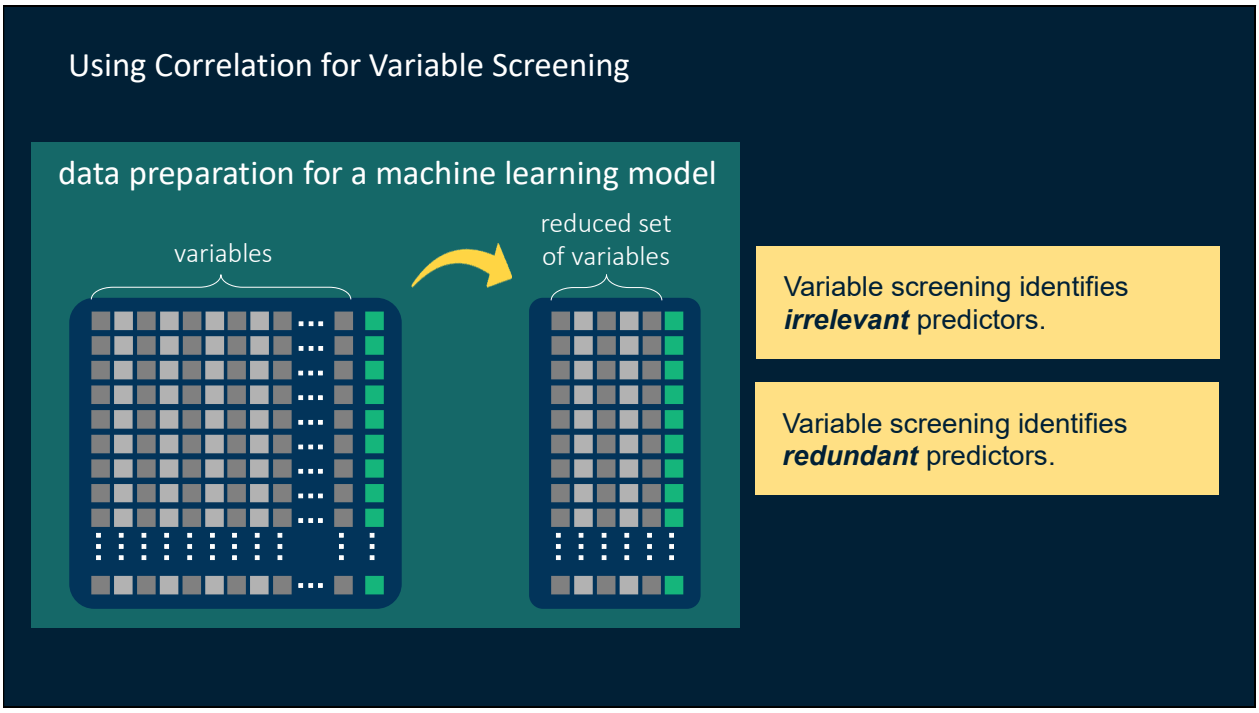


Several relationships between continuous variables are not linear in nature and thus would be inappropriate to quantify with correlation. One such relationship is shown on the left. This picture shows a strong curvilinear relationship between X and Y. Although it is possible to calculate a correlation (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.

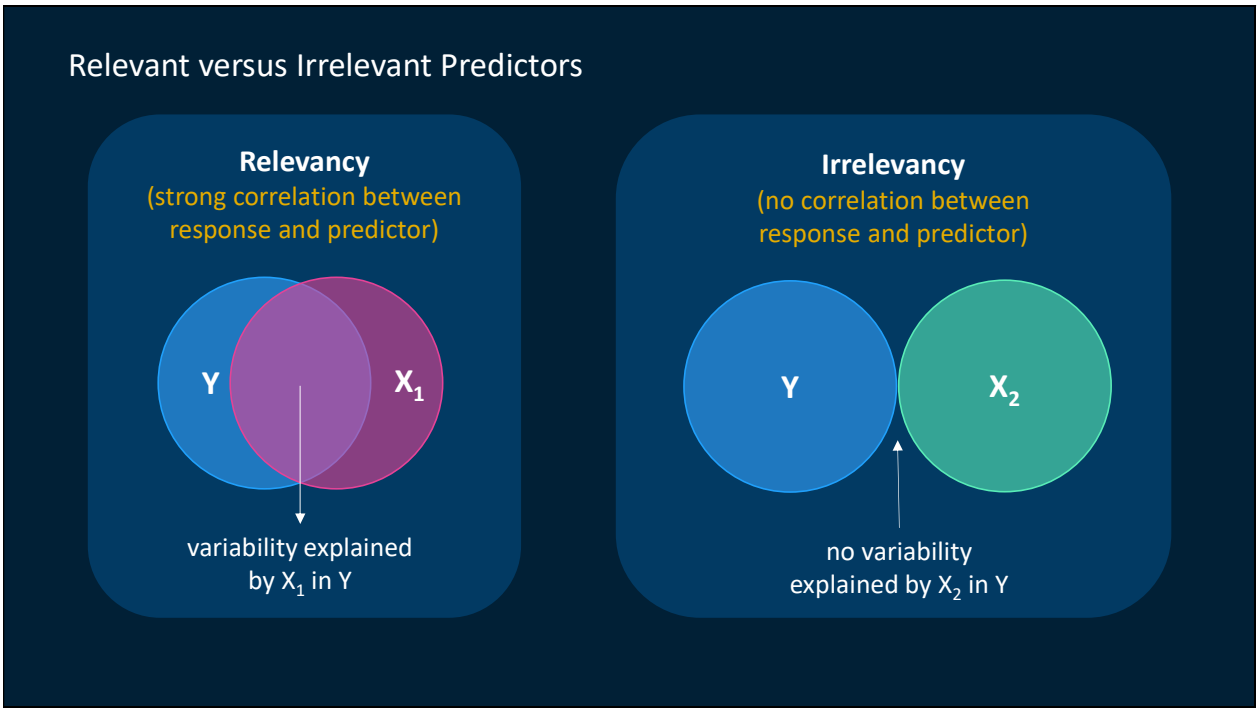
A correlation coefficient would also provide a misleading description of the data on the right. Here, there is essentially no relationship between X and Y. But the presence of one unusual observation would cause a strong positive X, Y correlation. Without more data to fill the gap in information between 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.



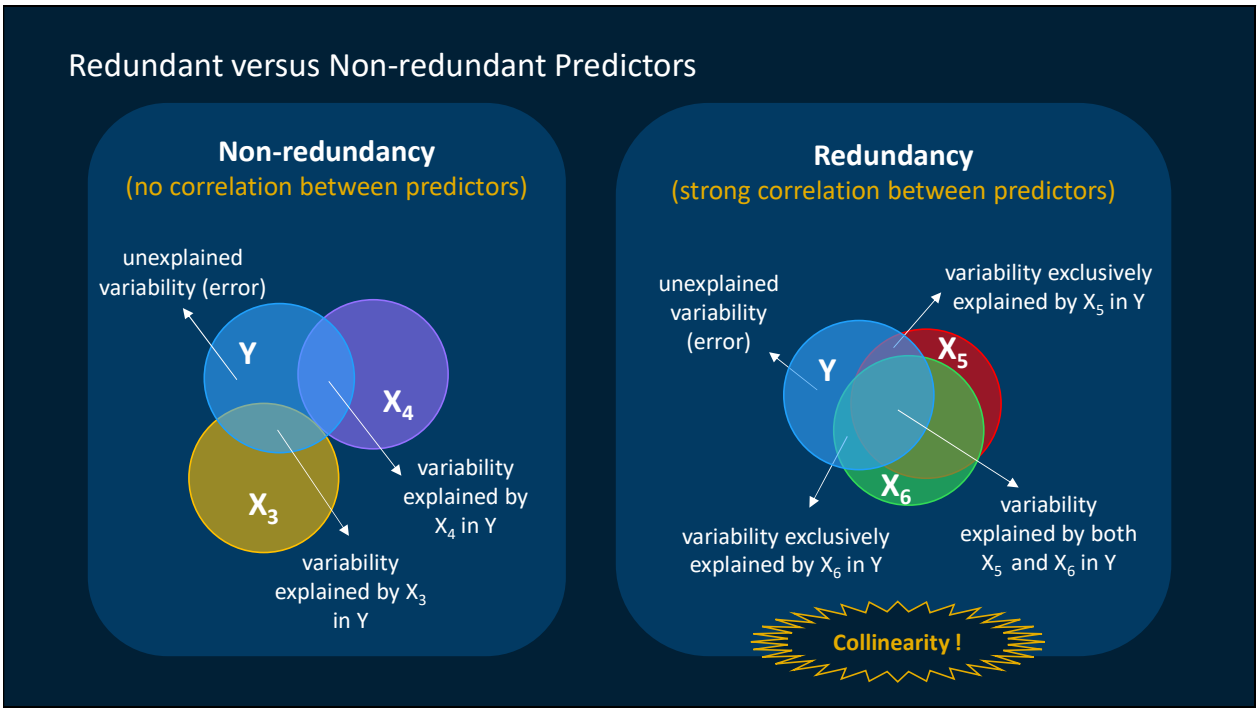
Correlations are not only useful for explanatory modeling. They are often used during data preparation before building a machine learning model as part of variable screening. When the researcher has many predictors, variable reduction can be important, particularly for models that have no built-in variable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



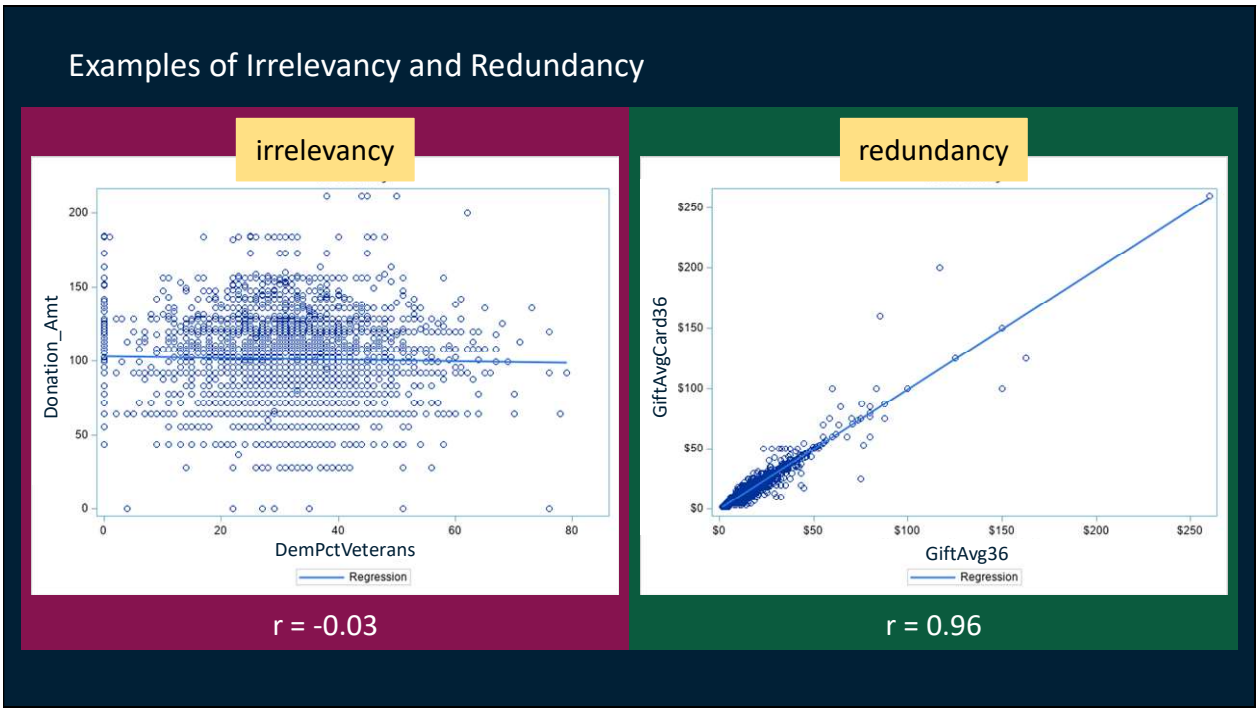
A predictor is said to be *relevant* when it has a strong correlation with the response variable, like $\mathbf{X_1}$ as shown in the illustration on the left. A relevant predictor will then explain a greater amount of variability in the response variable (**Y**).

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**).



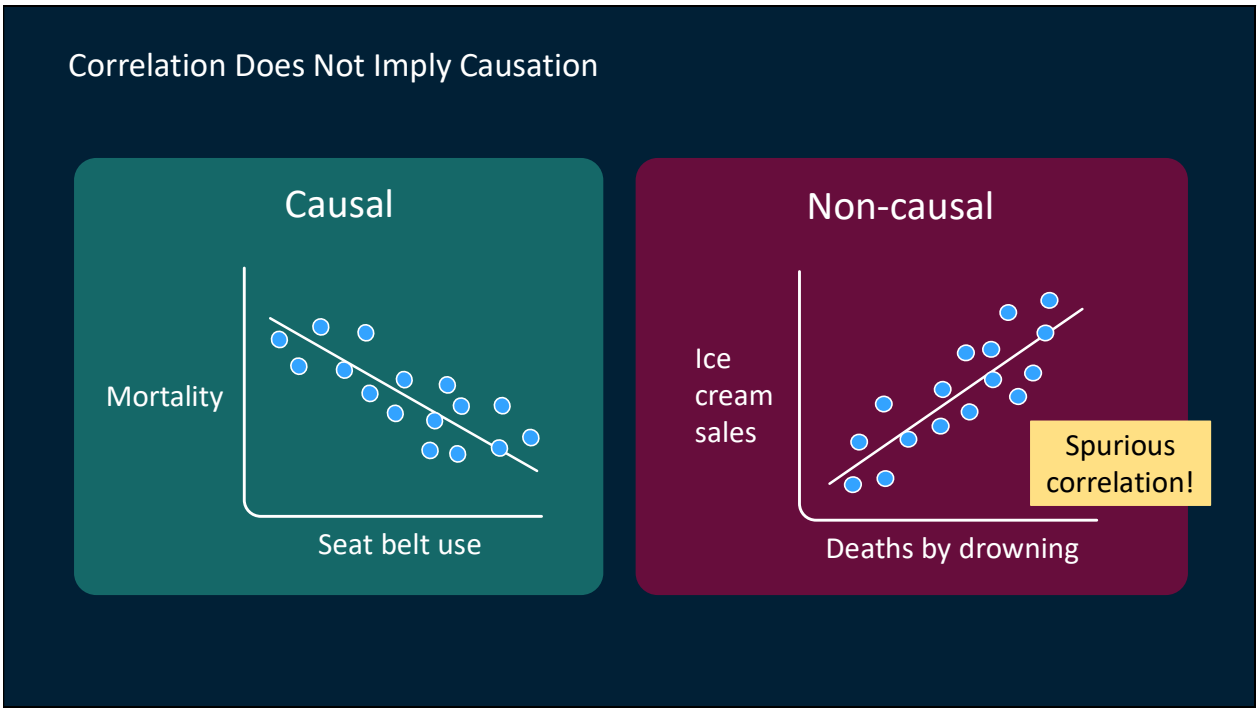
Two predictors are said to be *non-redundant* when they do not have a strong correlation with each other, like $\mathbf{X_3}$ and $\mathbf{X_4}$ as shown in the illustration on the left. Non-redundant or *independent* predictors do not poorly affect the variability explained by them in the response variable.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

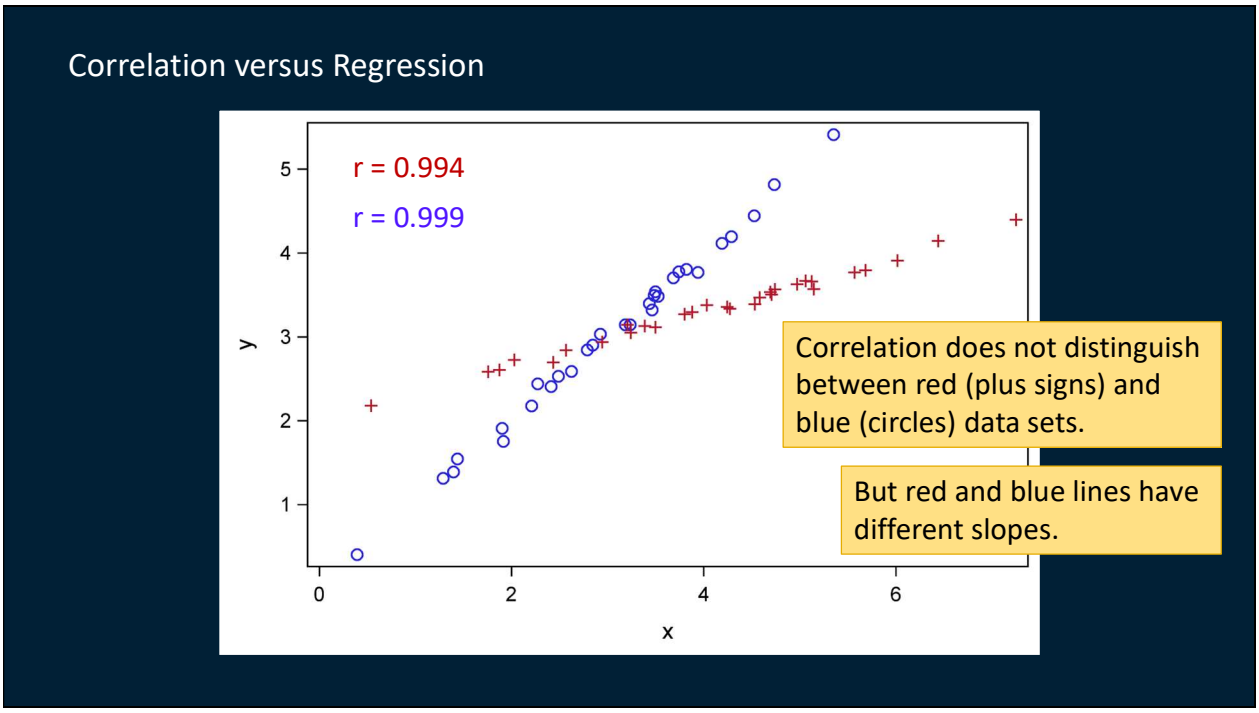


In the first graph shown here, the predictor **DemPctVeterans** has a near zero correlation with the target variable, **Donation_Amt**. This predictor could be considered irrelevant for prediction and discarded from analyses.

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.

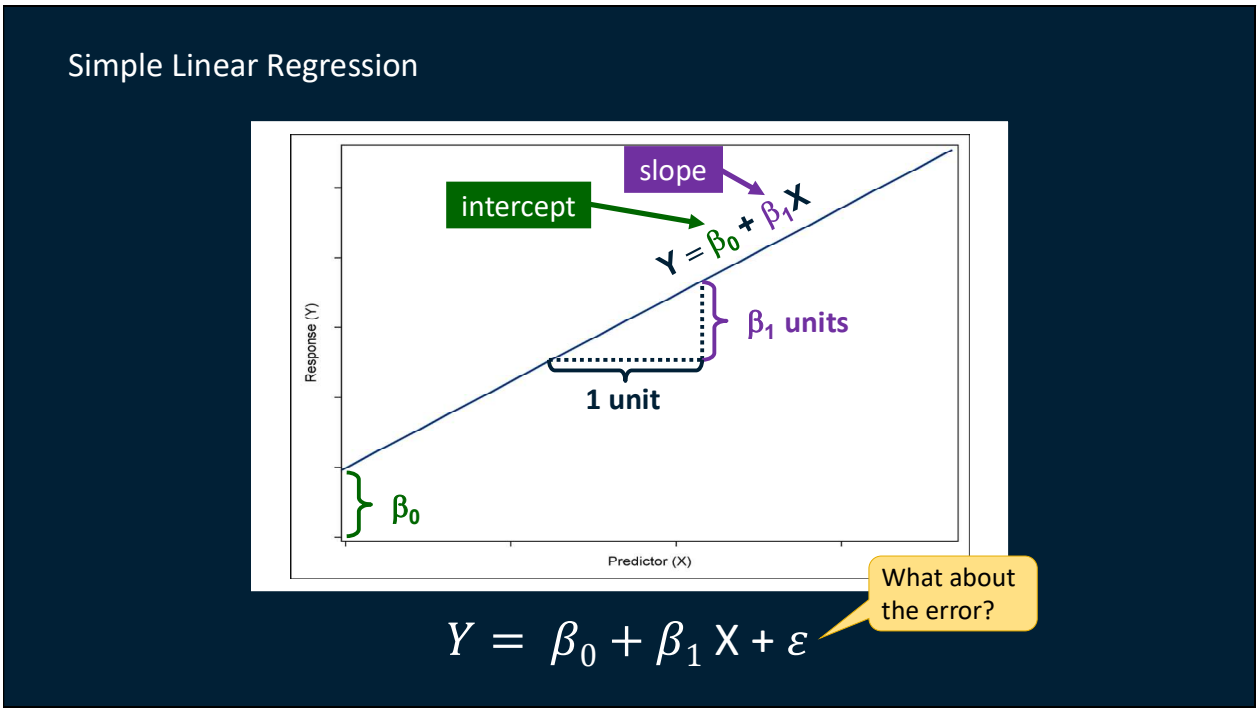


Some correlations are due to a causal relationship between variables. The negative correlation between seat belt use and mortality in vehicular accidents is a causal relationship. But many other associations are not due to causal relationships. For example, a strong correlation exists between ice cream 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.



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?

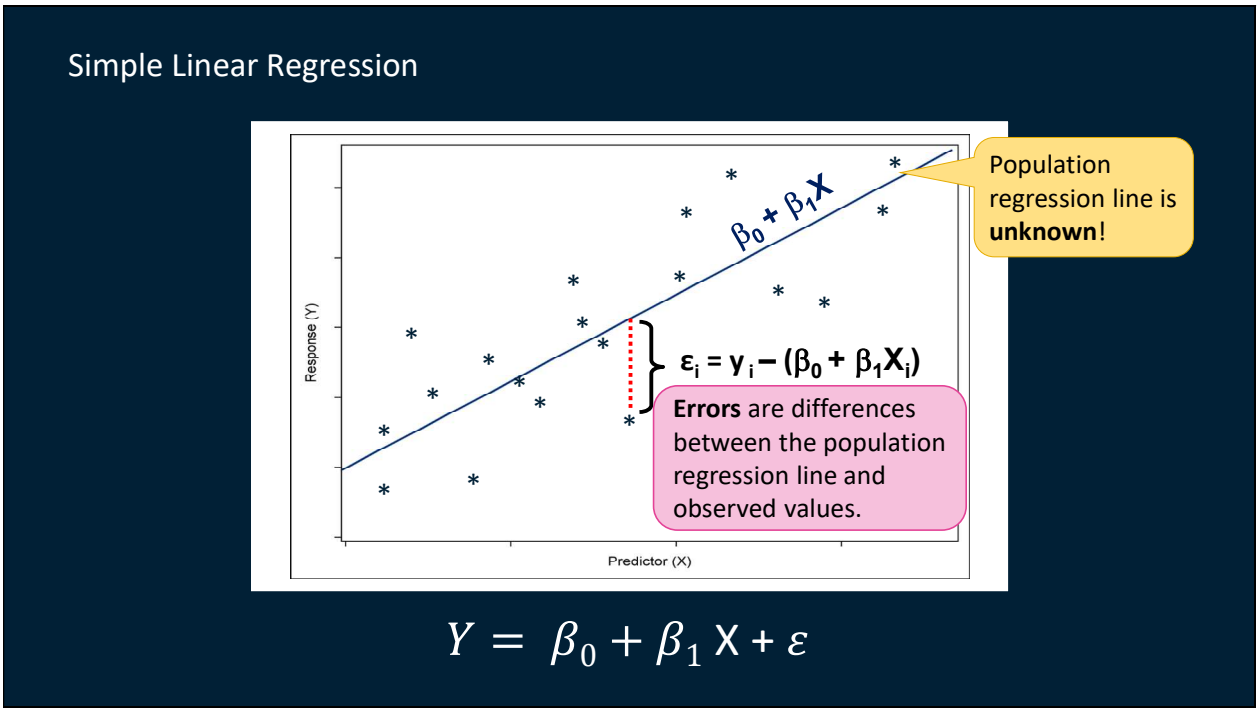


Here's a picture of the simple linear regression model.

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

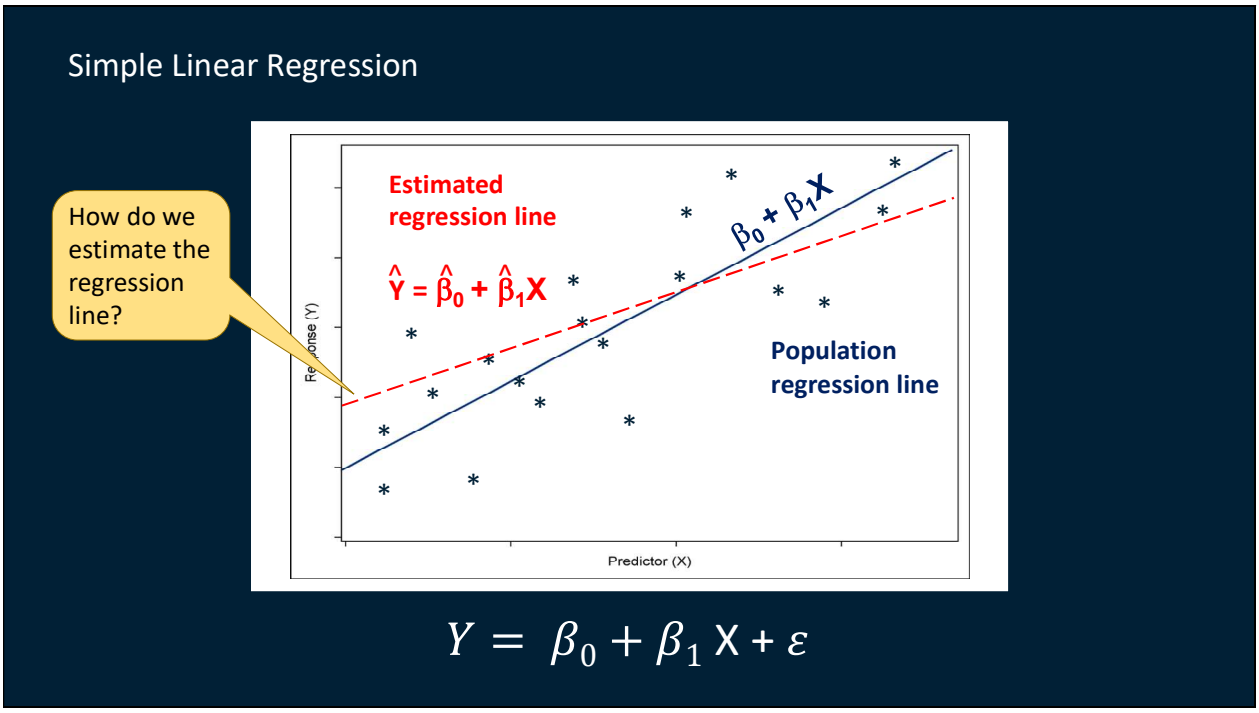
The term *simple* means that it involves only one predictor. Two or more predictors would make this a multiple regression model.

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.



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.

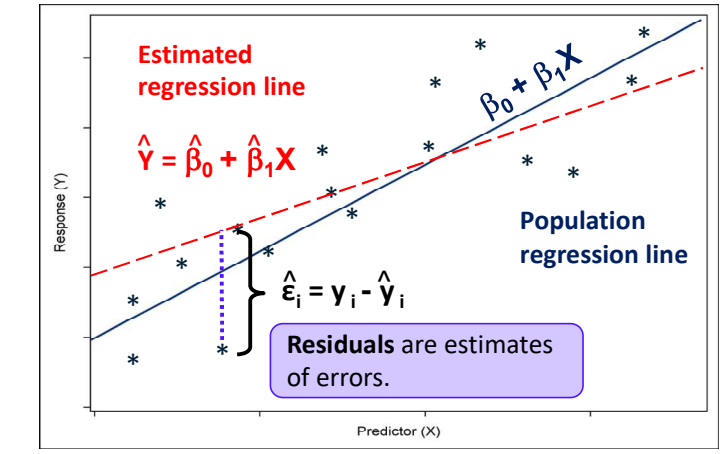
Errors are population parameters. Because we never know the true values of the population parameters β_0, β_1, we never know the true values of the errors. The parameters β_0, β_1, and ε need to be estimated from a sample.



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

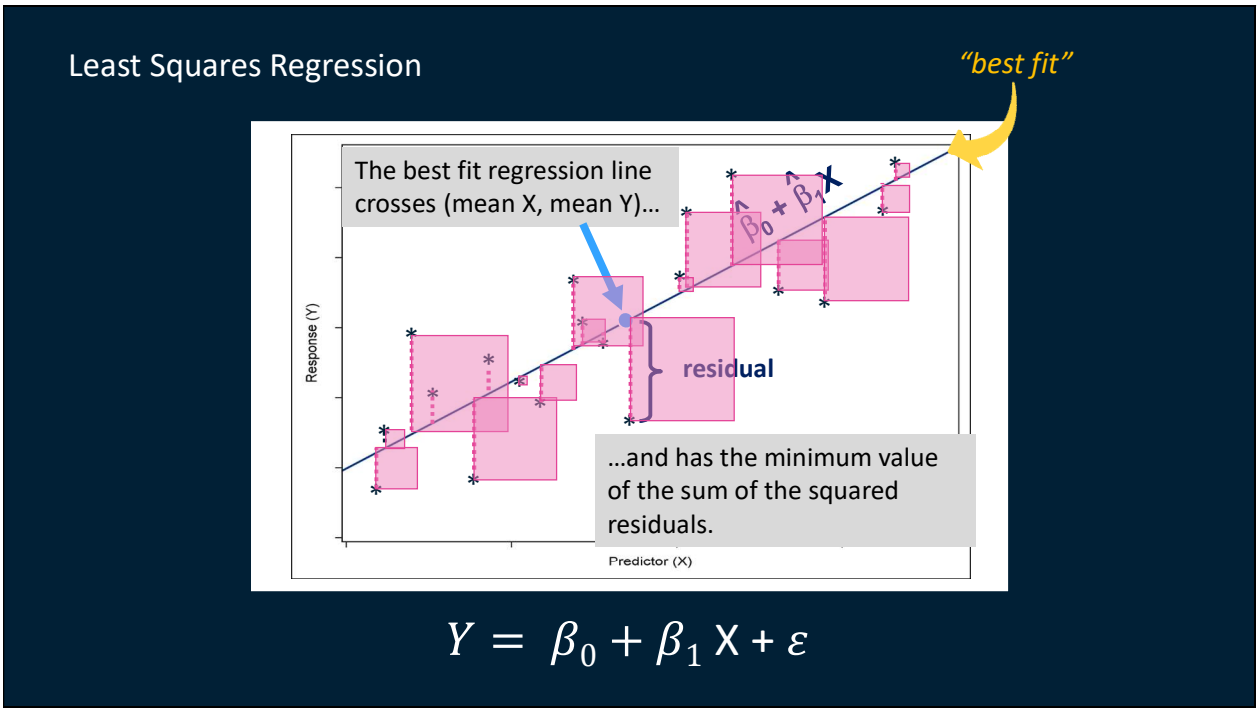
Estimating and assessing regression involves residuals.



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

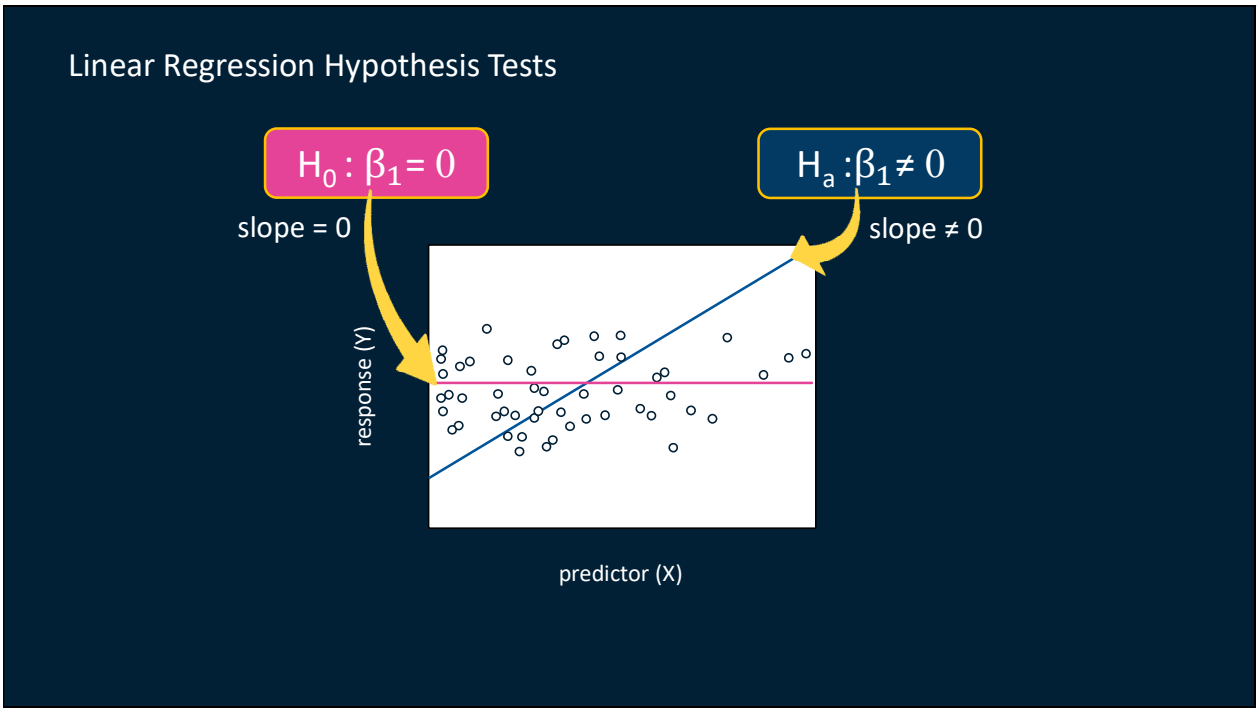
A commonly used approach to estimating the regression line is to use ordinary least squares. The process involves the residuals, the difference between predicted values and actual values of the response variable Y. Residuals are estimates of the population errors and thus can be observed, whereas population 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.



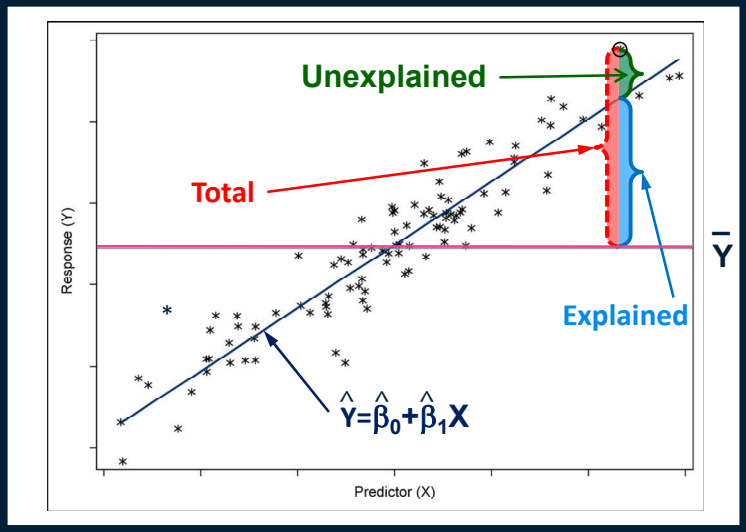
The hypothesis test for a simple linear regression model is a test of whether the slope (β_1) is significantly different from zero. A zero slope for a regression corresponds to a horizontal line where Y equals the mean value of Y. A flat regression line indicates that the mean value of Y does not depend 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 statistically different from zero might still be too small to be meaningful to the researcher.

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.

Coefficient of Determination

How well is my regression model fitting the data?

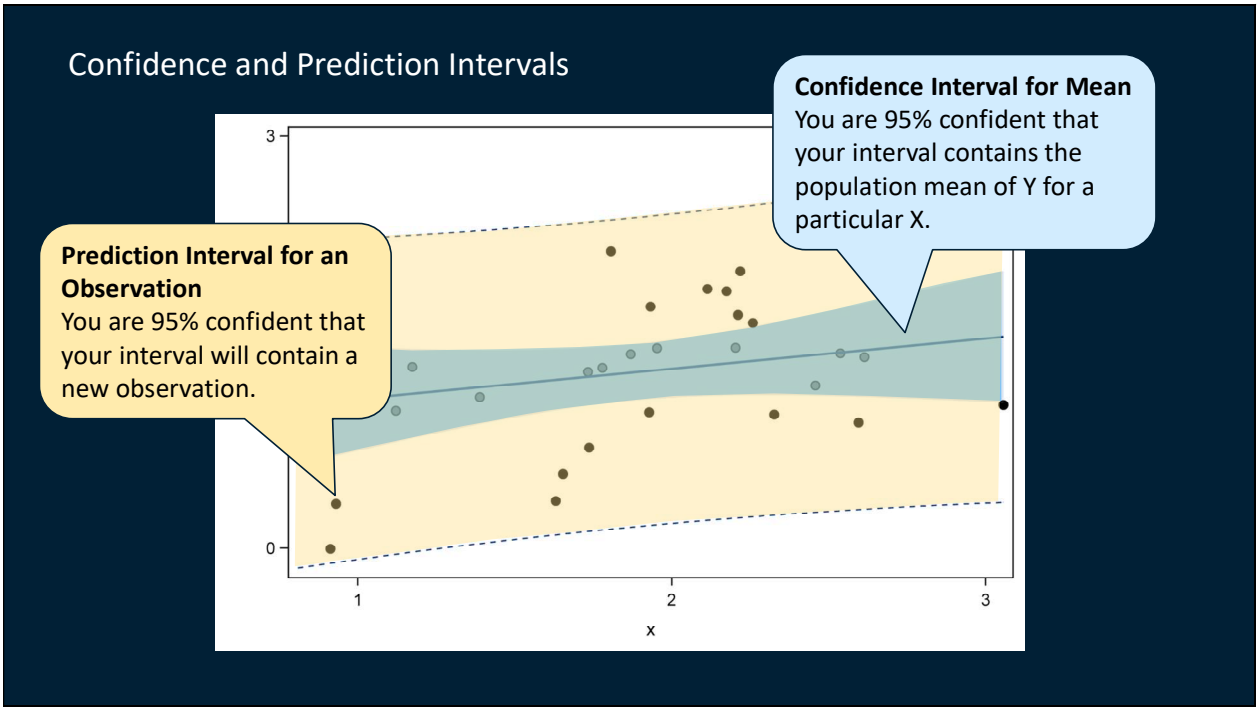
$$R^2 = \frac{Explained\ variablitiy}{Total\ variability}$$

"the proportion of the variability in Y that can be explained by the model"

$$0 \leq R^2 \leq 1$$

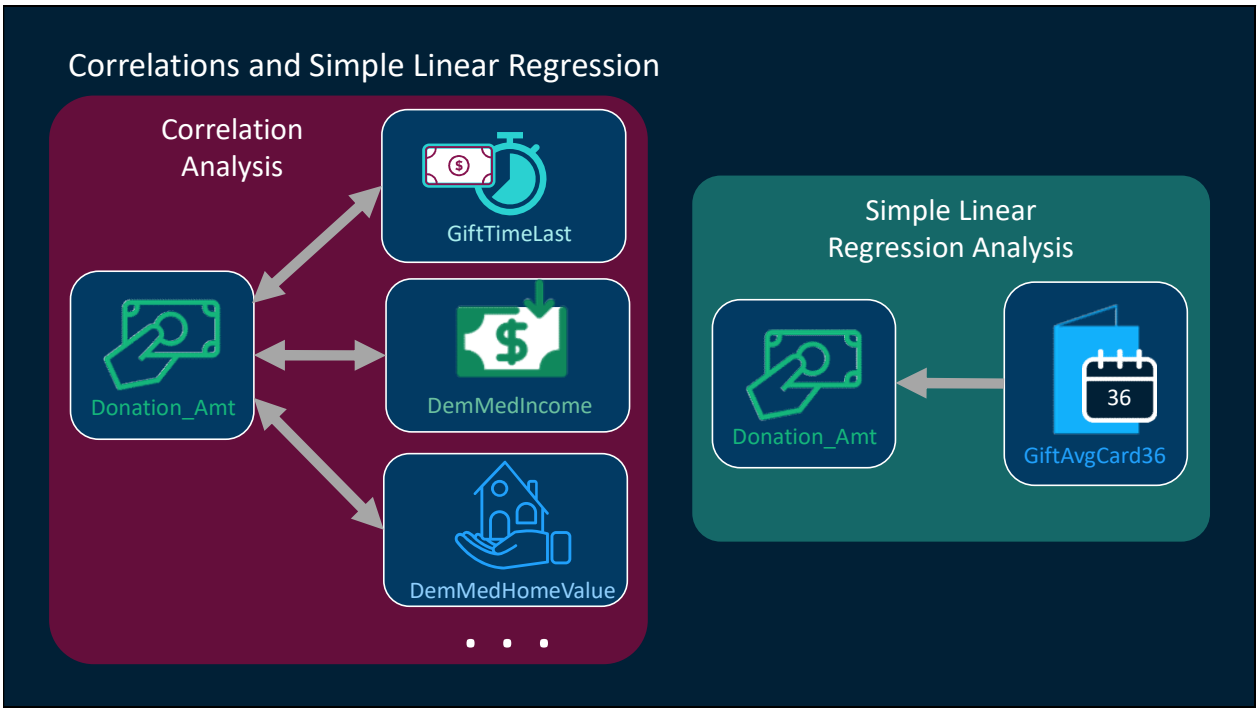
The *coefficient of determination*, R-square, is a measure of the proportion of variability in the response variable explained by predictor variables in the model.

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.



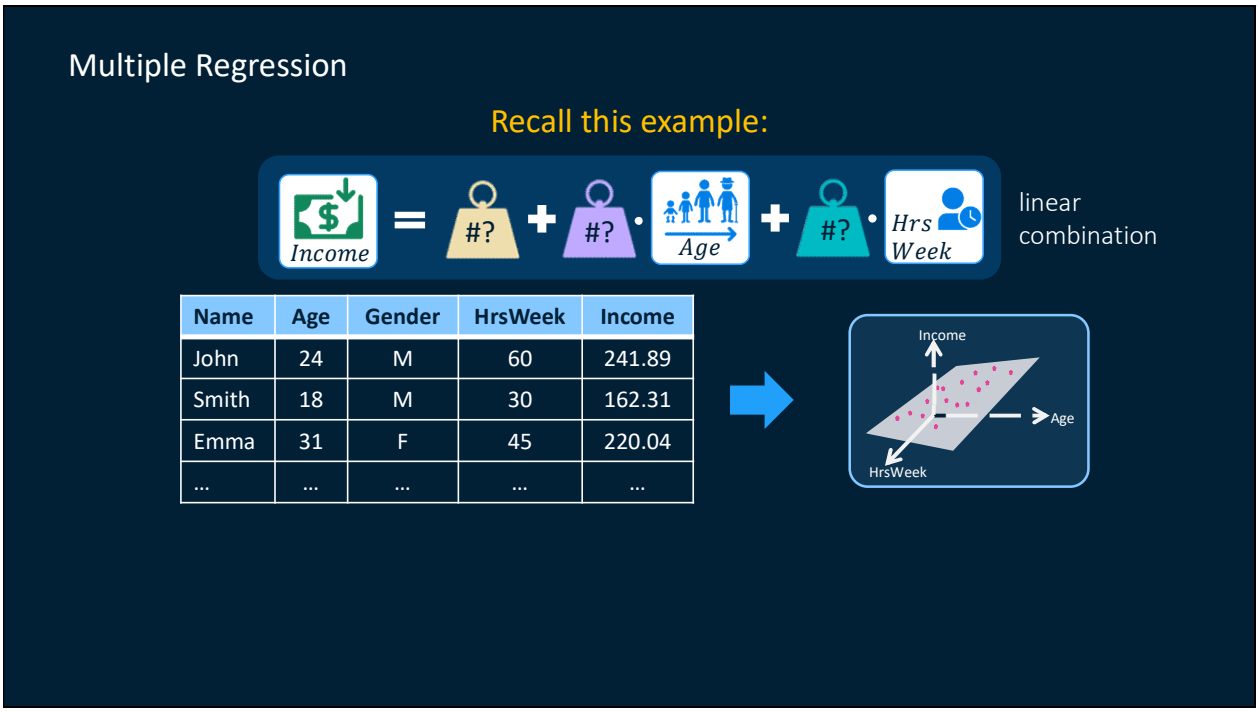
Confidence intervals and prediction intervals can be constructed around the estimated regression line. Often these are 95% confidence limits or 95% prediction limits, but other values can also be used. The confidence interval for the mean enables you to say with a given level of confidence where you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



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**.

3.2 Multiple Regression and Model Selection



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

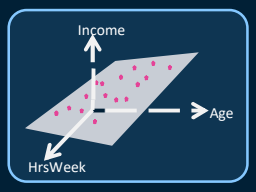
Recall an example we considered in lesson 1, a simple data set containing information on age, gender, working hours per week, and income. Where you have two predictors, the relationship between predictors and the response can be modeled with a plane. Here we see a multiple regression model with the dependent 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:

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

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



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

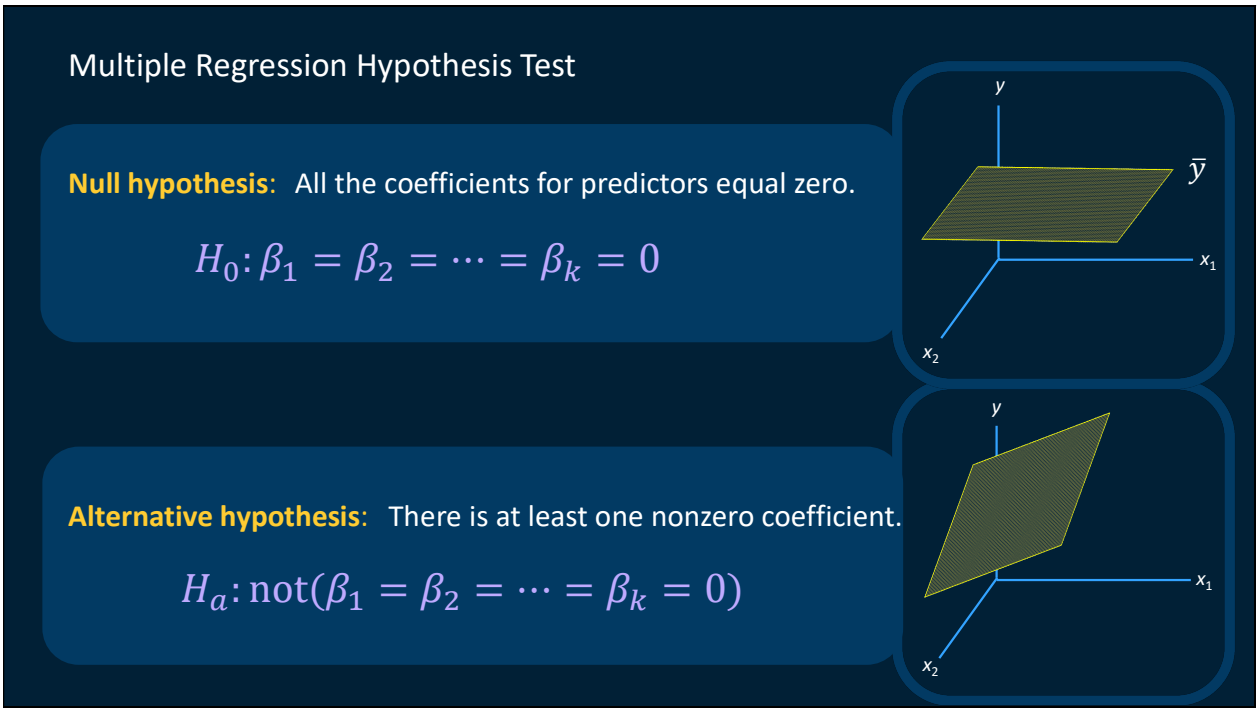
$\hat{\beta}_2$ = effect of **HrsWeek** on **Income** while holding **Age** constant

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.



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.

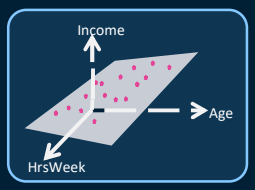
If the null hypothesis of all βs = 0 is rejected, then the best fit regression plane has at least one nonzero slope, and the best estimate of Y depends on the values of the X's.

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
…	…	…	…	…

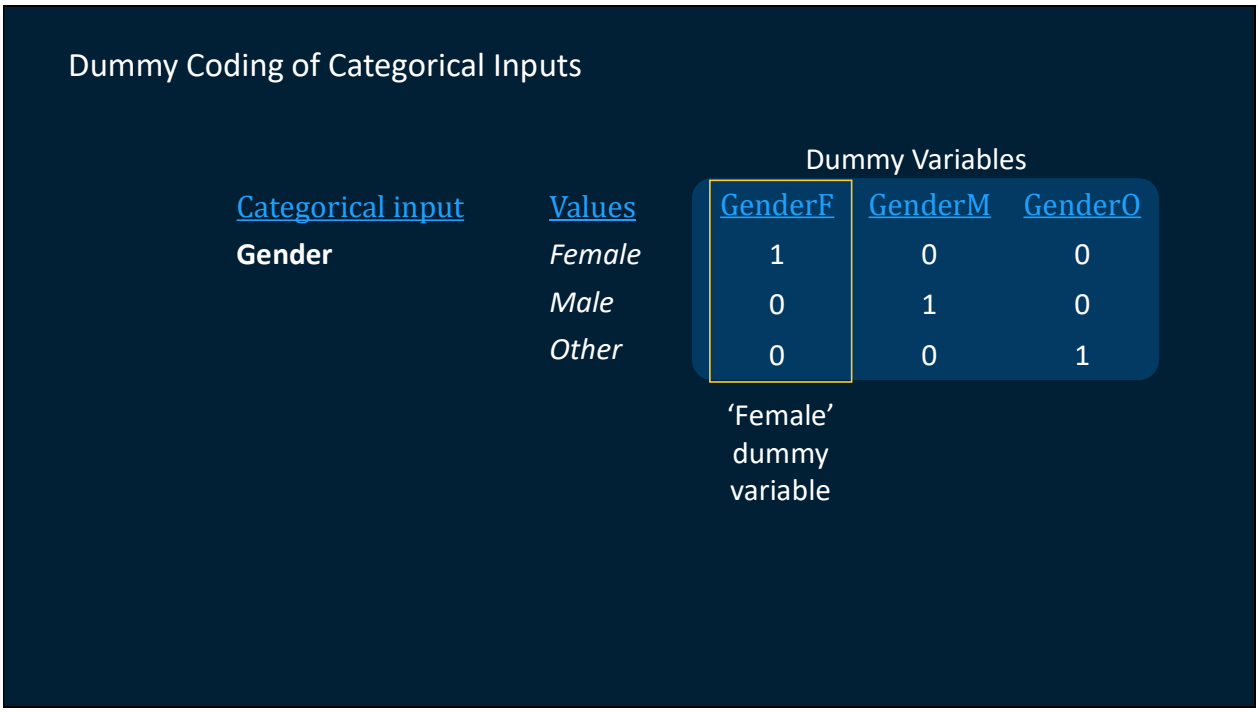
How are these converted to a numeric variable?



a regression model

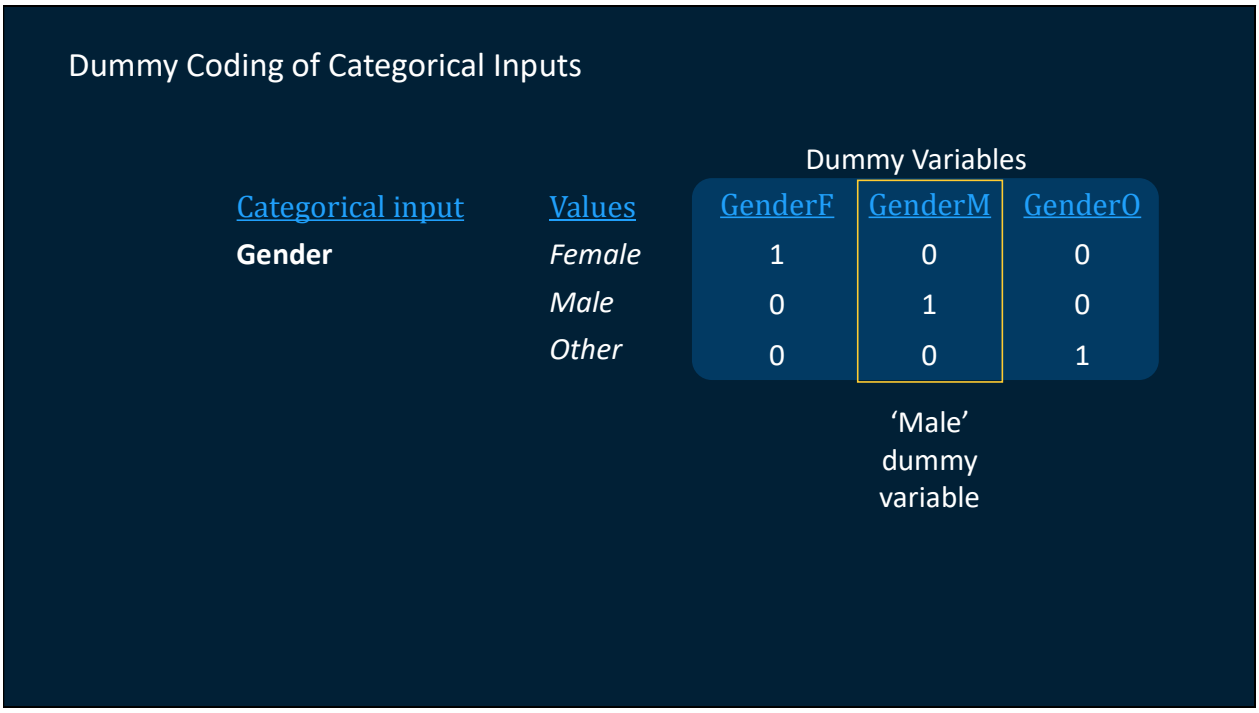
The term *regression* originally referred only to modeling continuous predictors, but categorical predictors can be used as well. These predictors need to be converted into numeric variables that represent the different levels.

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

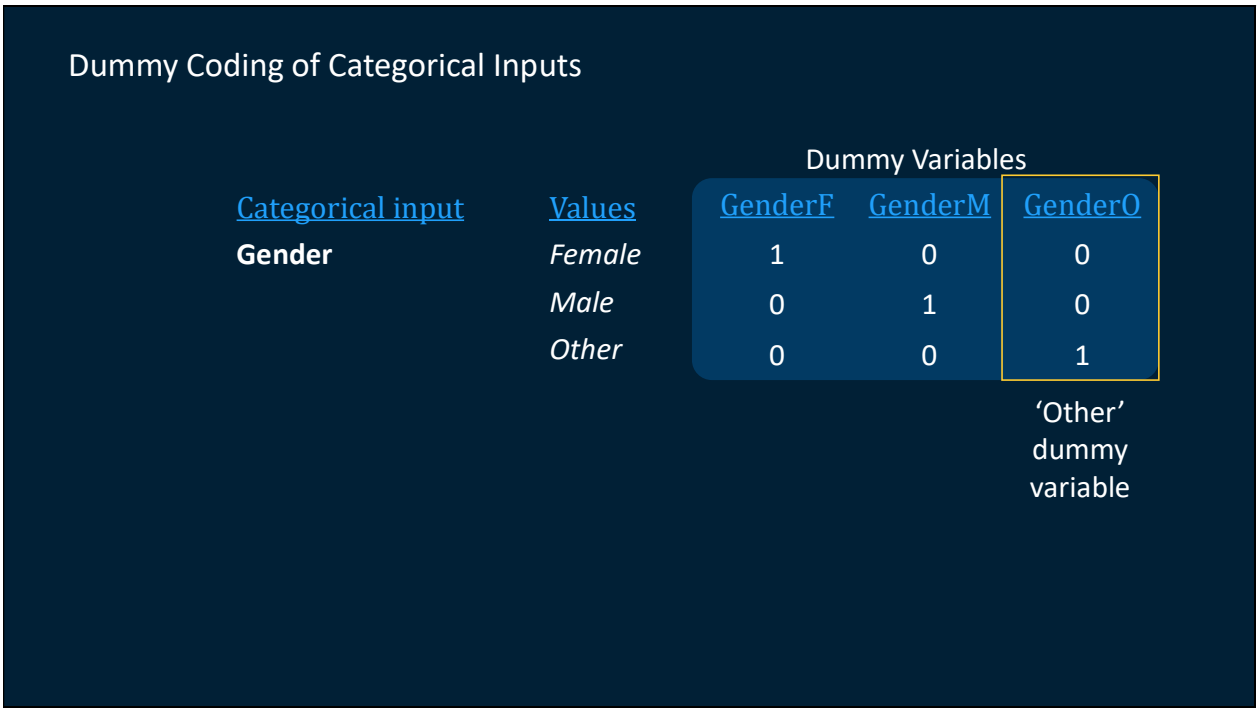


Dummy variables (sometimes called *design variables* or *indicator variables*) are constructed numeric variables that represent the levels of a categorical predictor. These can be created in variety of ways, and one of the most common ways or parameterizations is shown here.

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.

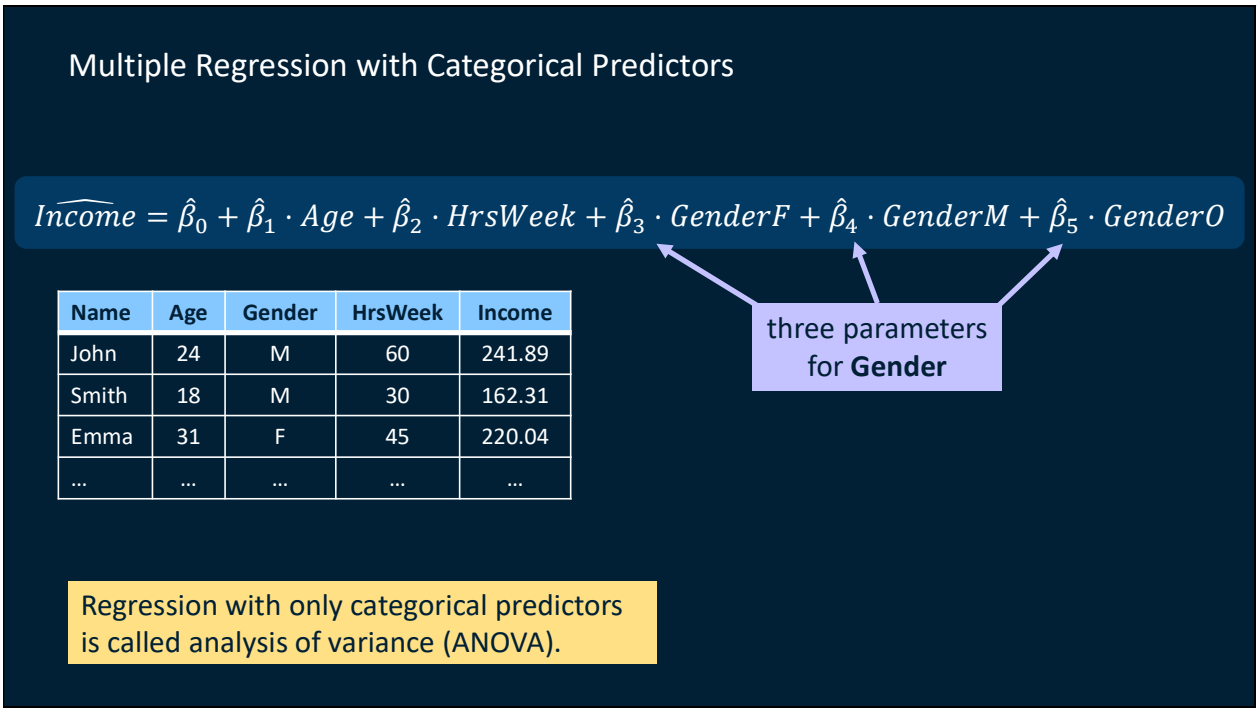


In a similar manner, a dummy variable for males takes the value of 1 for people in the *Male* category and 0 for everyone else.



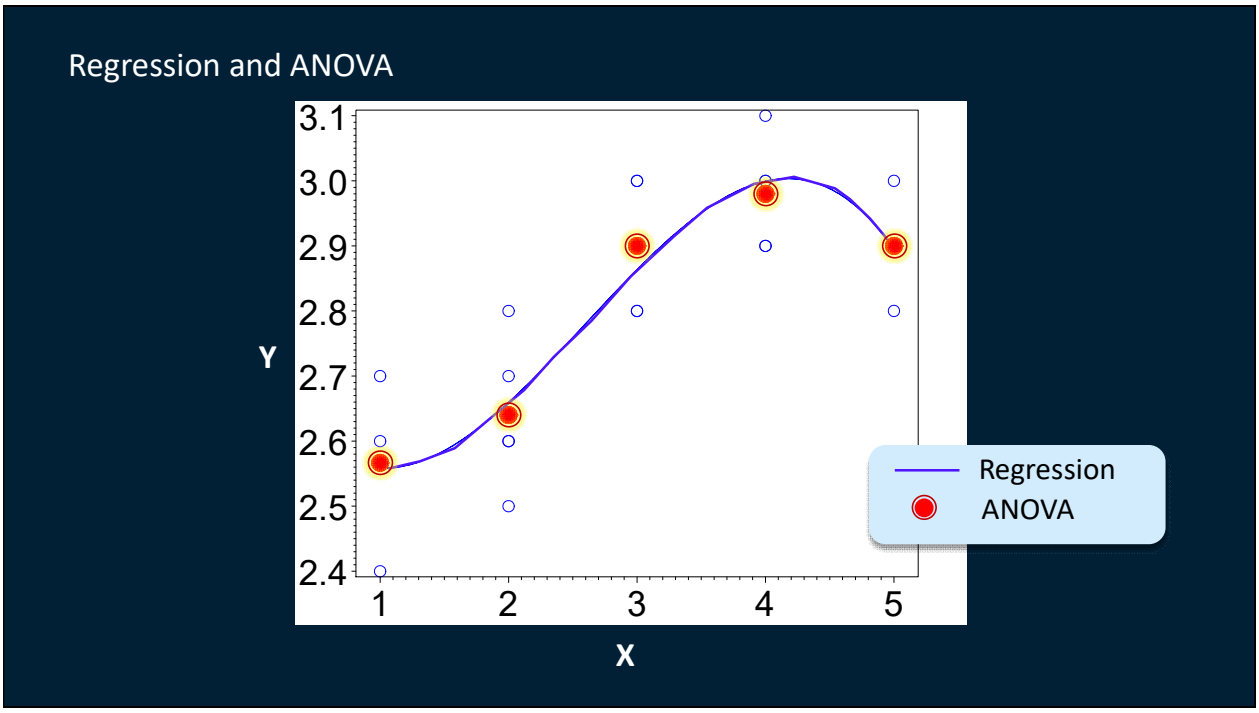
Finally, members of the *Other* category are assigned a value of 1 for the Other dummy variable, and everyone else gets a 0.

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



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

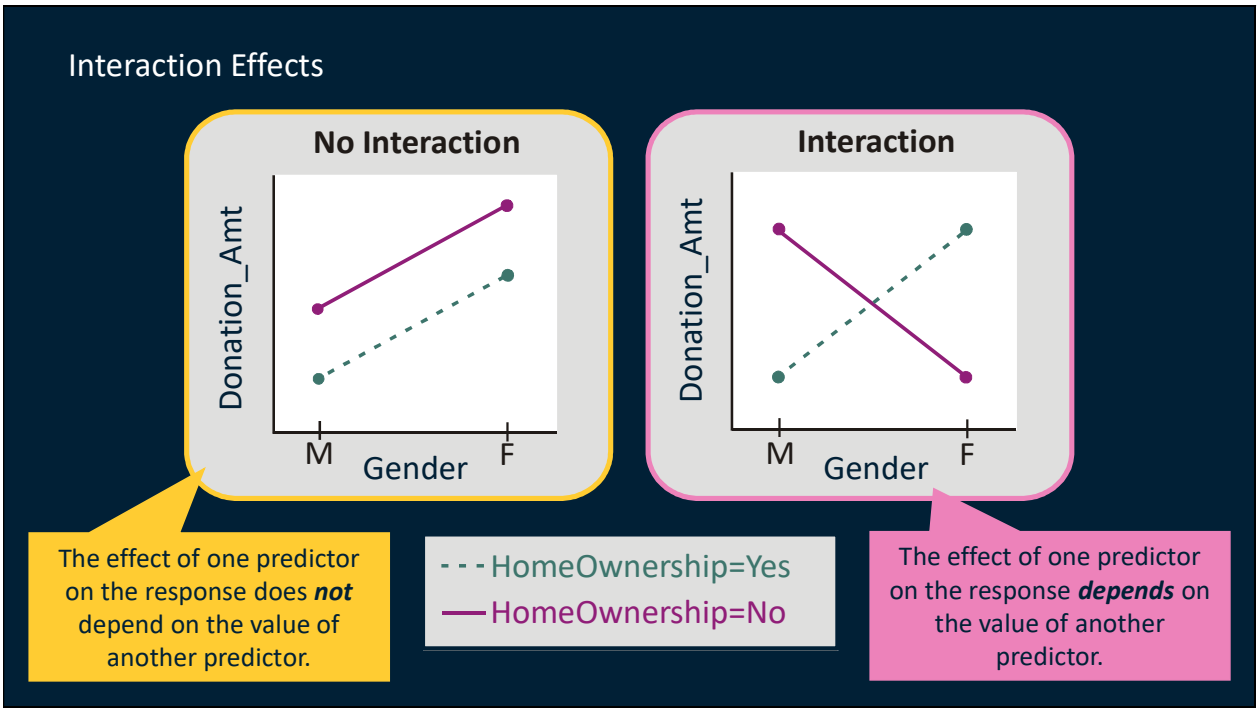
It is these three dummy variables and not the original variable **Gender** that would be used directly in a multiple regression model. Regression with only categorical predictors is called analysis of variance (or ANOVA).



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.

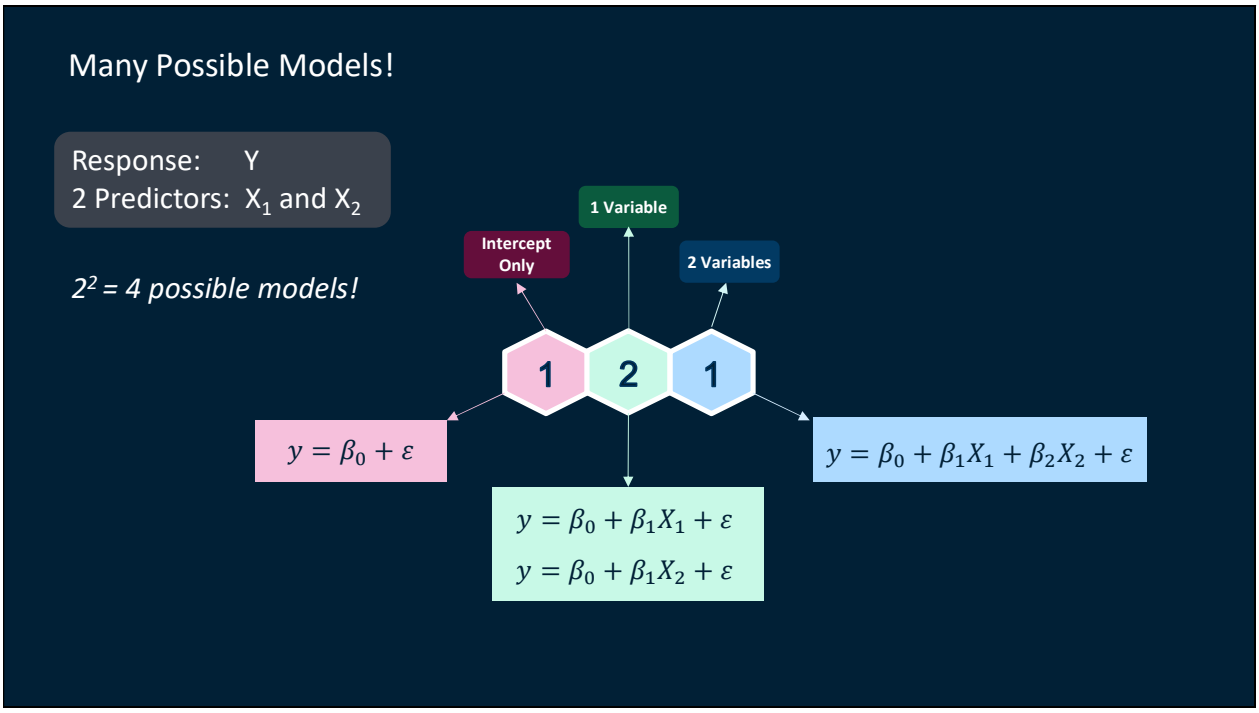


When there are multiple predictors in a regression model, there is the potential for predictors to interact. When interactions exist, modeling them can greatly increase the explanatory power and predictive ability of a model.

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.

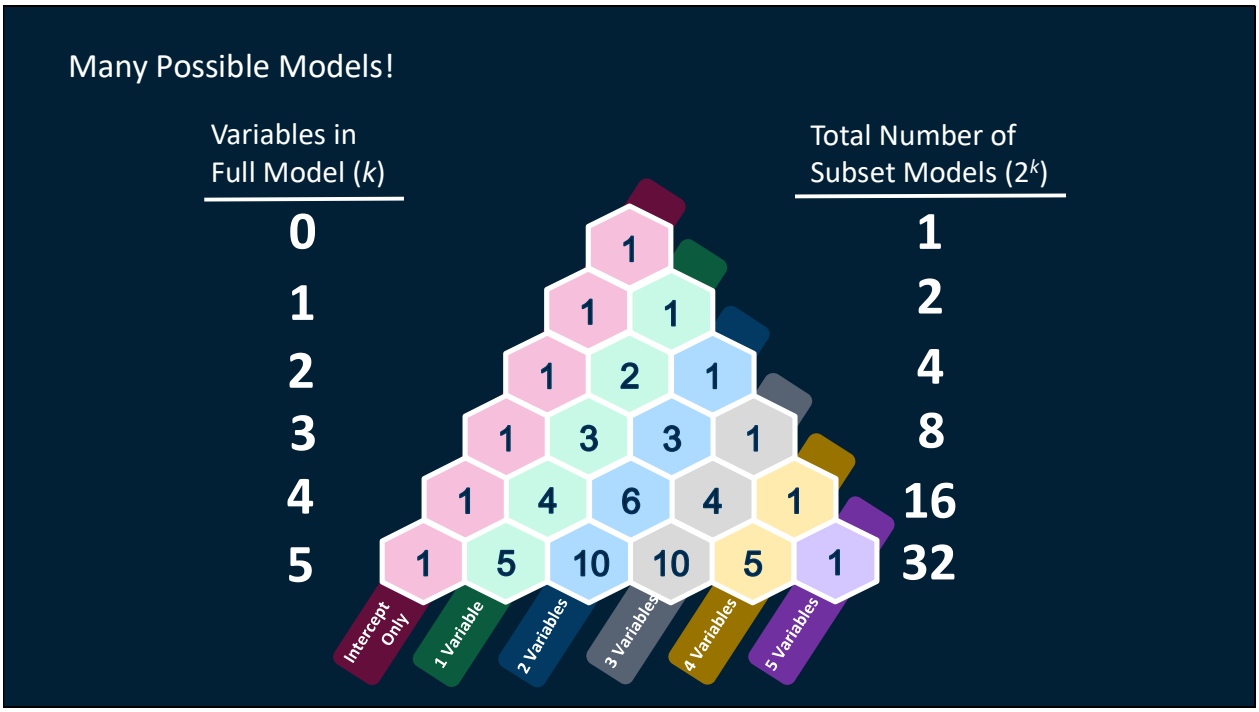
The plots displayed here are called *means plots*. The average **Donation_Amt** values over different levels of **Gender** were plotted and then connected for **HomeOwnership** *Yes* and *No*.

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**.



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.

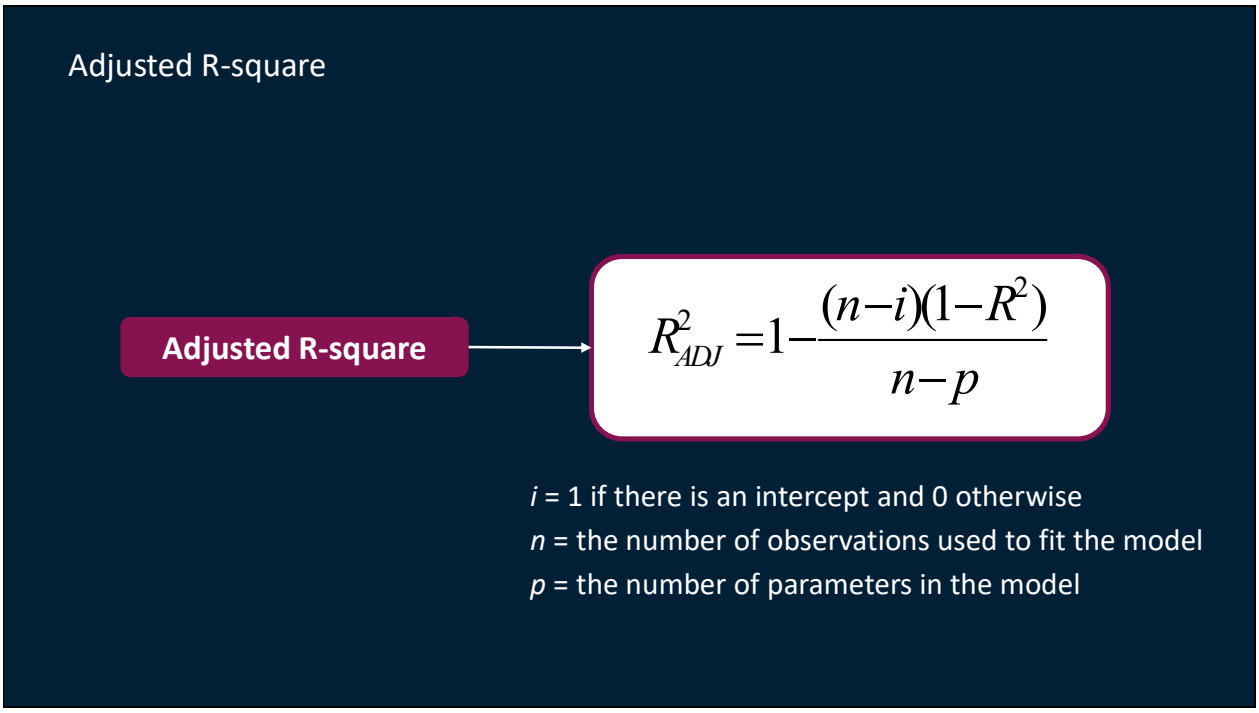


The number of models explodes quickly with increasing inputs and is even larger when considering polynomial and interaction effects.

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.

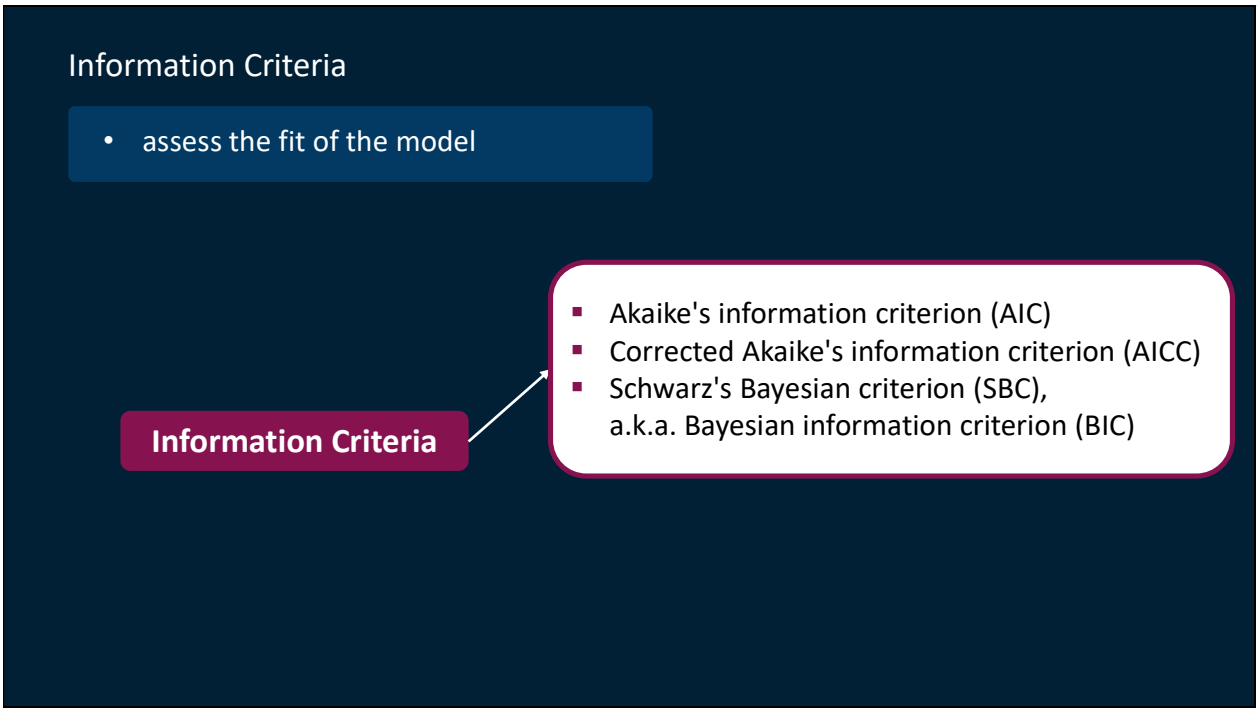


One statistic for comparing multiple regression models is the adjusted R-square. Adjusted R-square is a measure similar to R-square, but it takes into account the number of parameters in the model. It can be thought of as a penalized version of R-square, with the penalty increasing with each parameter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

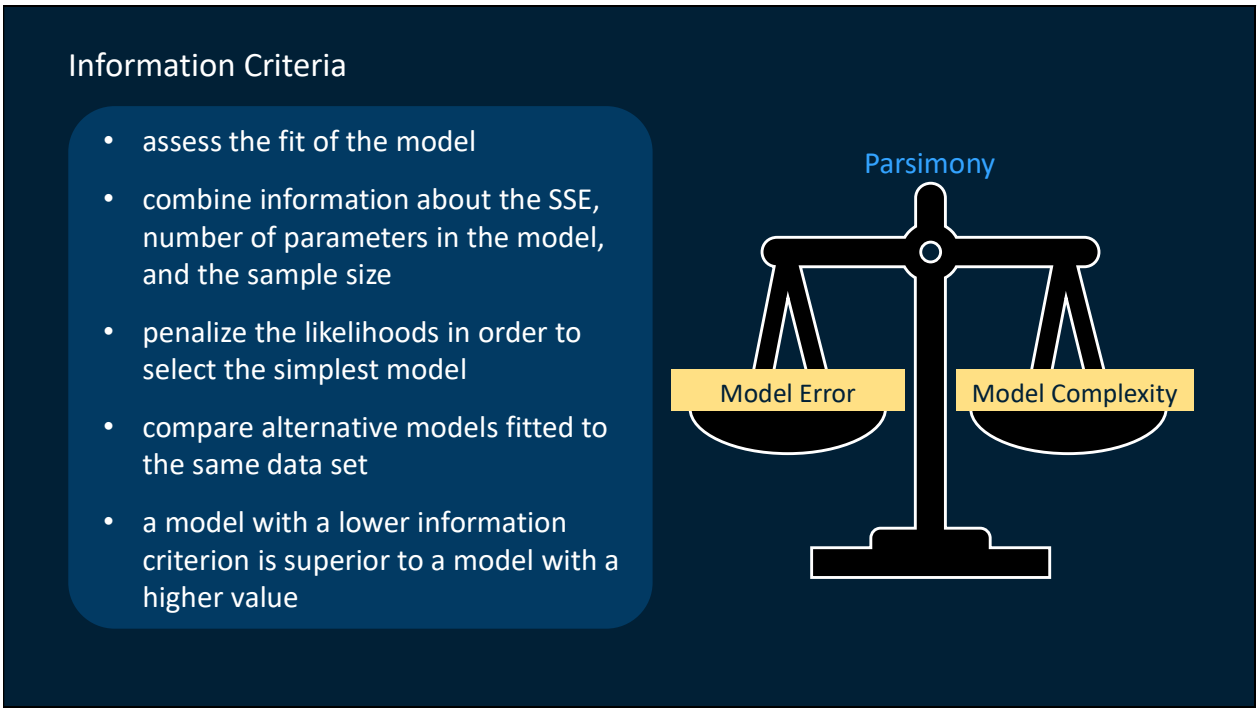
Why is this modified version of R-square useful? When comparing models with different numbers of effects, R-square is not an ideal choice. R-square increases as you include more terms in the regression model. A model with 10 predictors might have a higher R-square than a model with two predictors just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



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.

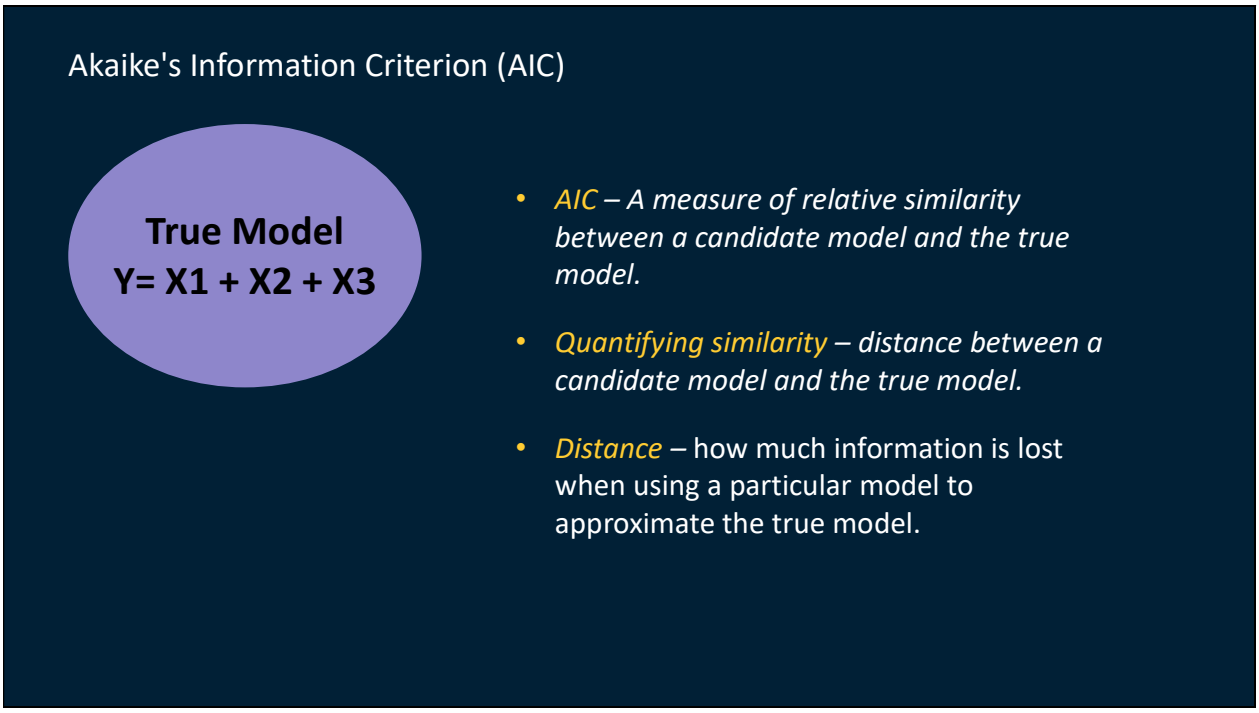
Information criteria penalize the likelihoods in order to select the simplest model. These criteria are based upon concepts of information and entropy.

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.

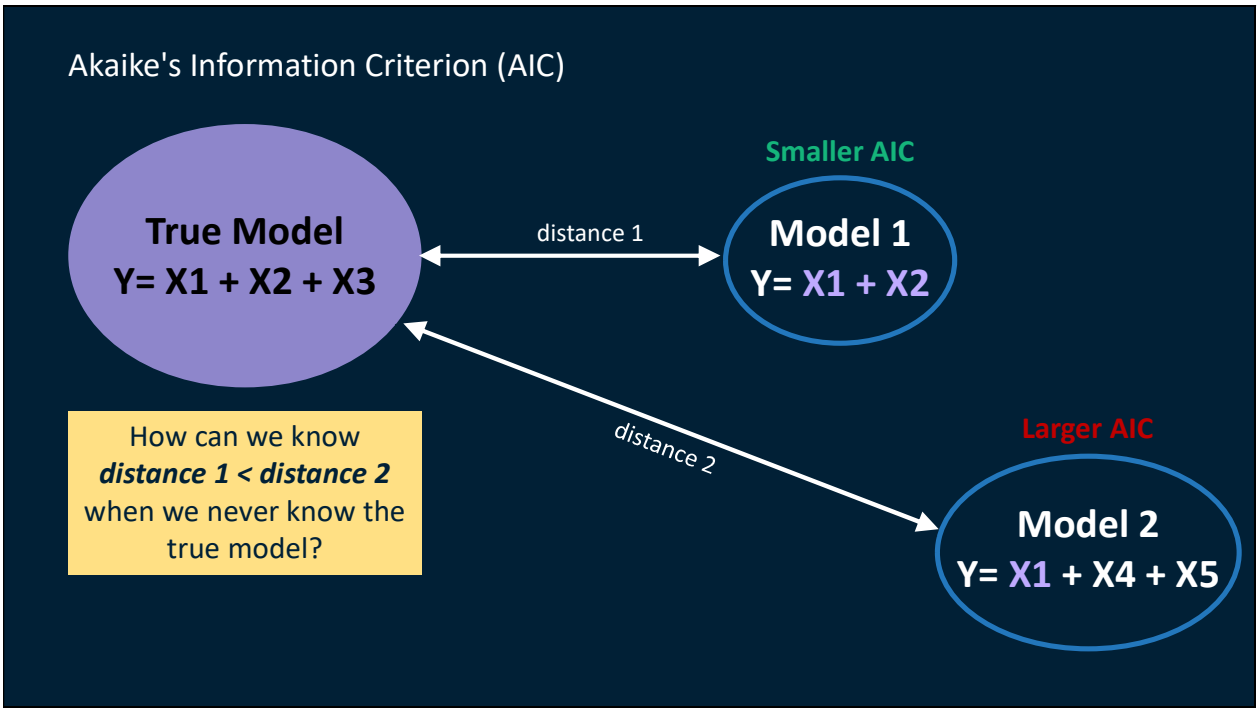
And all else being equal, a model with a lower information criterion is superior to a model with a higher value.

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



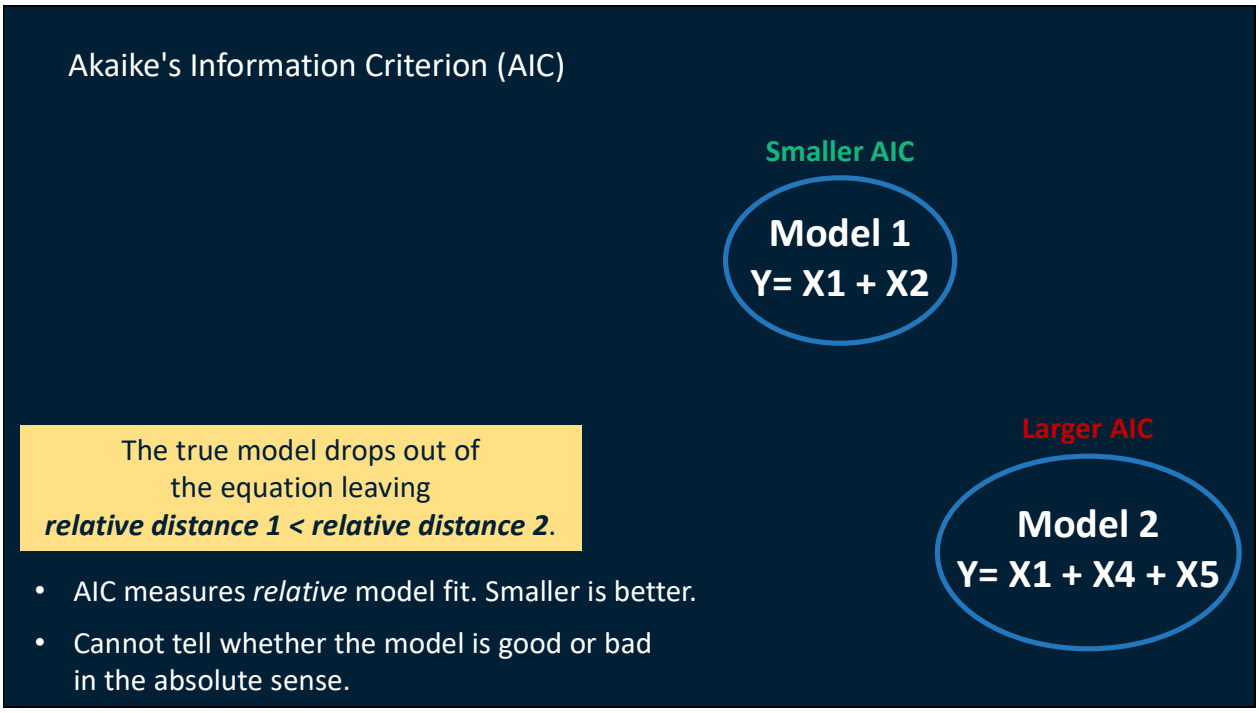
Understanding Akaike's information criterion requires reference to a true model. "True model" is actually an abstract concept. All models are simplifications by definition, and thus there is no such thing as a true model. So "true model" is really short for "best approximating model," but it is helpful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

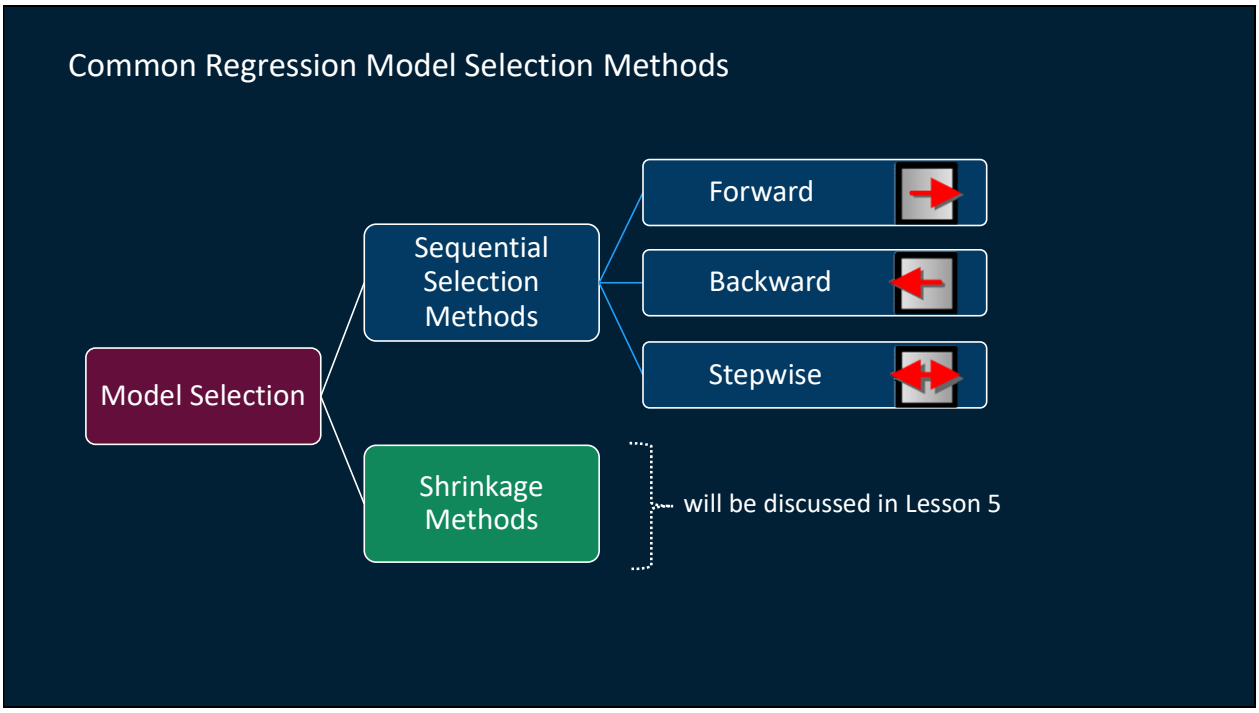


Akaike showed that it is possible to calculate the relative distances to the true model, despite not knowing the population parameters. This is possible because the true model drops out of the equation that compares the models, and what is left is the *relative distance* of each model to the true model.

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.

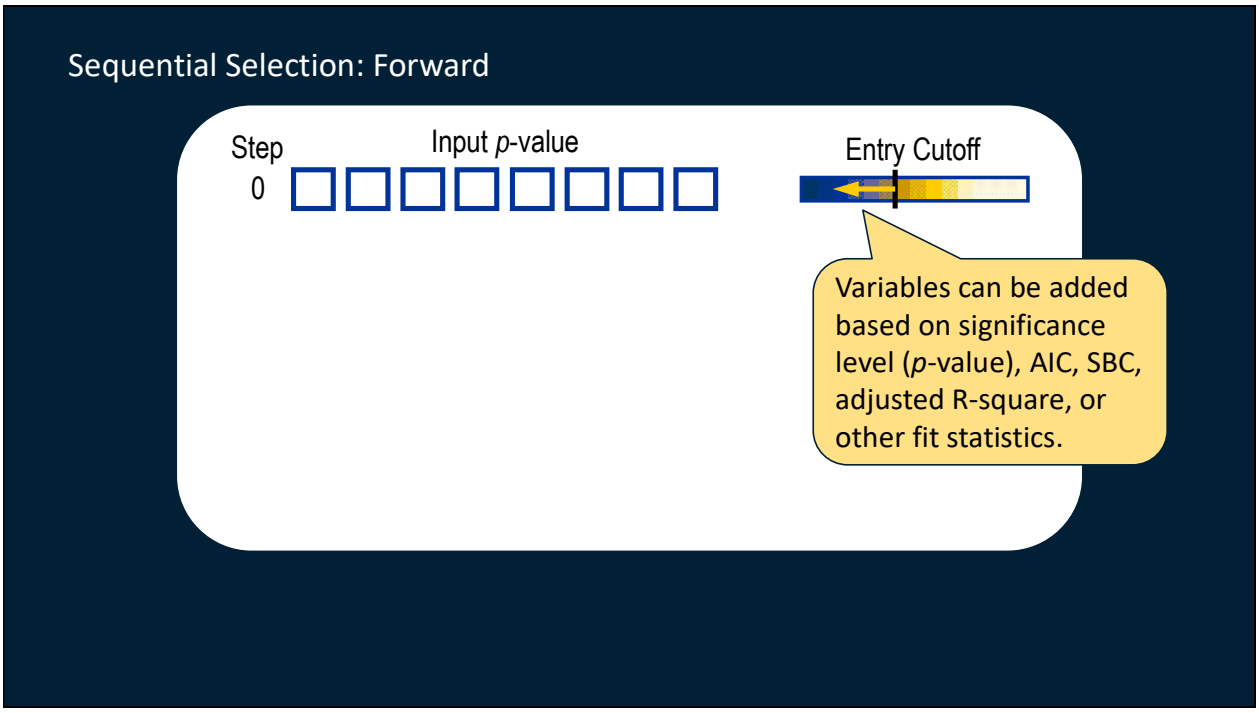
For a readable and detailed explanation of AIC, see Model Selection and Multimodel inference: A Practical Information-Theoretic Approach by Burnham and Anderson.



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?

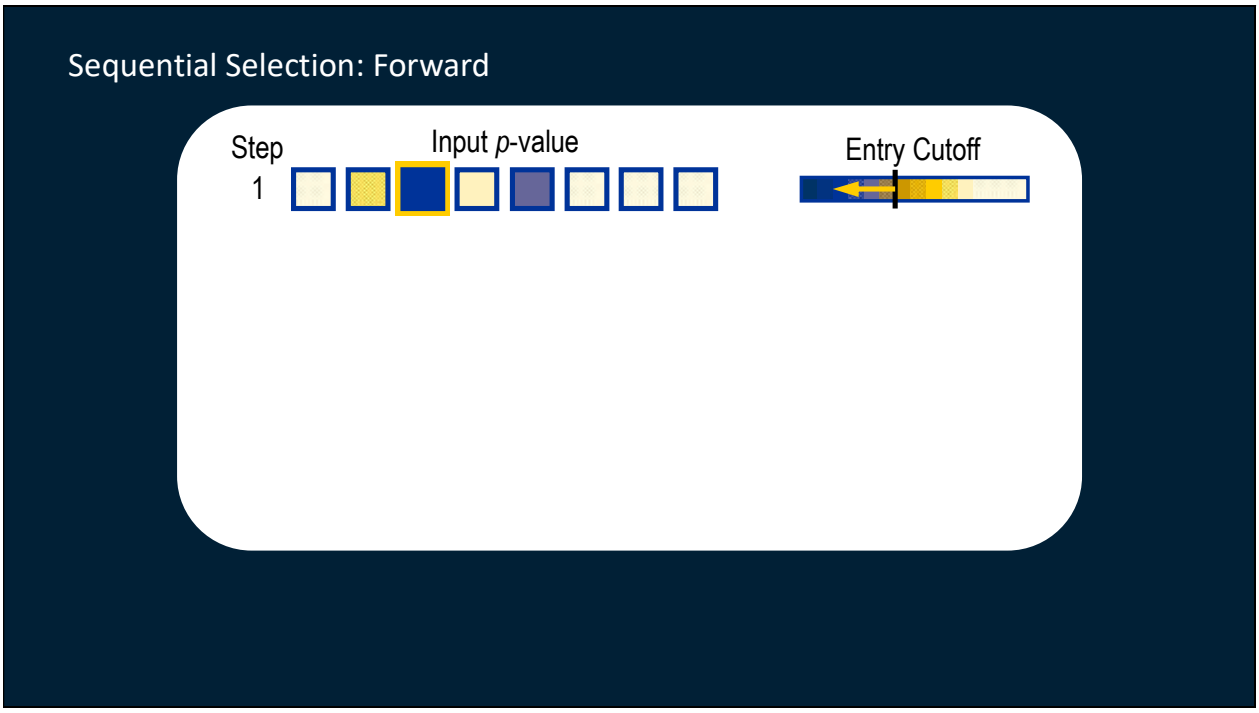
With many predictors that could potentially be included in a model, trying all possible models is often not feasible. We need a computationally efficient way of searching through possible models to find a subset that will fit the data well. Many of these model selection methods exist for regression models. 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.

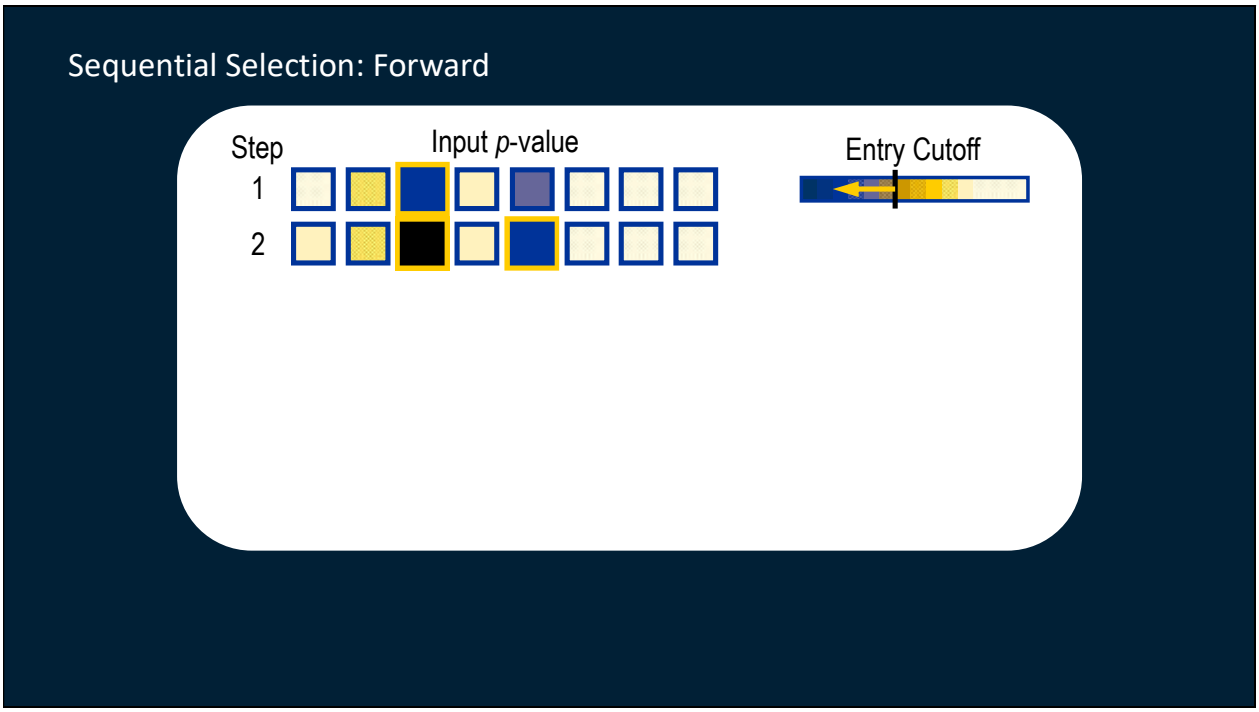


Let's look more closely at some of the traditional model selection methods. Forward selection creates a sequence of models of increasing complexity. The sequence starts with the baseline model, an intercept-only model predicting the overall average target value for all cases. We'll describe this process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

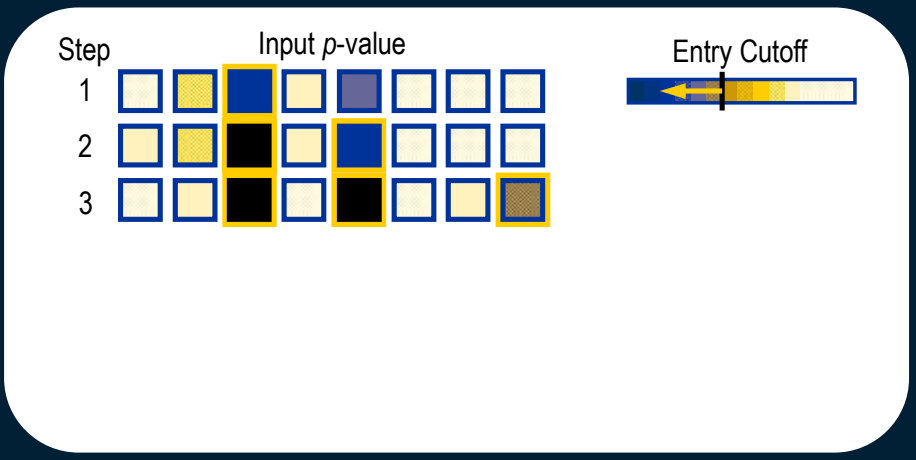
For example, here out of eight potential predictors, the most significant variable will be represented by the darkest blue color, whereas the most insignificant will be represented by the lightest gold color.



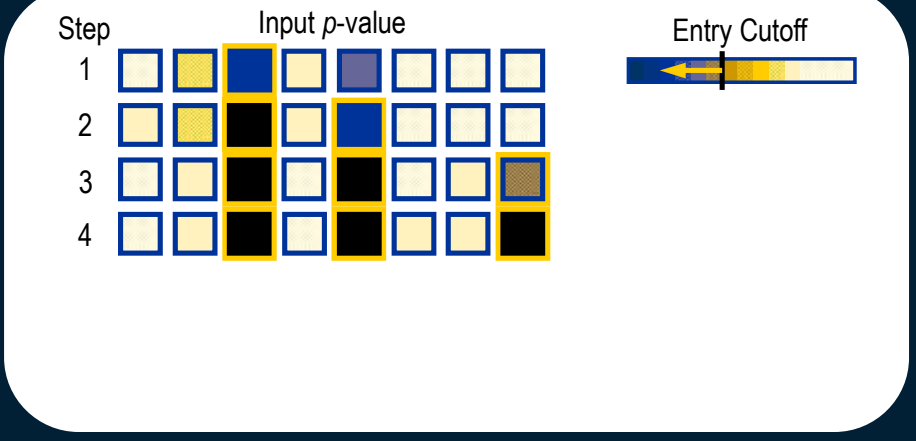
Forward selection adds predictors one at a time based on having the lowest *p*-value, as long as it is under a predetermined threshold value, called the *entry cutoff*. The algorithm searches the set of one-input models and selects the model that most improves on the baseline model.

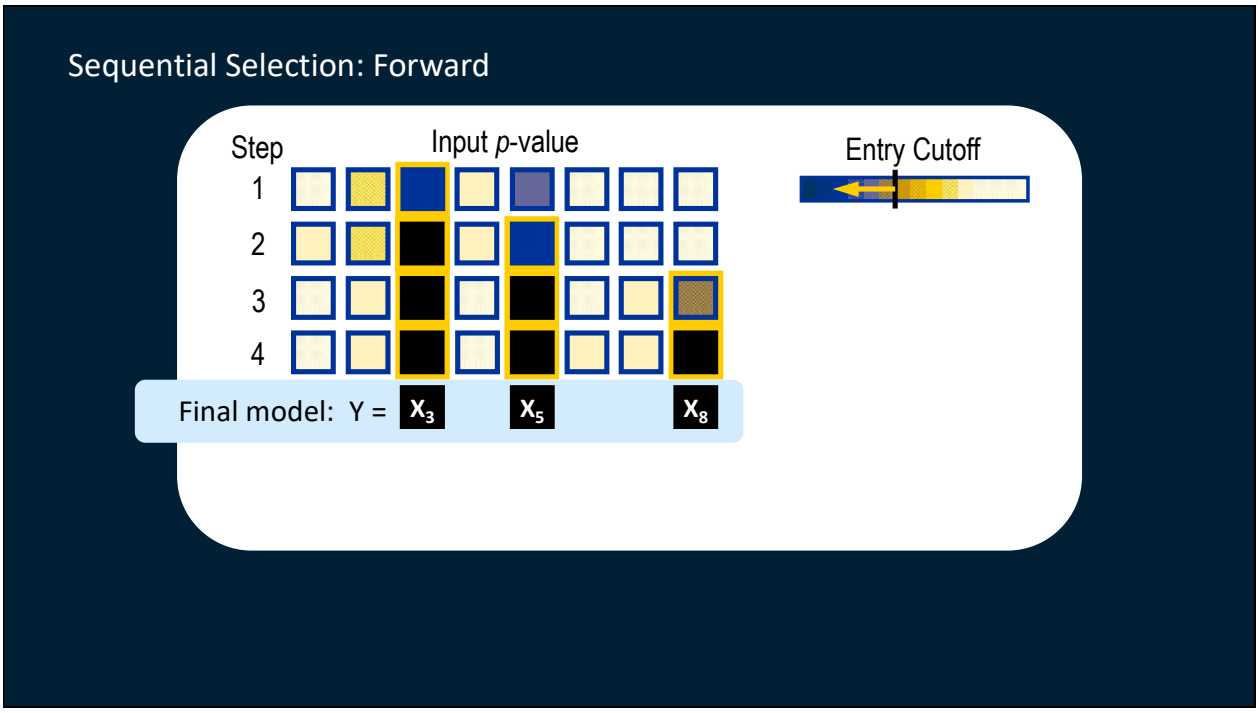


Sequential Selection: Forward

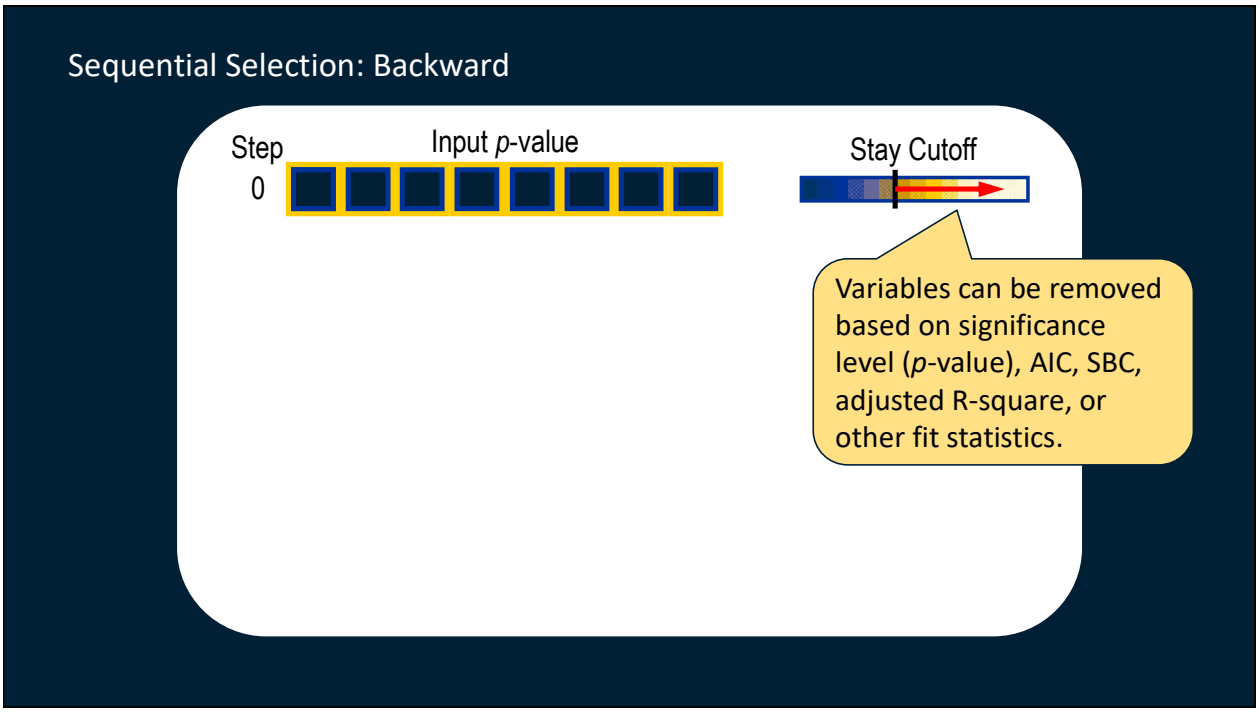


Sequential Selection: Forward

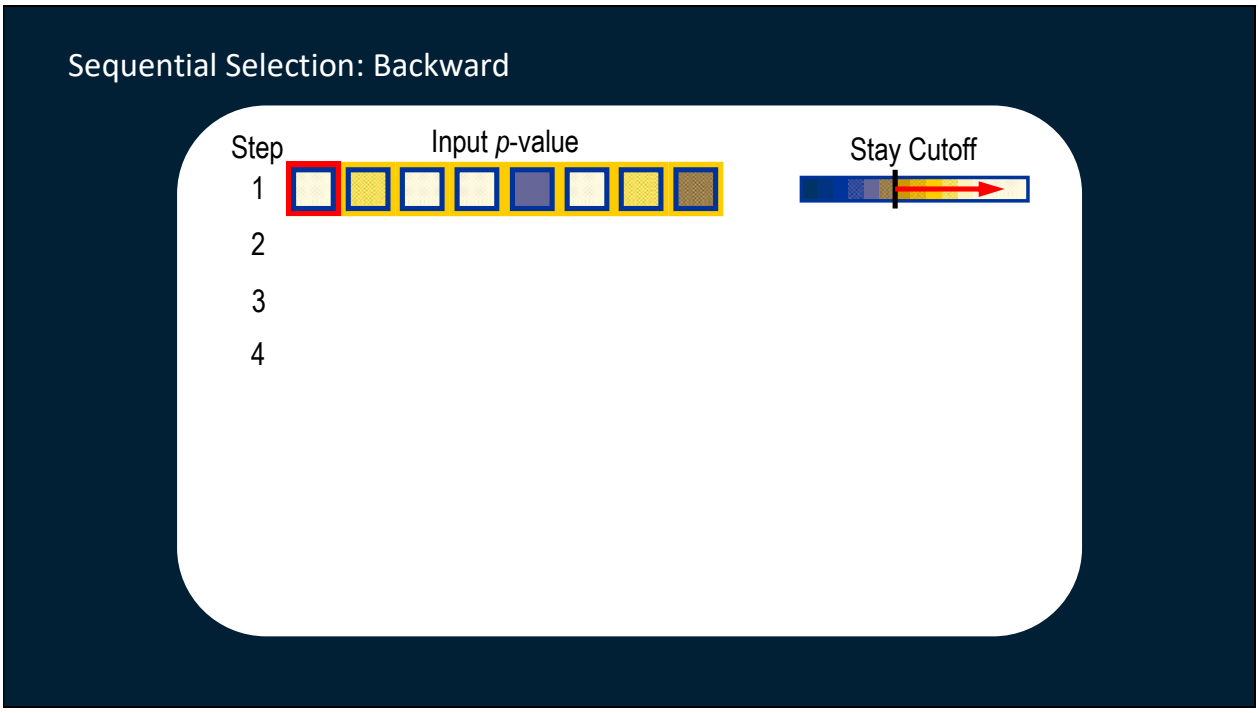




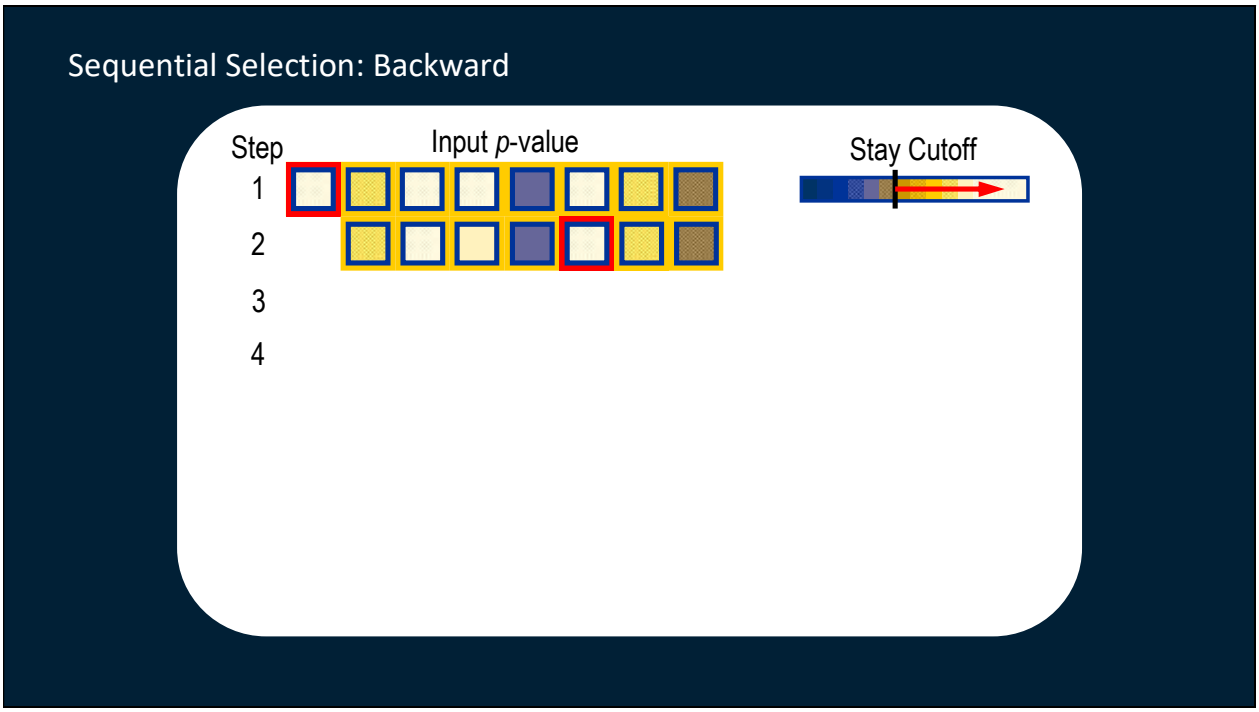
Eventually, all variables outside of the model have *p*-values too high to enter, and the forward selection process stops.



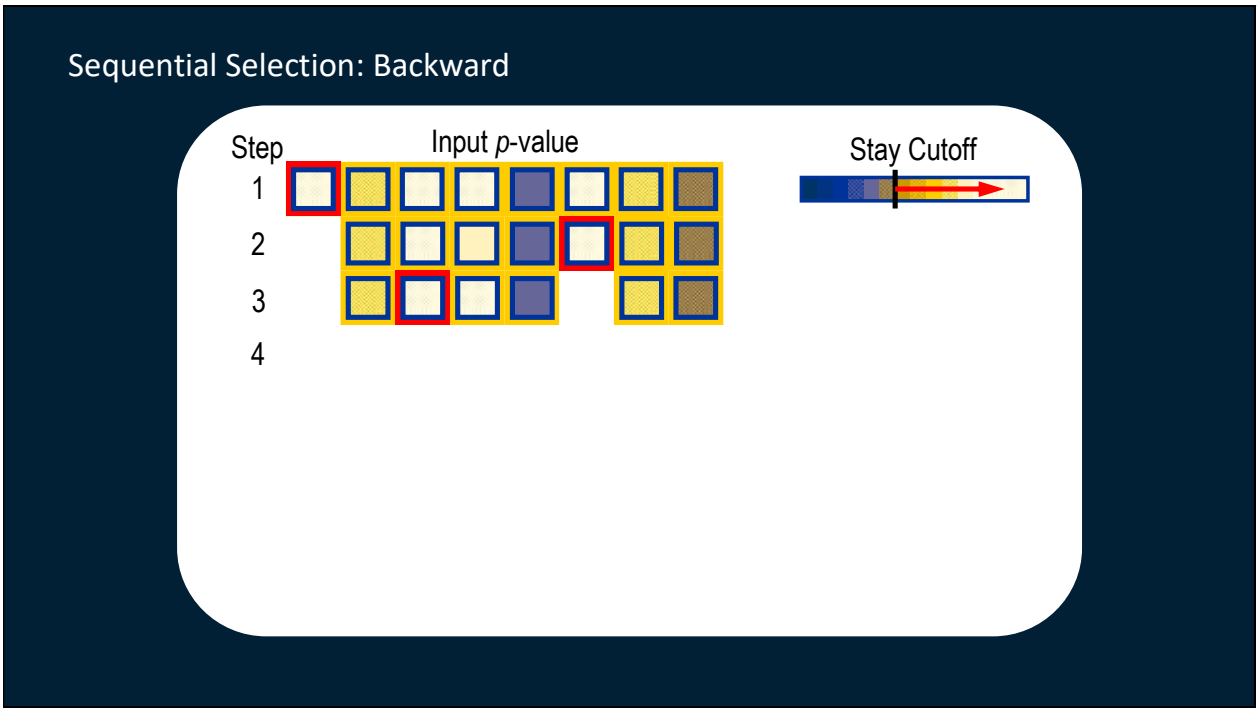
In contrast to forward selection, backward selection creates a sequence of models of decreasing complexity. The sequence starts with a saturated model, which is a model that contains all available inputs, and therefore, has the highest possible fit statistic.

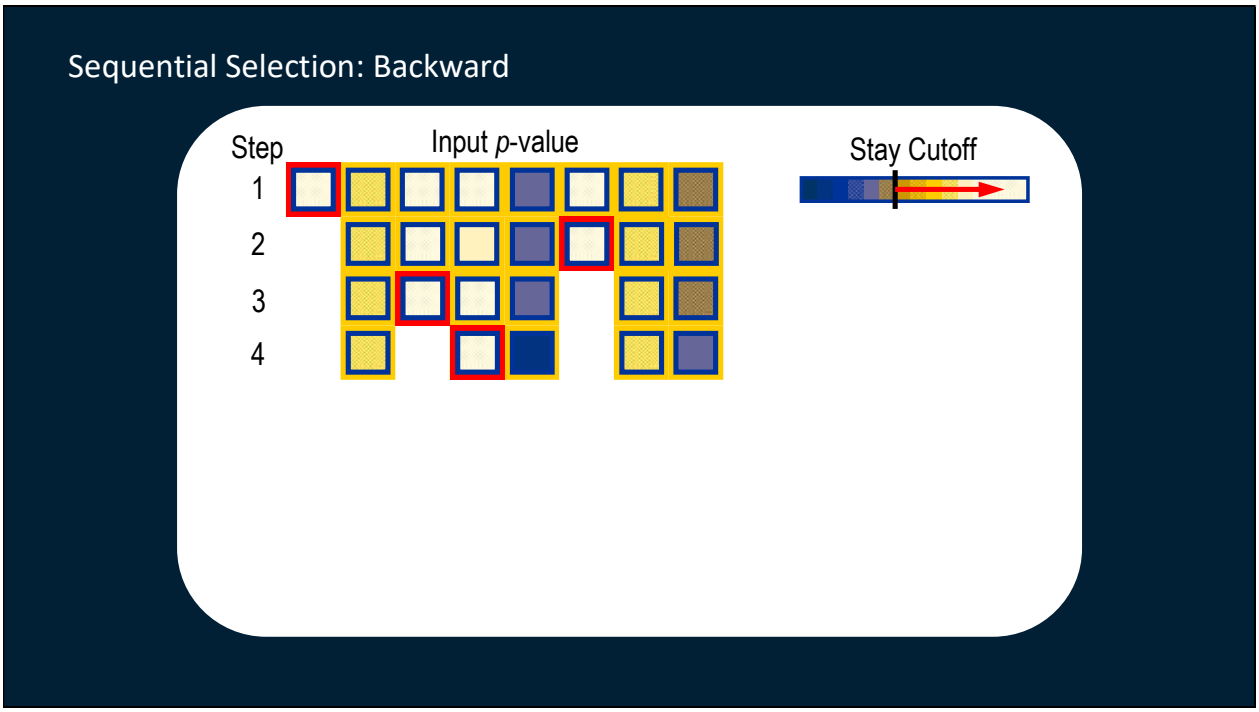


When significance level is the selection criterion, the variable with the highest *p*-value is removed first, as long as the *p*-value is above a predetermined threshold, called the *stay cutoff.*



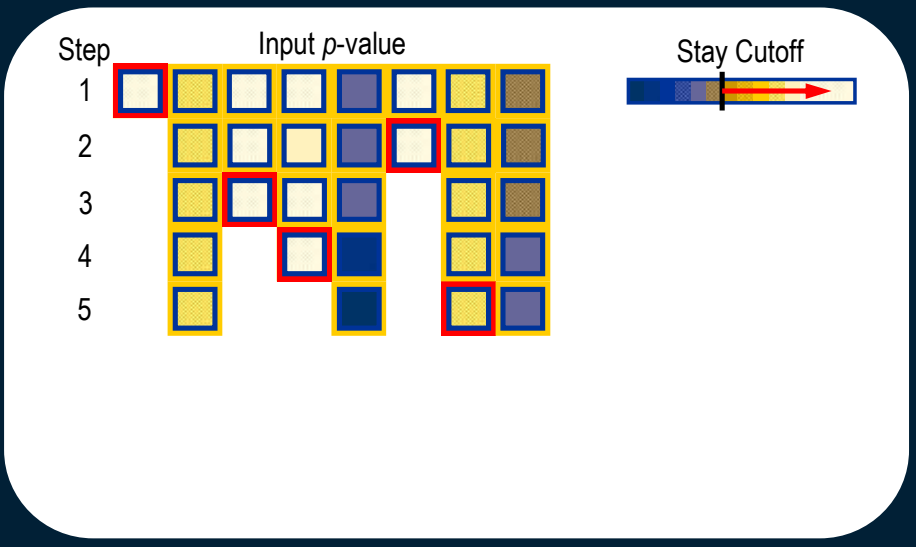
Variables continue to be removed one at a time based on *p*-values. Variables are not allowed to reenter the model.



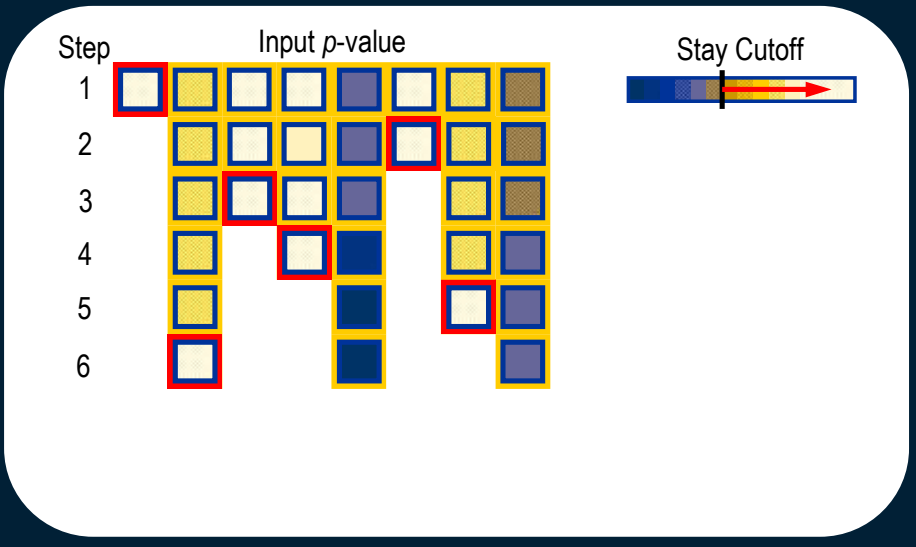


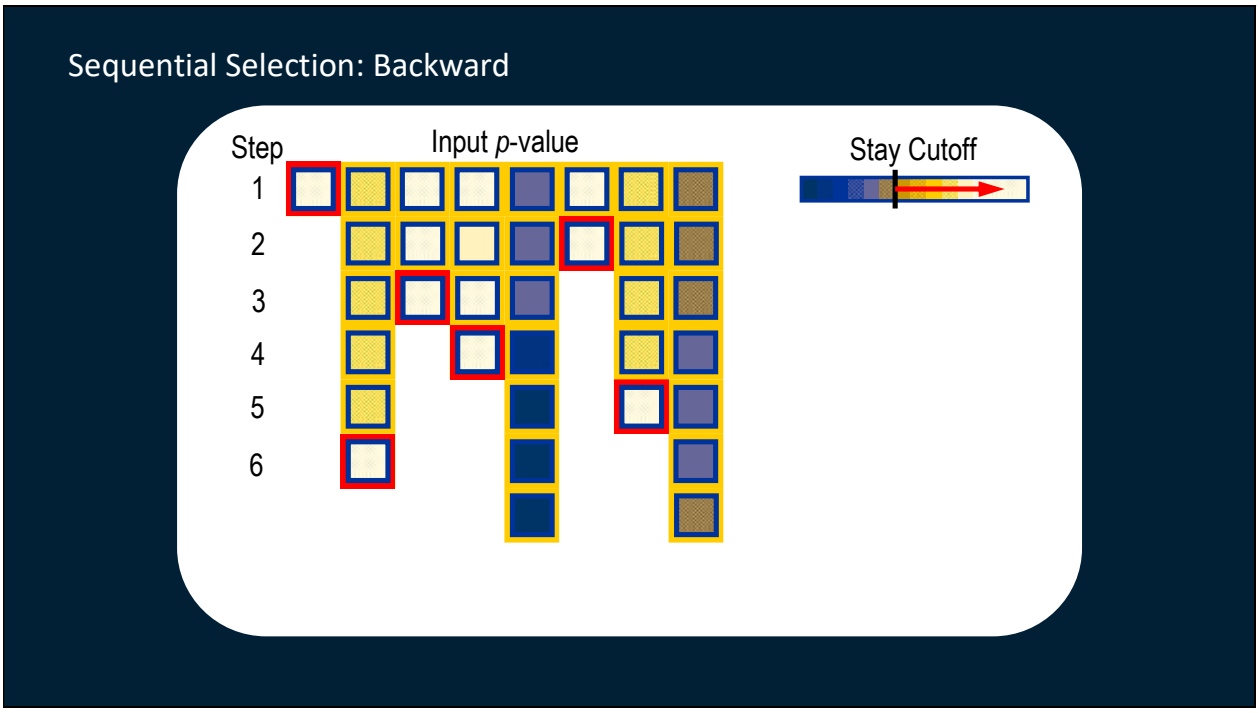
Inputs are sequentially removed from the model. At each step, the input chosen for removal least reduces the overall model fit statistic.

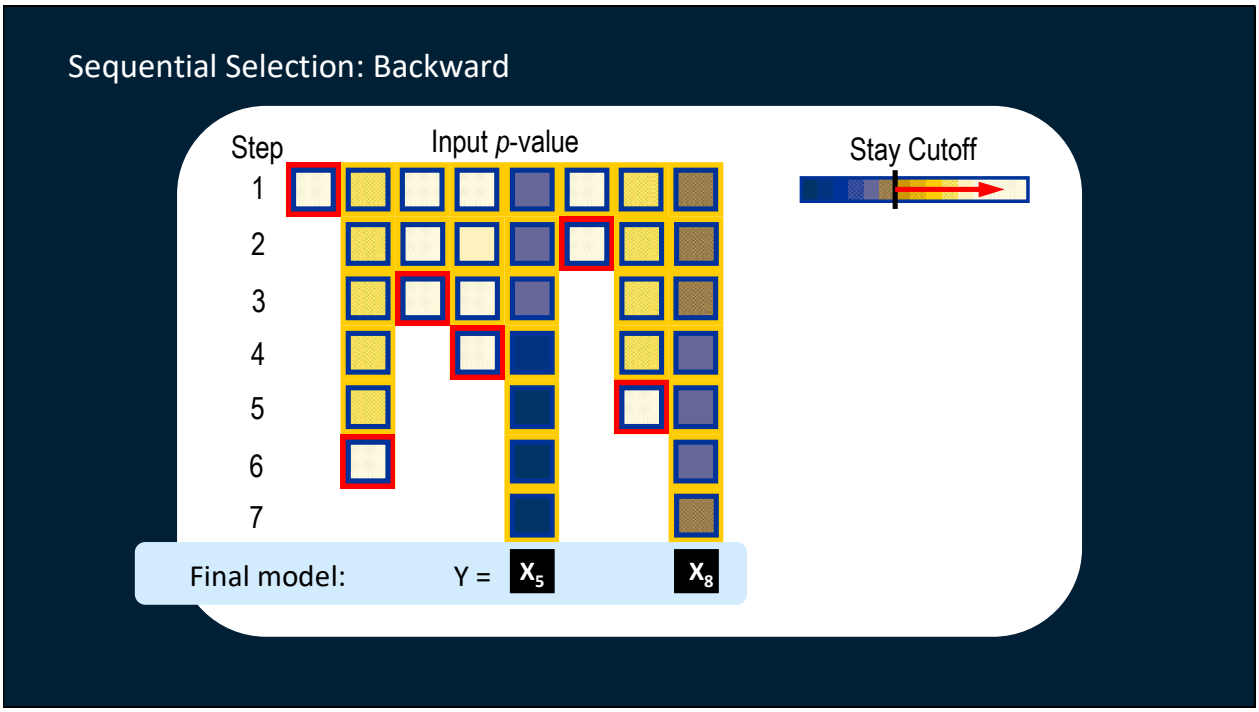
Sequential Selection: Backward



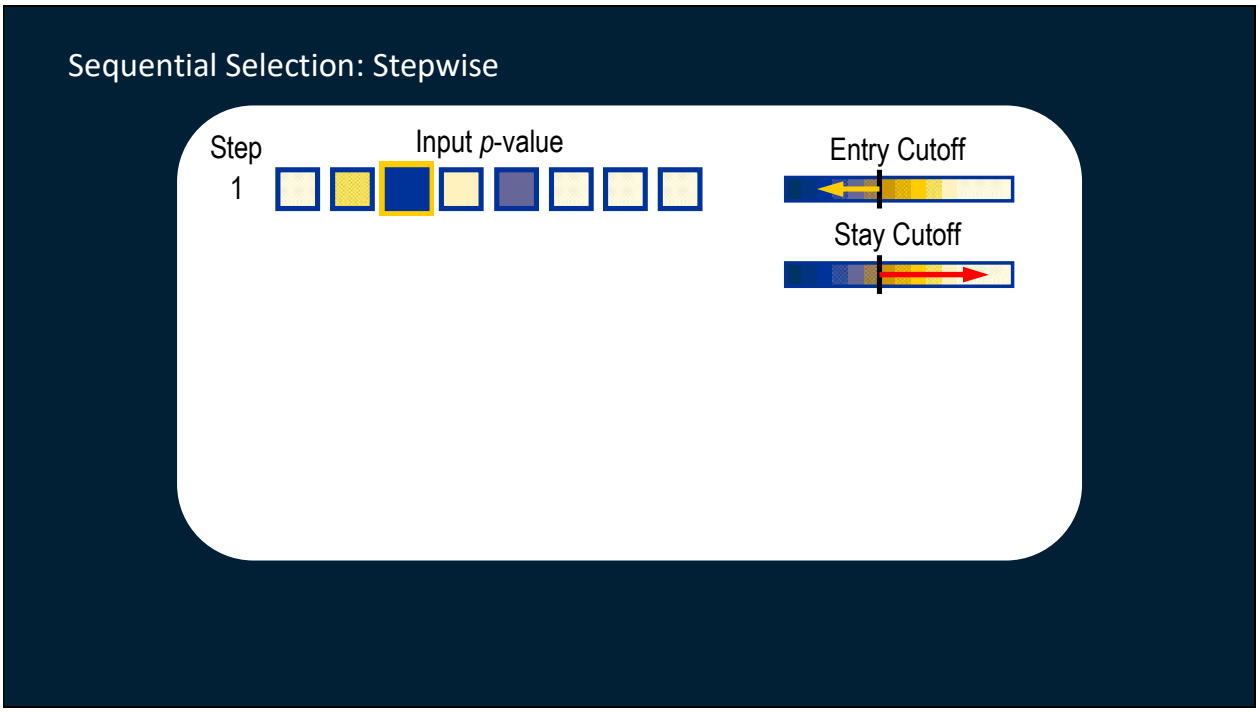
Sequential Selection: Backward



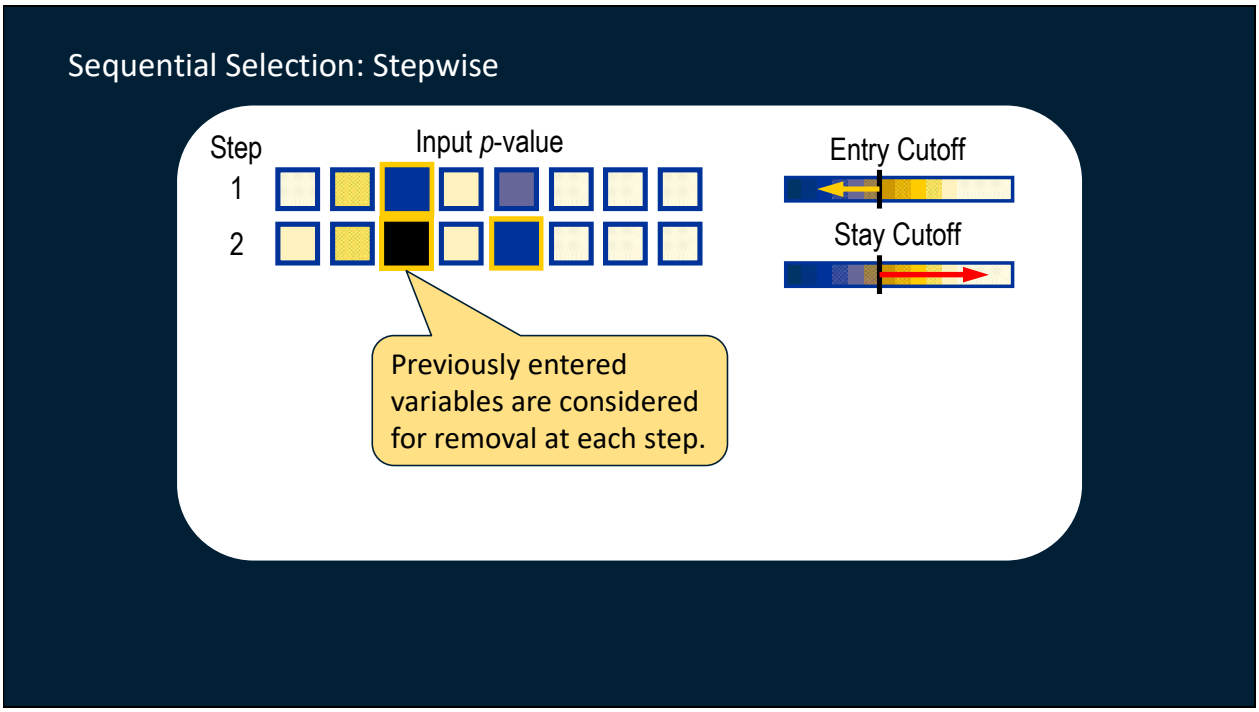




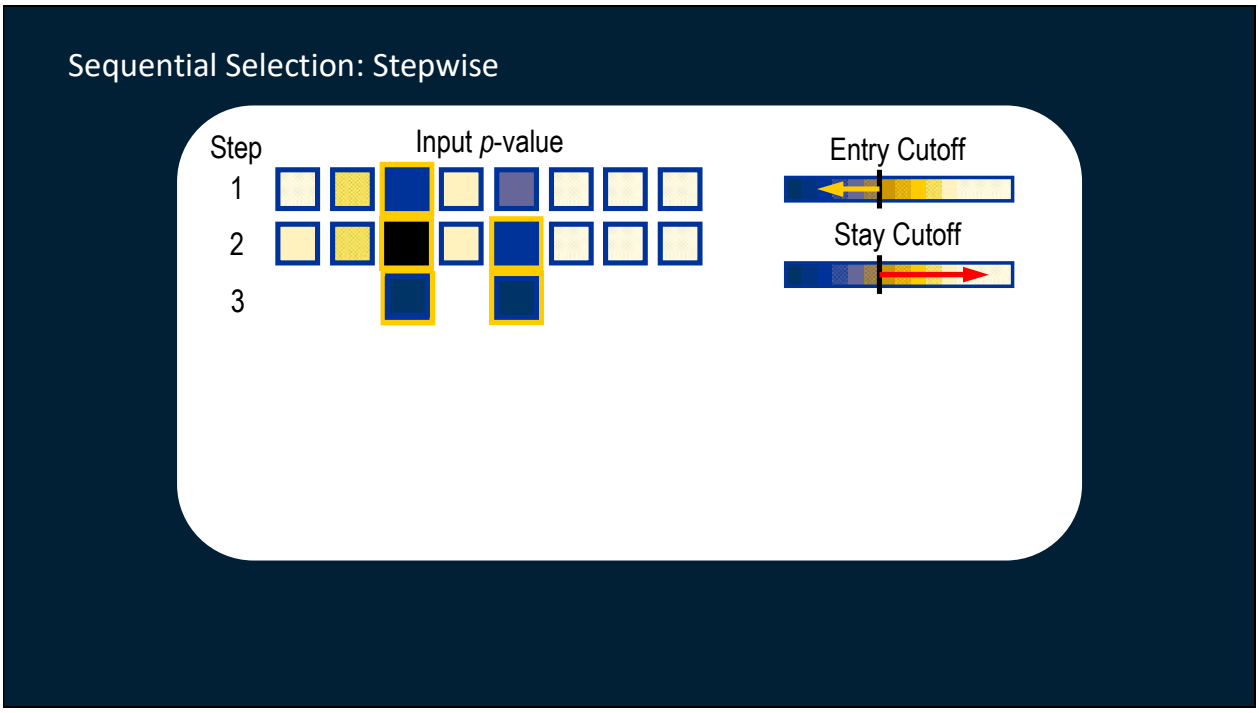
When all variables in the model have *p*-values low enough to stay, the backward elimination process stops.



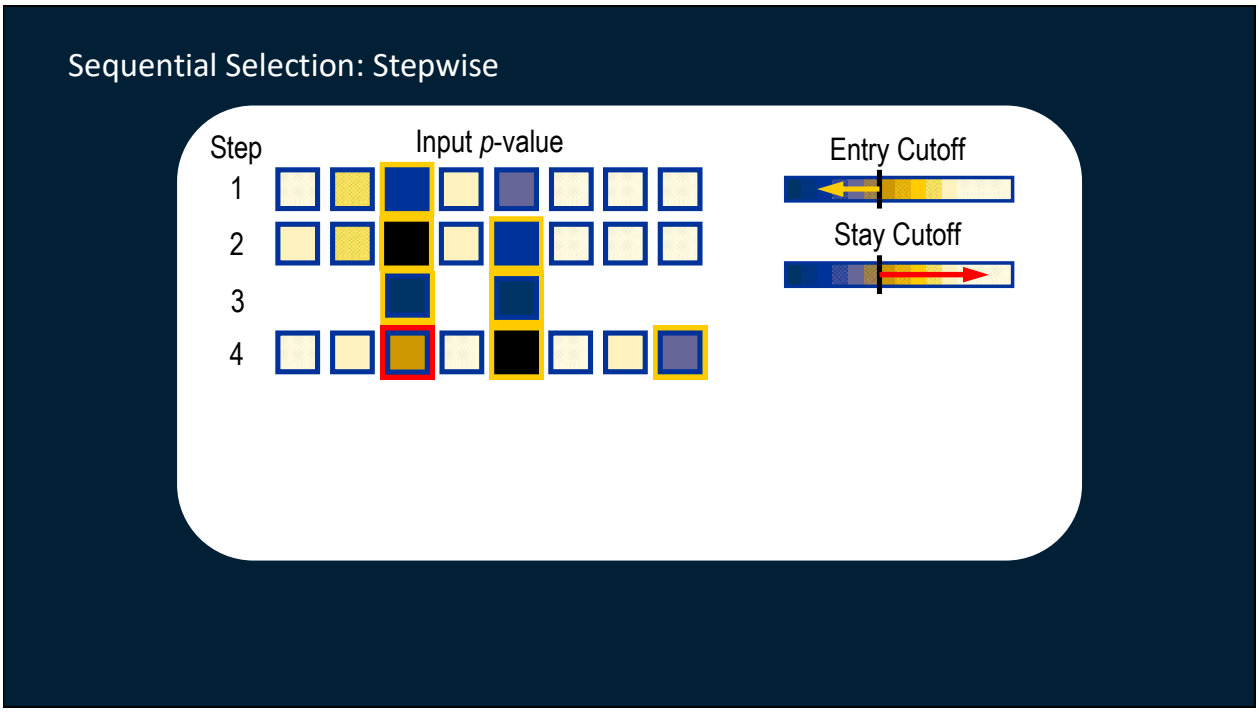
Stepwise combines elements of forward selection and backward elimination. It begins with an intercept-only model. Variables are added one at a time as in forward selection.



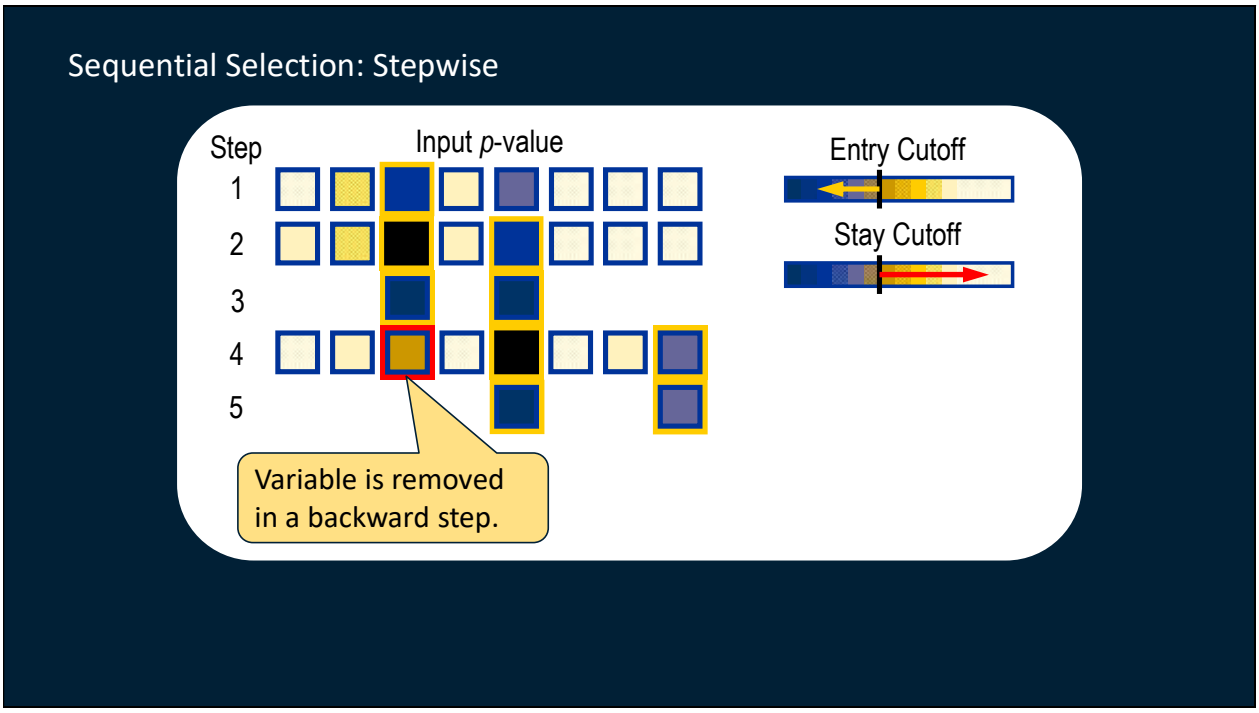
Once a second variable is added, previously added variables are removed in a backward step if they no longer meet the criterion for remaining in the model.



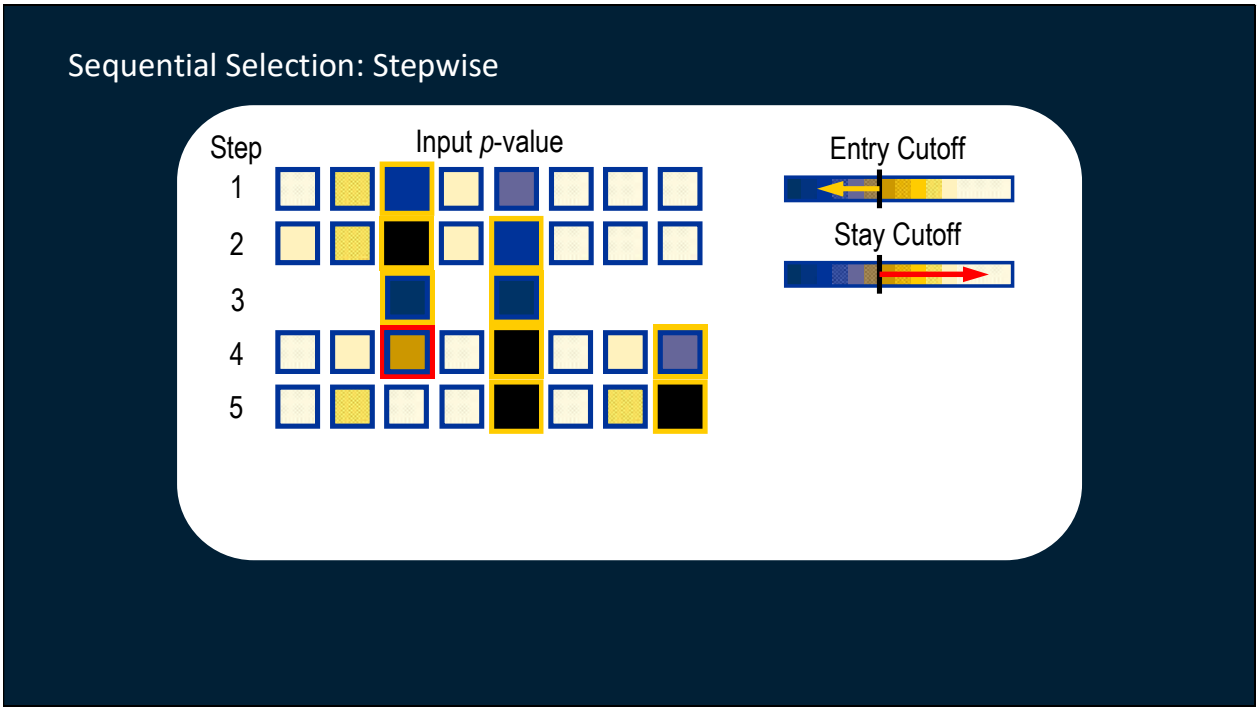
In the backward elimination stage of this example, neither of the two variables were removed because they both met the stay cutoff.

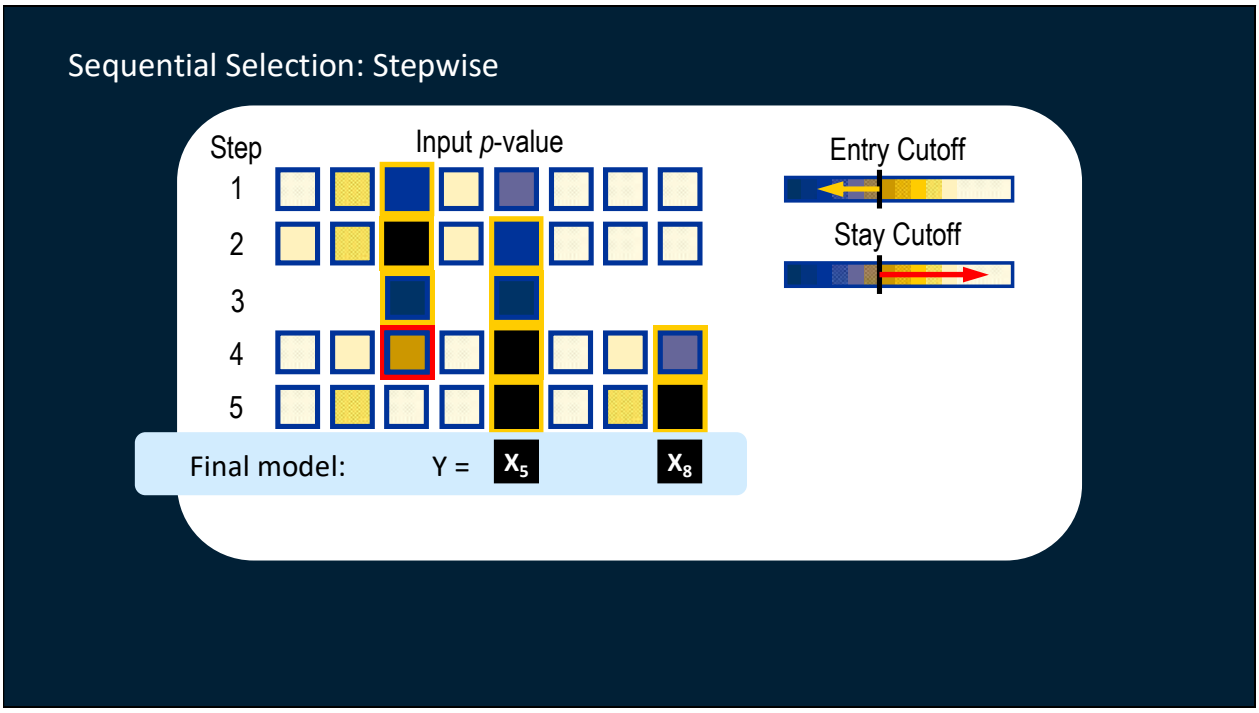


In the next step, when another variable entered in the model, one of the previous variables that was present in the model became insignificant.

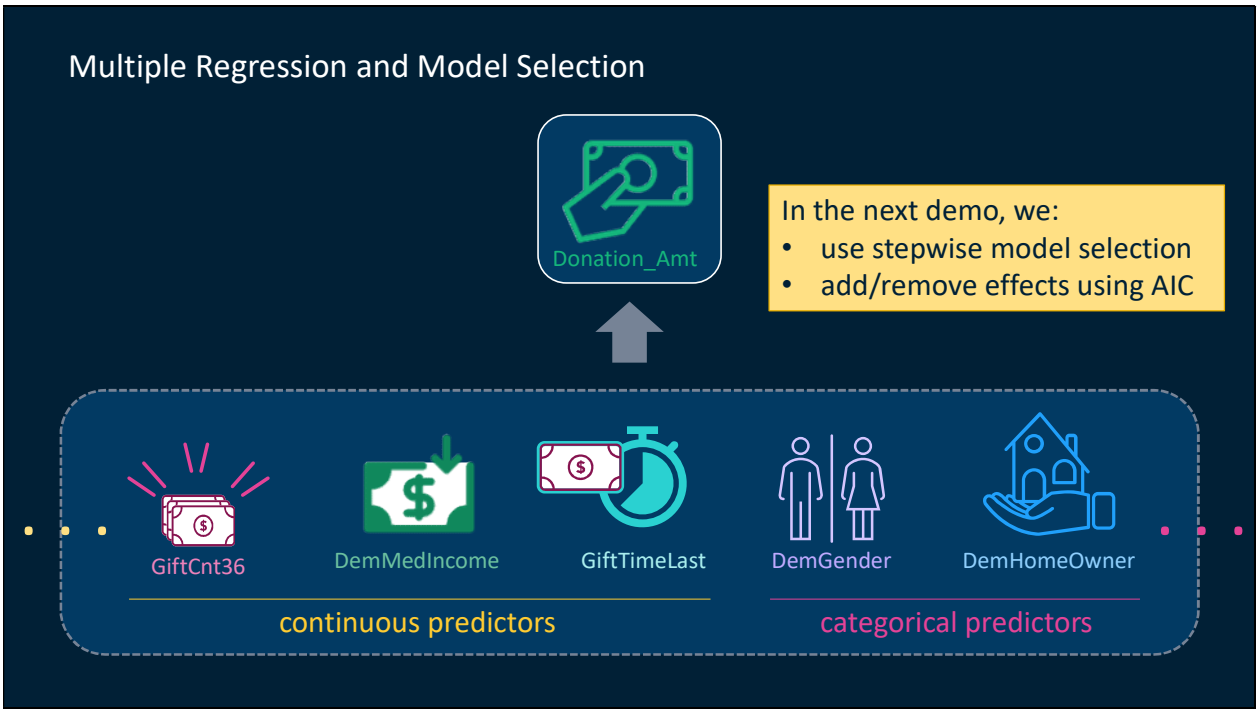


In step 4, a previously added variable is removed in a backward step.





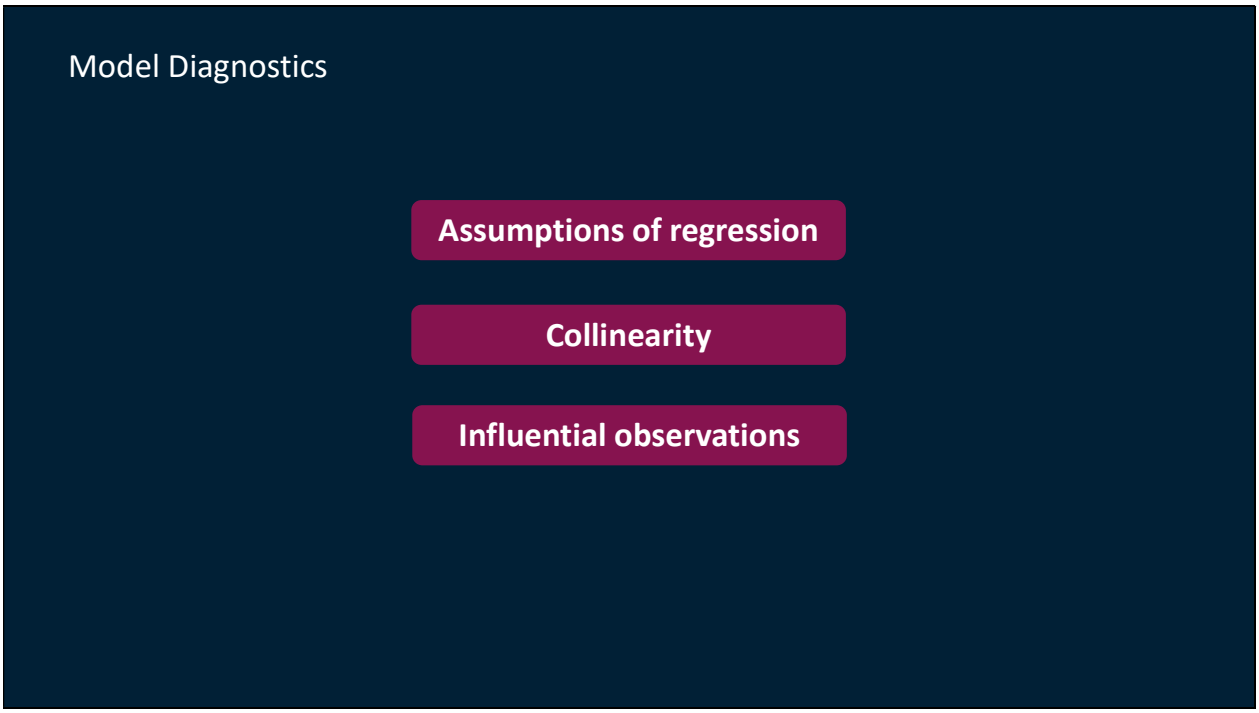
The process terminates when all inputs available for inclusion in the model have *p*-values in excess of the entry cutoff, and all inputs already included in the model have *p*-values below the stay cutoff.



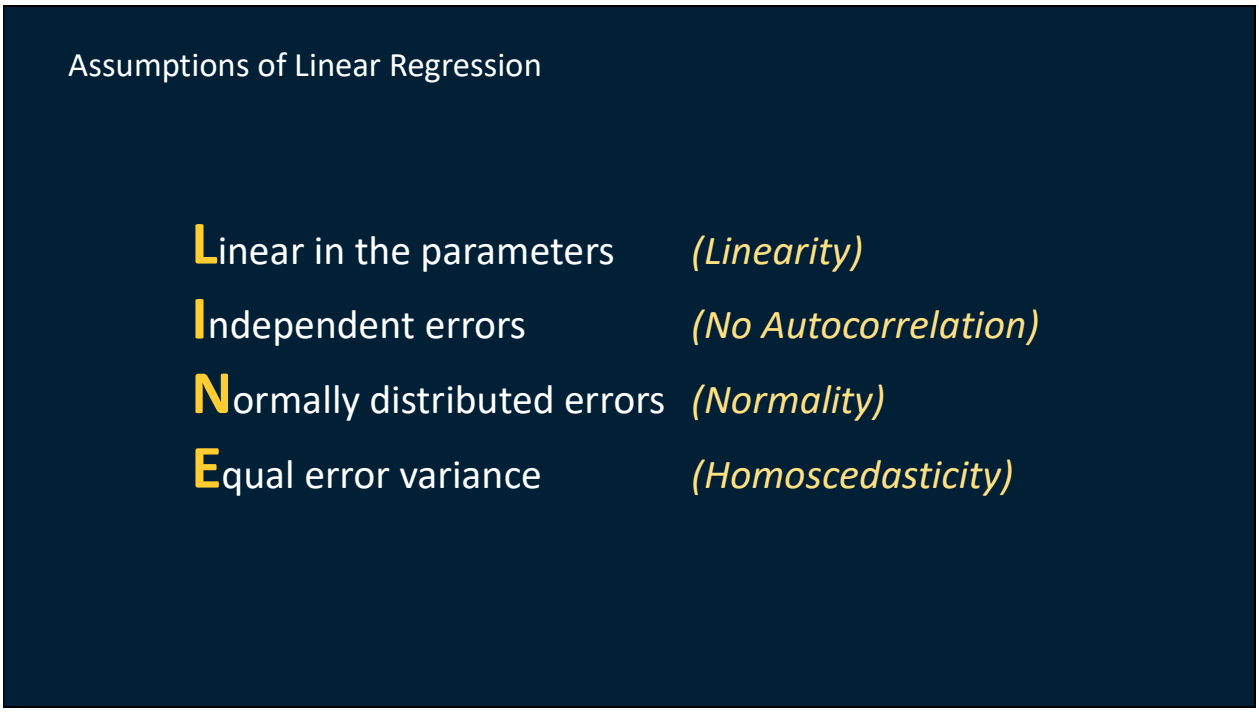
In the next demonstration, we fit a multiple linear regression model. We will include continuous predictors such as **Gift Count 36 Months**, **DemMedIncome**, and **GiftTimeLast** and categorical

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.

3.3 Model Diagnostics



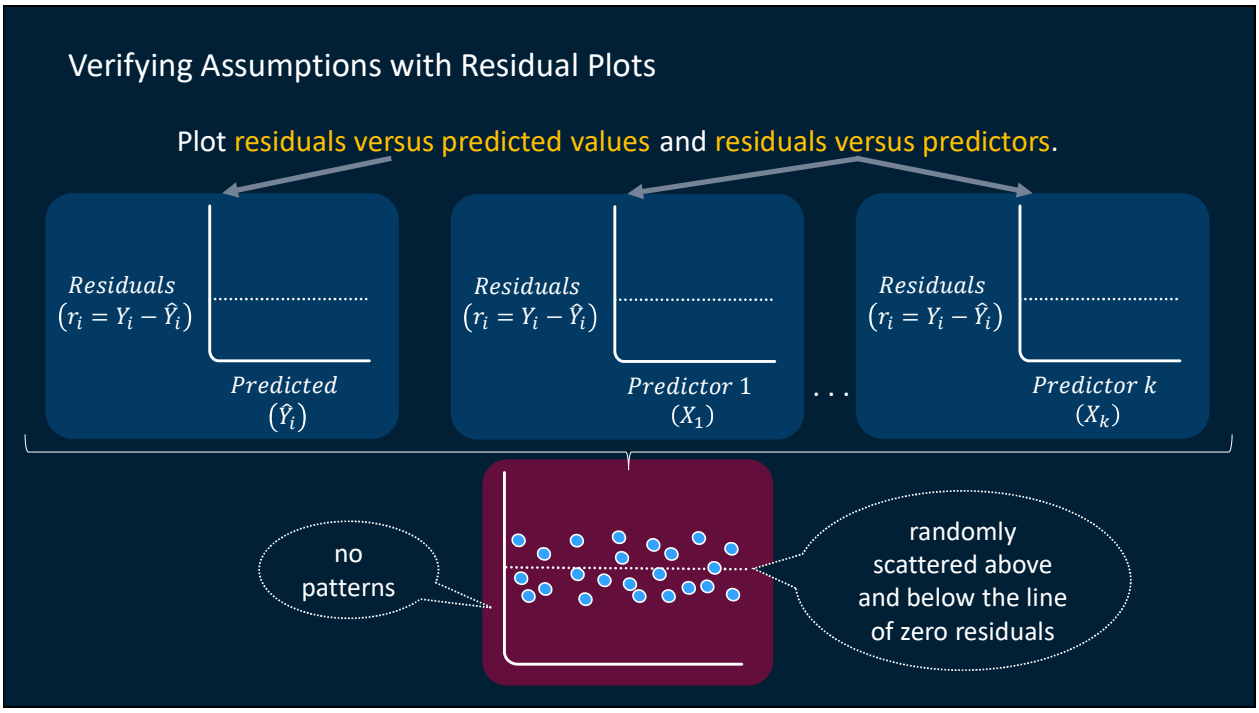
Model diagnostics for regression can involve checking the assumptions of the model, testing for collinearity (strong correlations among sets of predictors), and looking for outliers and influential observations.



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.

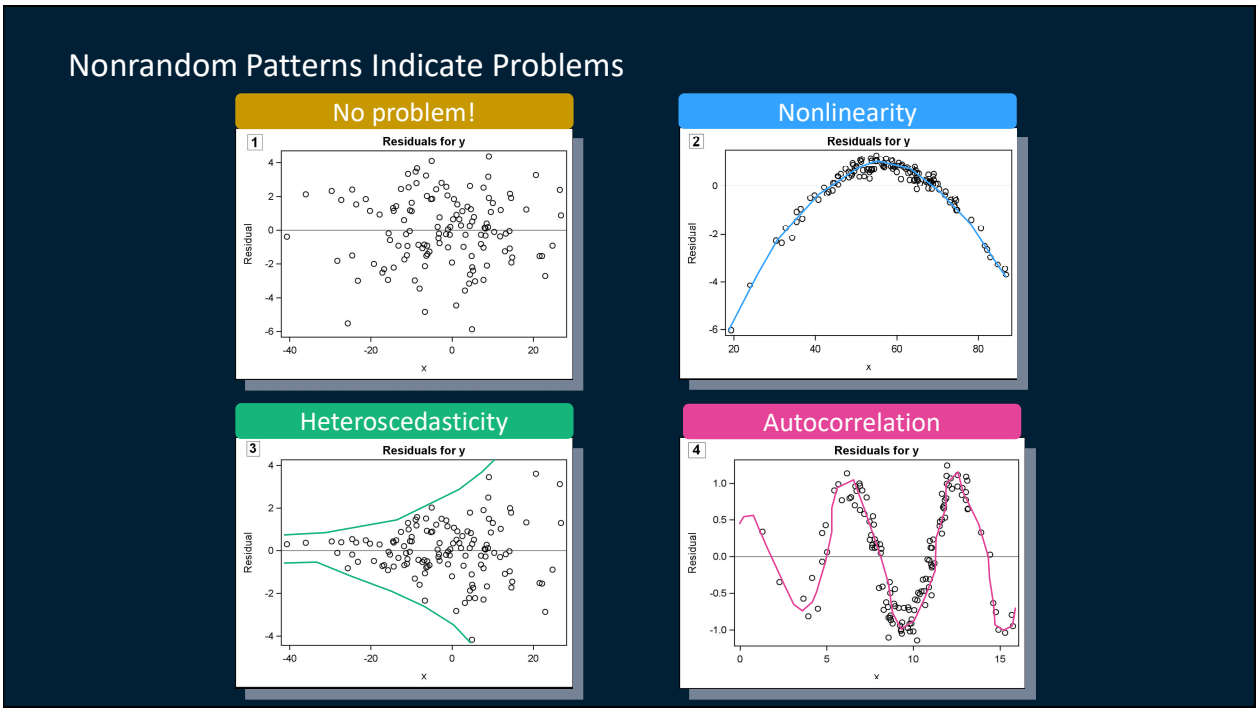
Independent normally distributed errors with constant or equal variance errors are all required for statistical inference. Conveniently, these assumptions form the acronym LINE, making them easy to remember.



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.



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.

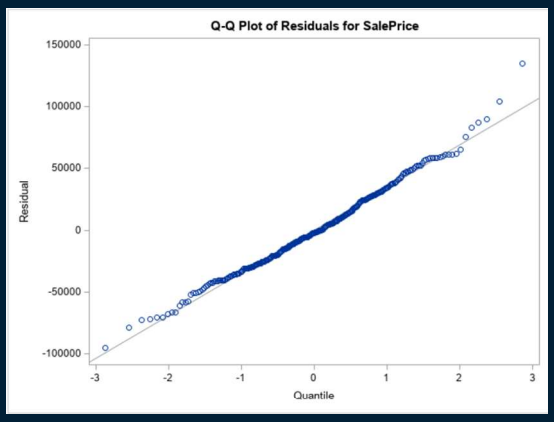
A curvilinear residual pattern (top right) can indicate a misspecified model.

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

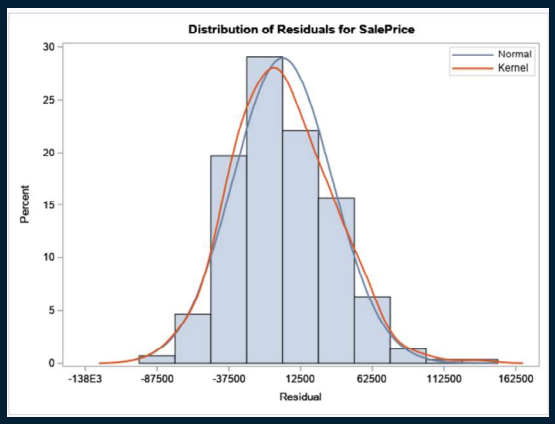
Violations of these assumptions can limit the inferences that an explanatory modeler can make from a regression, but are less problematic when the regression is used for pure prediction.

Graphical assessment of normality requires different plots.

Check Normality with Other Plots



normal quantile-quantile plot

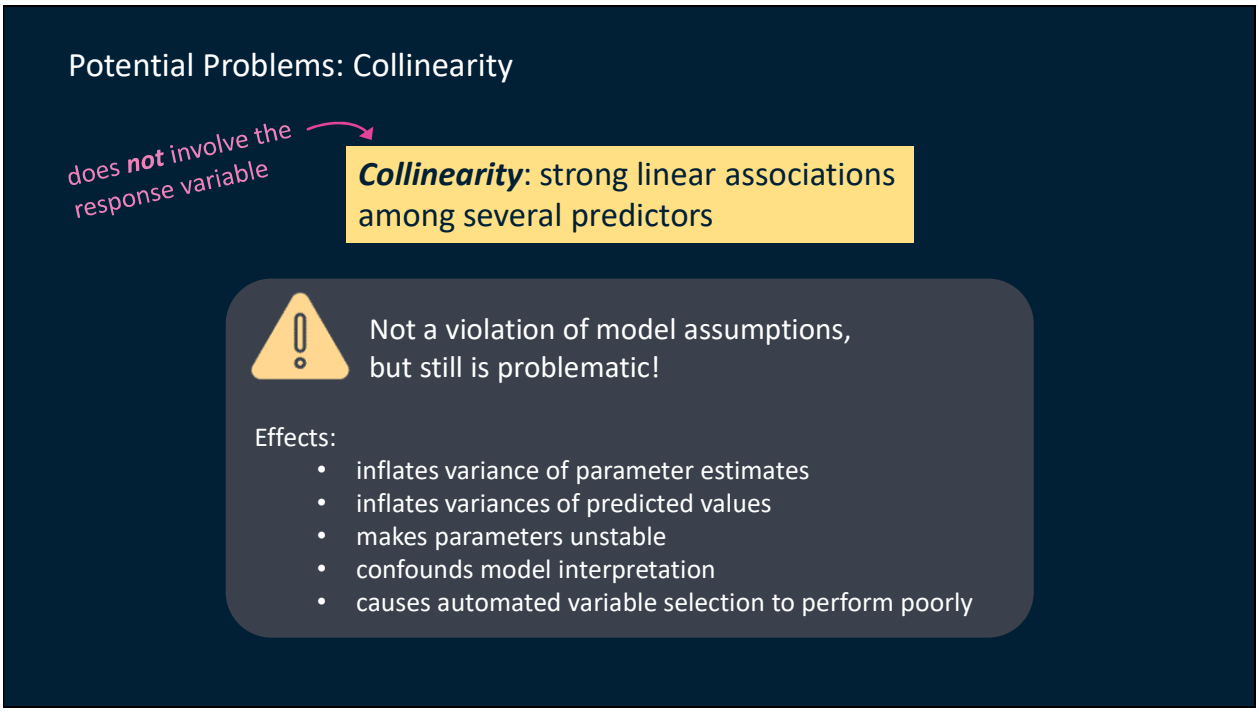


histogram of residuals

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.

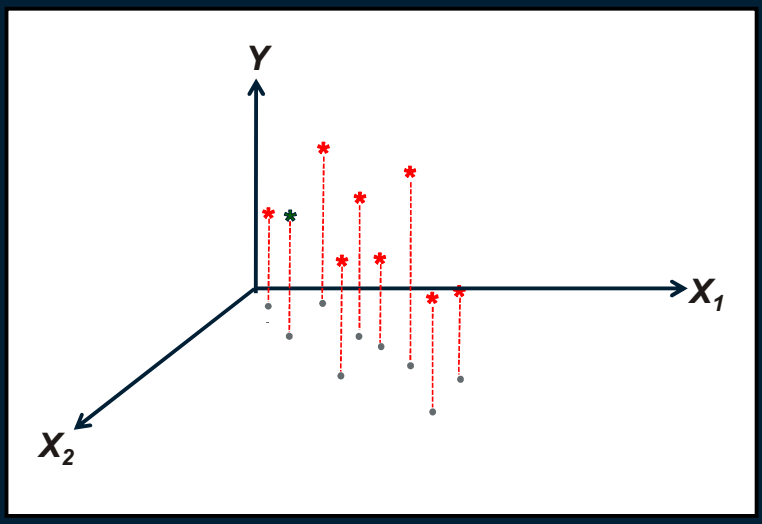


Besides checking regression model assumptions, model assessment can include checking for collinearity. Collinearity (also called multicollinearity) means strong associations among sets of predictors. Collinearity is an unsupervised concept. That is, it involves only predictors, not the response variable. 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.

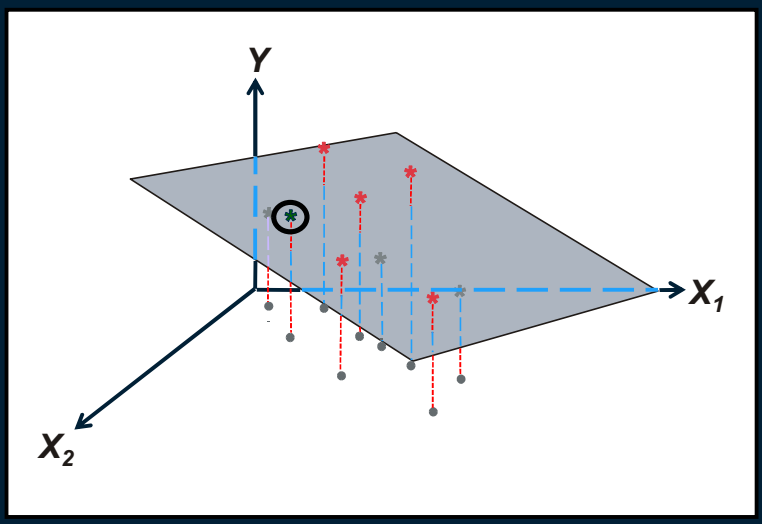
Let's see a graphical depiction of collinearity.

Illustration of Collinearity



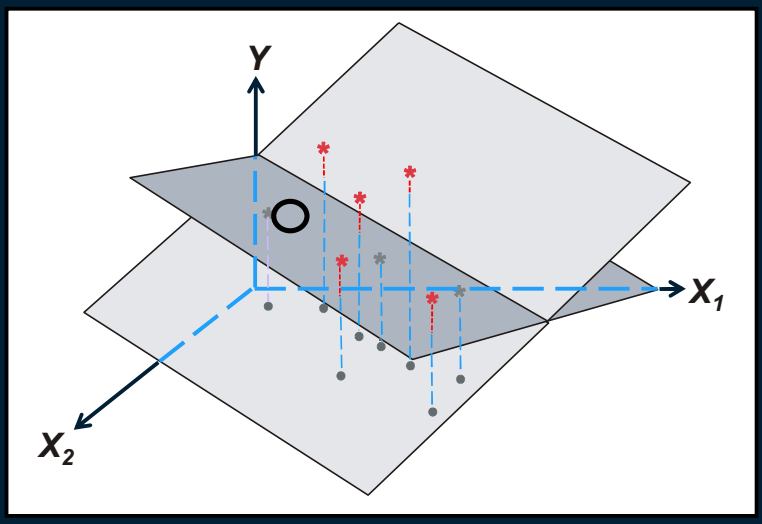
In this illustration of collinearity, the drop shadows for the data points show that X1 and X2 are correlated. This correlation makes the regression coefficients unstable. What does instability really mean? It means small changes in the data can lead to large changes in the parameter estimates.

Illustration of Collinearity

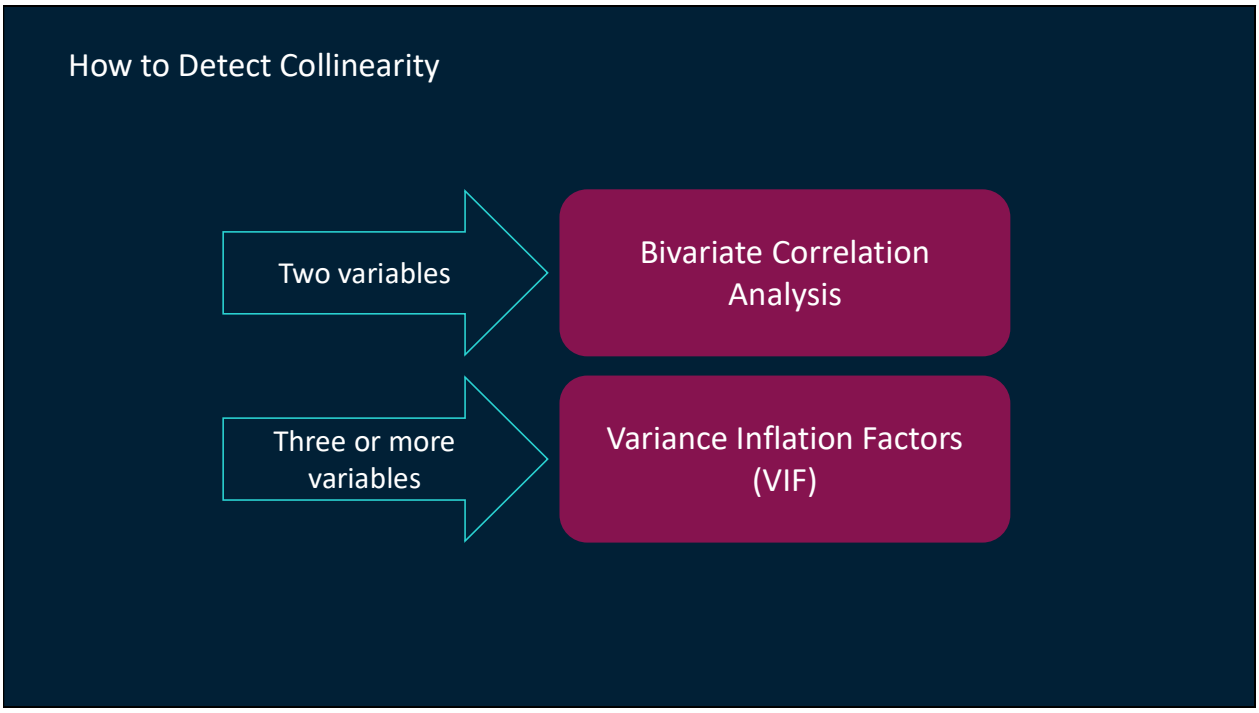


If we fit a multiple regression model to these data, the estimated coefficients can be visualized as the tilt of the best-fit regression plane. With collinearity, these regression coefficients are unstable. What would happen if we remove a single data point and refit the model?

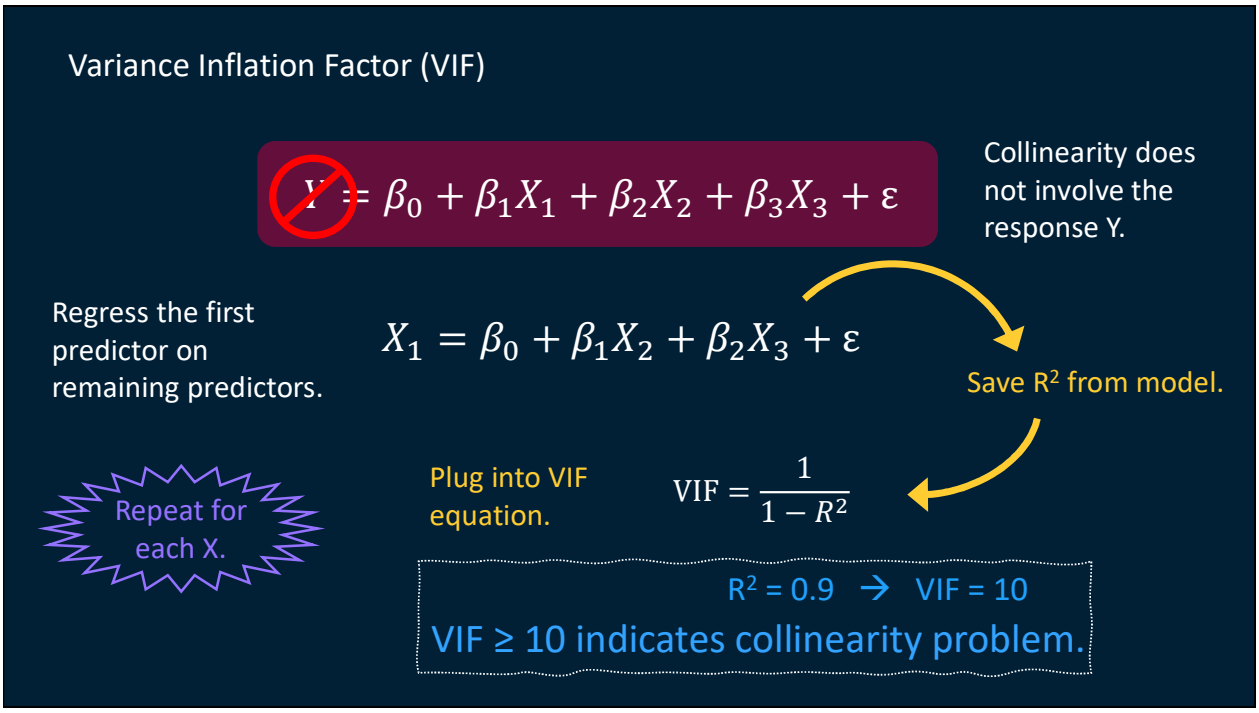
Illustration of Collinearity



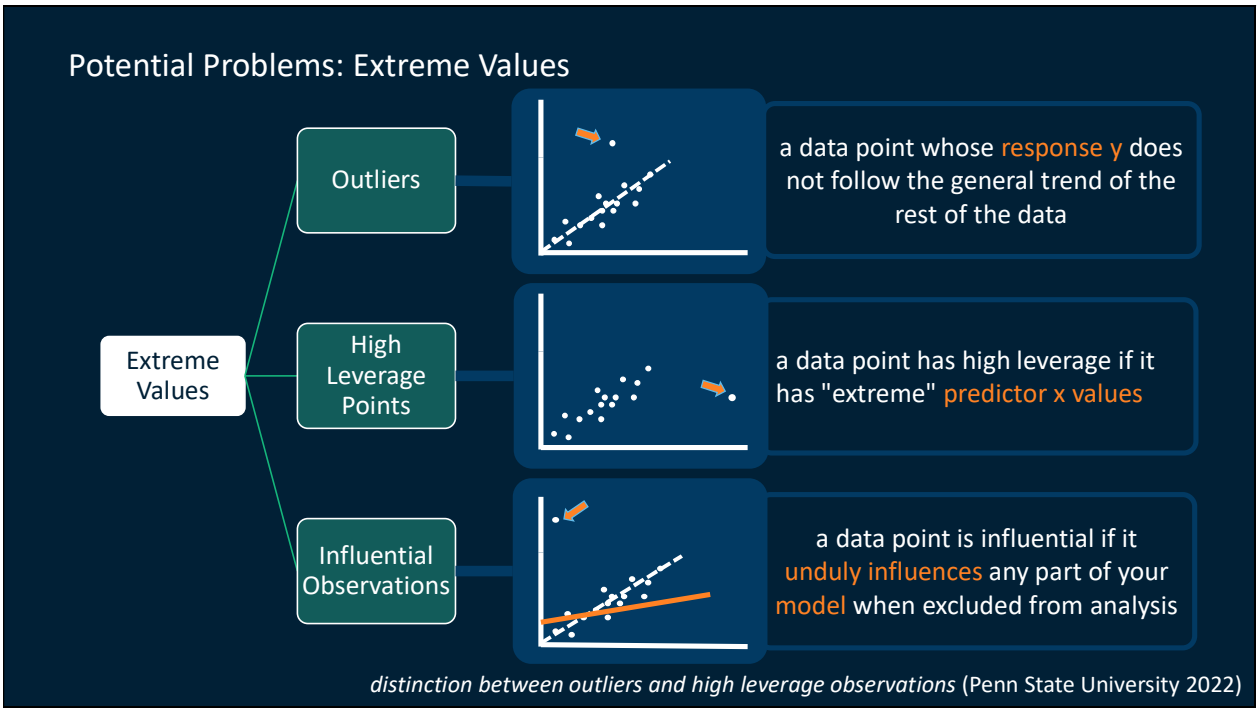
Removal of a single data point causes the slopes to change sign and magnitude! Small changes in the data lead to changes in the parameter estimates that would drastically alter interpretation of this model.



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.



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

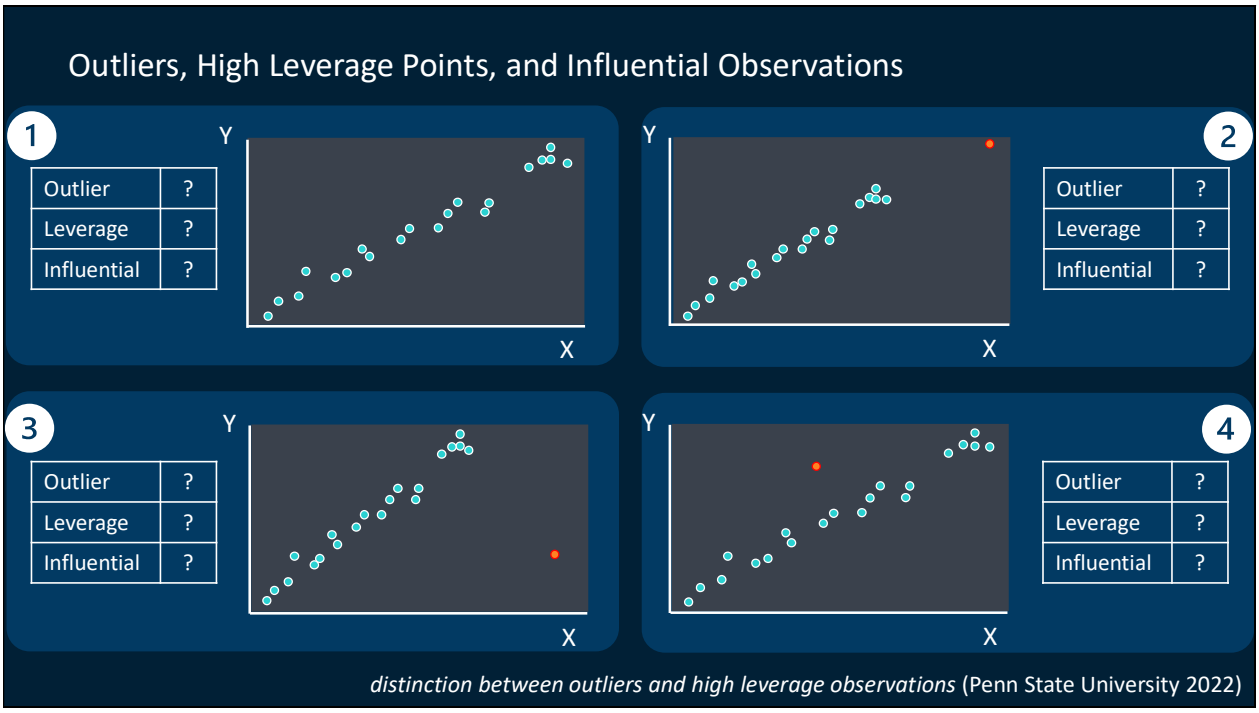
An *outlier* is a data point whose response y does not follow the general trend of the rest of the data.

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.

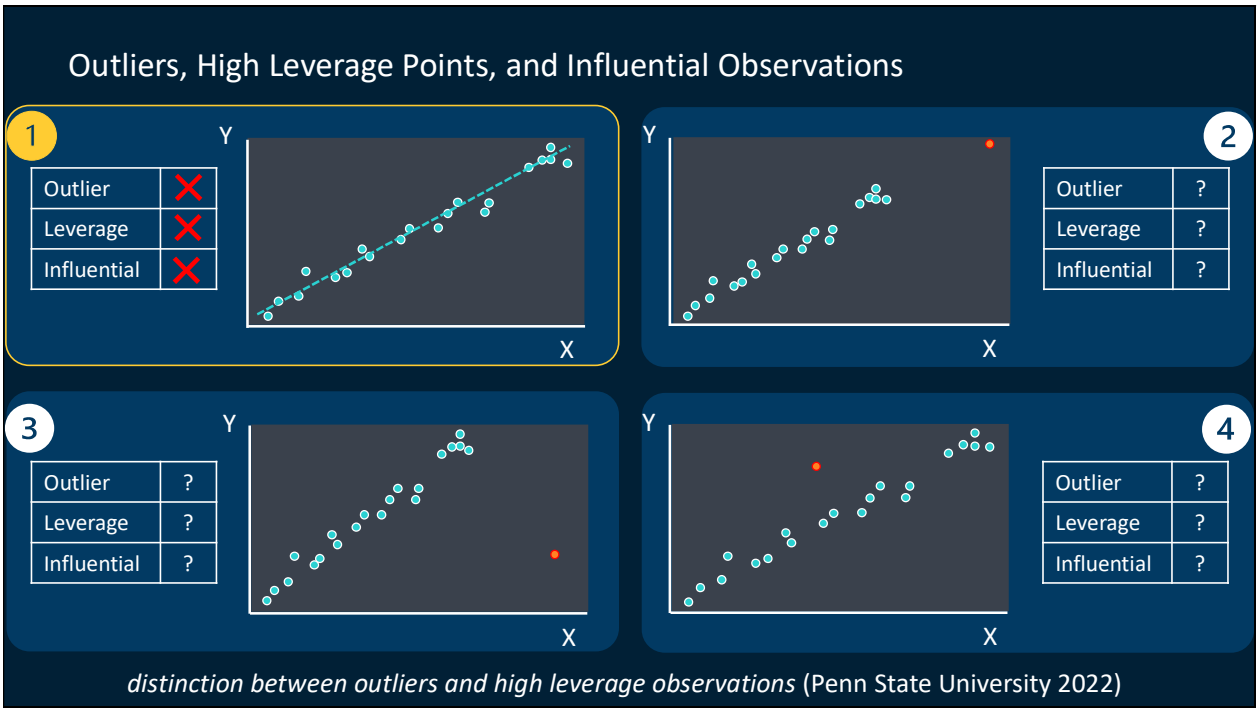
Note that — for our purposes — we consider a data point to be an outlier only if it is extreme with respect to the other y values, not the x values.

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.

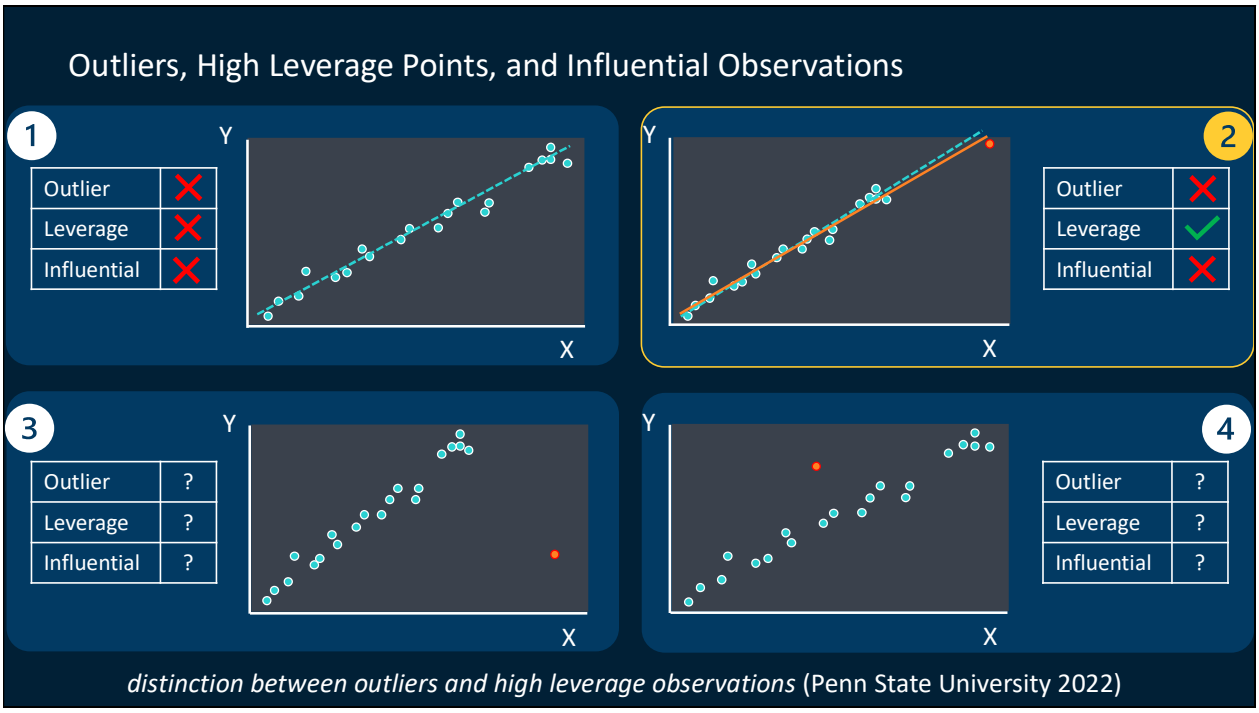
The presence of outliers or influential observations does not violate the assumptions of linear regression directly. But the unusual data points can affect the linearity assumption or distributional assumptions of the model.



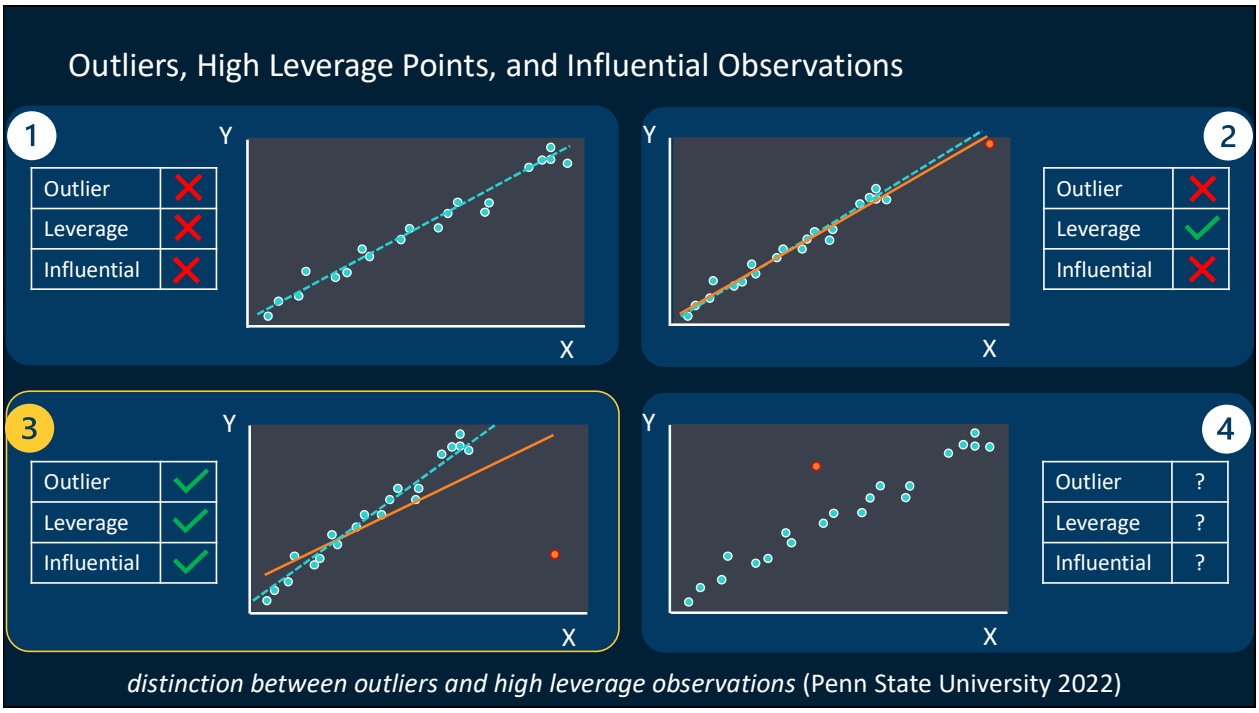
Based on the definitions just now discussed, do you think the four illustrations contain any outliers? Or any high leverage points or influential observations?



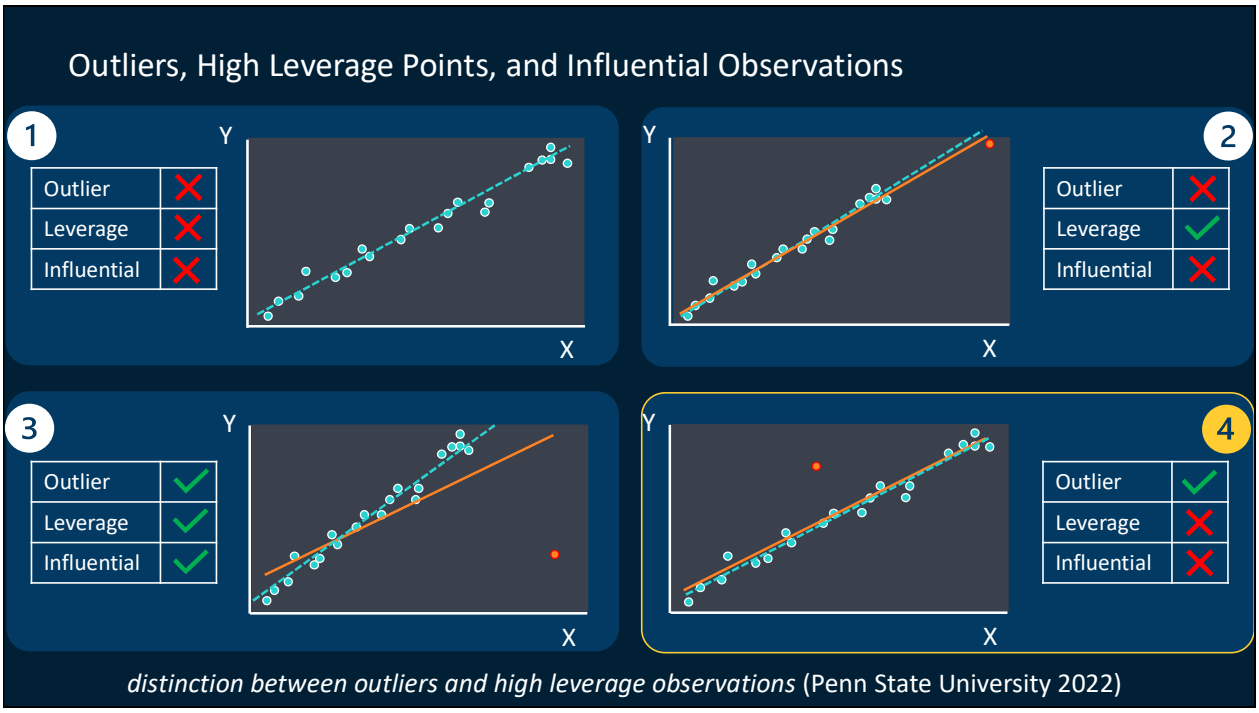
Example (1): All the data points follow the general trend of the rest of the data, so there are no outliers (in the y direction). And none of the data points are extreme with respect to x, so there are no high leverage points. Overall, none of the data points would appear to be influential with respect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



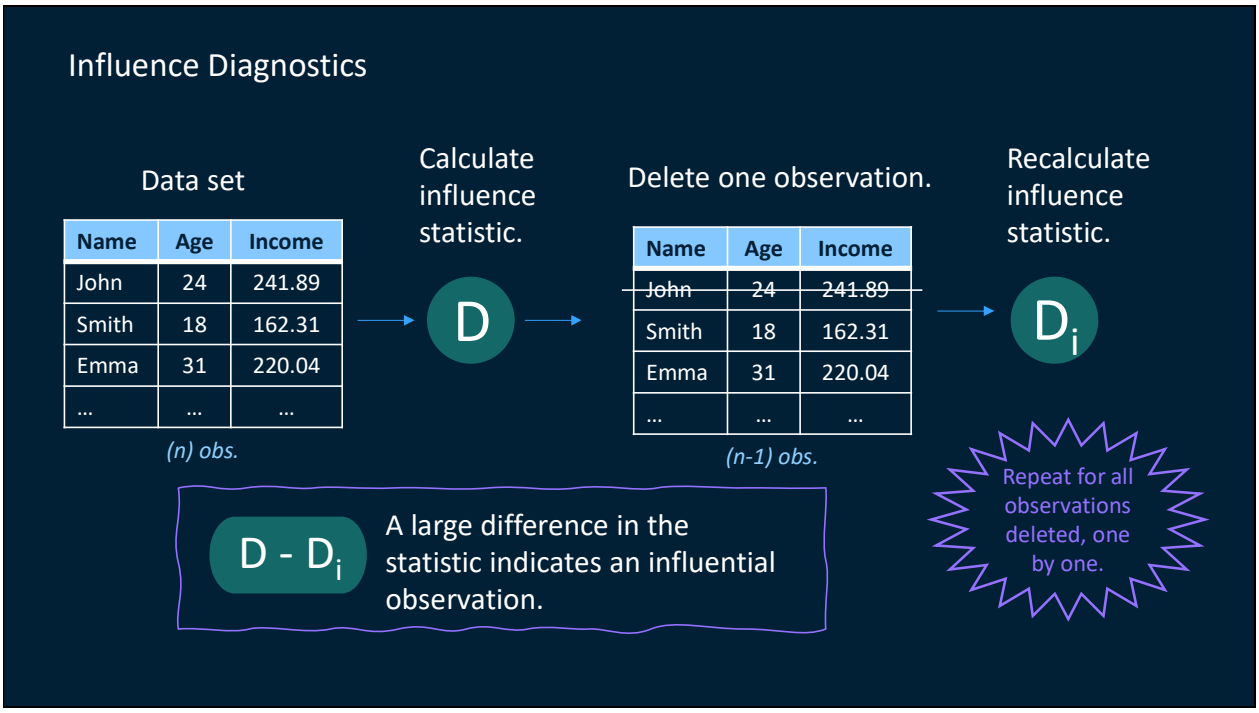
Example (2): In this case, the orange data point does follow the general trend of the rest of the data. Therefore, it is not deemed an outlier here. However, this point does have an extreme x value, so it does have high leverage. Is the orange data point influential? An easy way to determine whether 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.



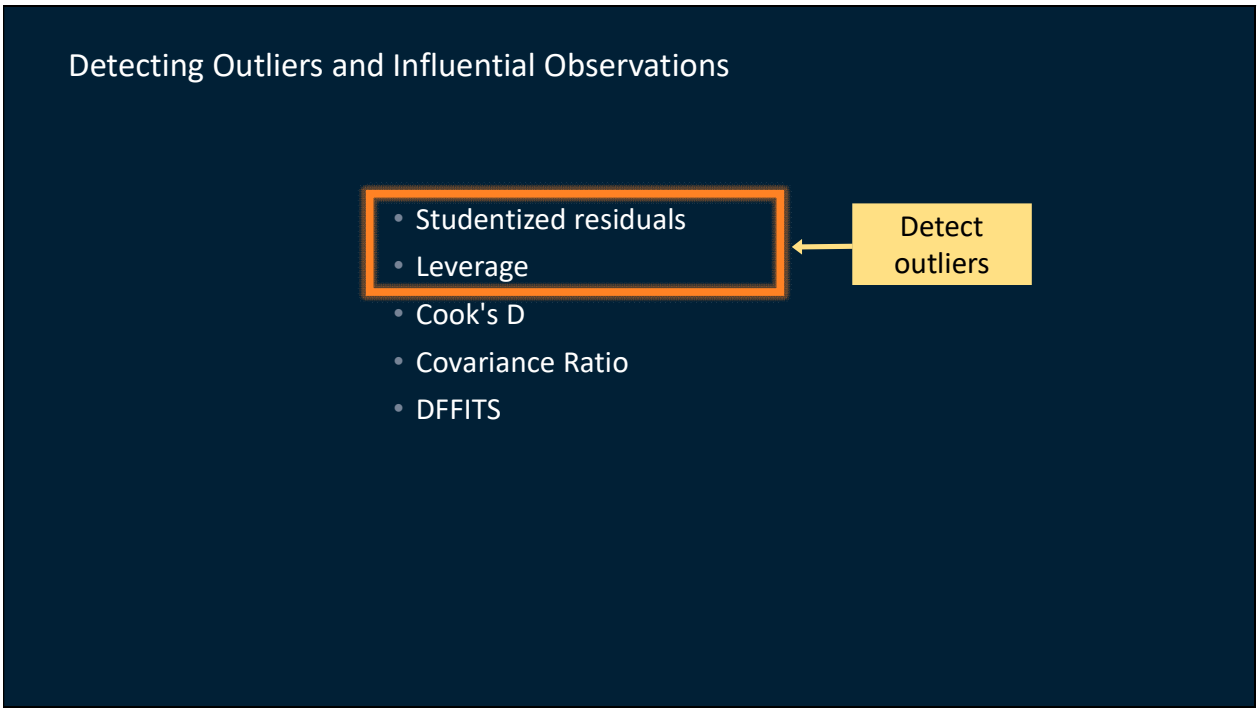
Example (3): In this case, the orange data point is most certainly an outlier and has high leverage! The orange data point does not follow the general trend of the rest of the data, and it also has an extreme x value. And, in this case the orange data point is influential. The two best fitting lines, 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.



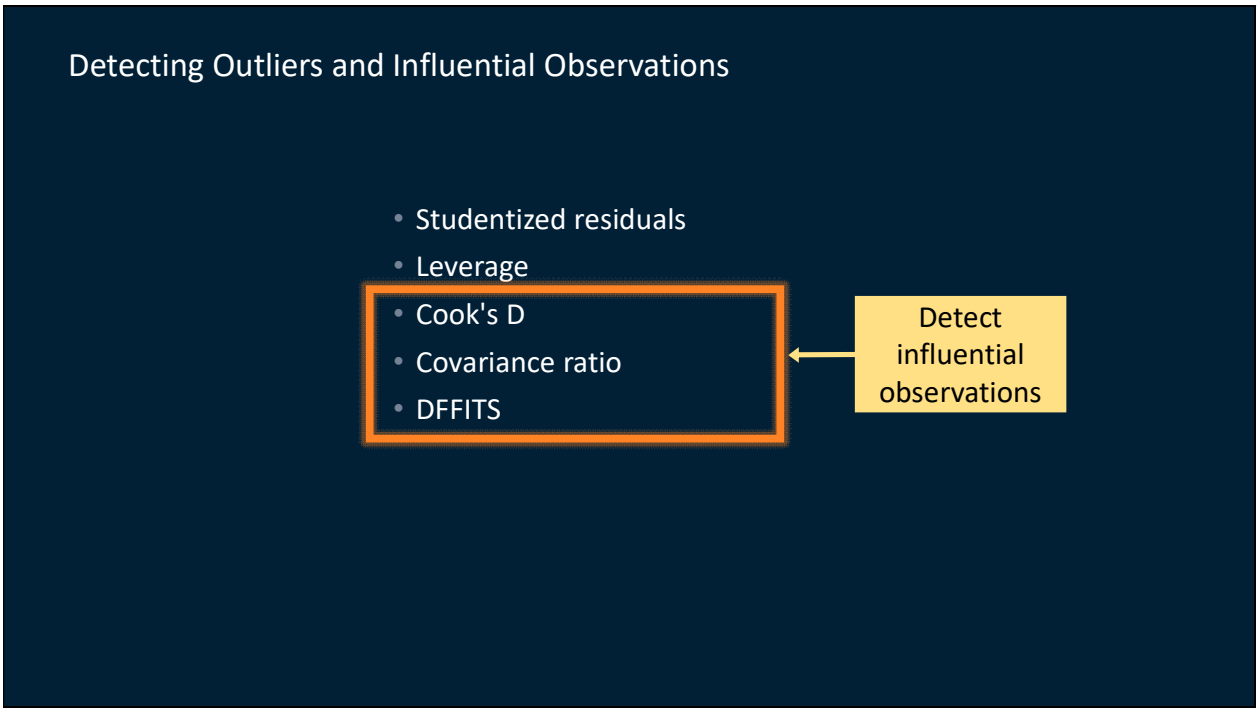
Example (4): Because the orange data point does not follow the general trend of the rest of the data, it would be considered an outlier. However, this point does not have an extreme x value, so it does not have high leverage. It's hard to even tell the two estimated regression lines apart! So, it's not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



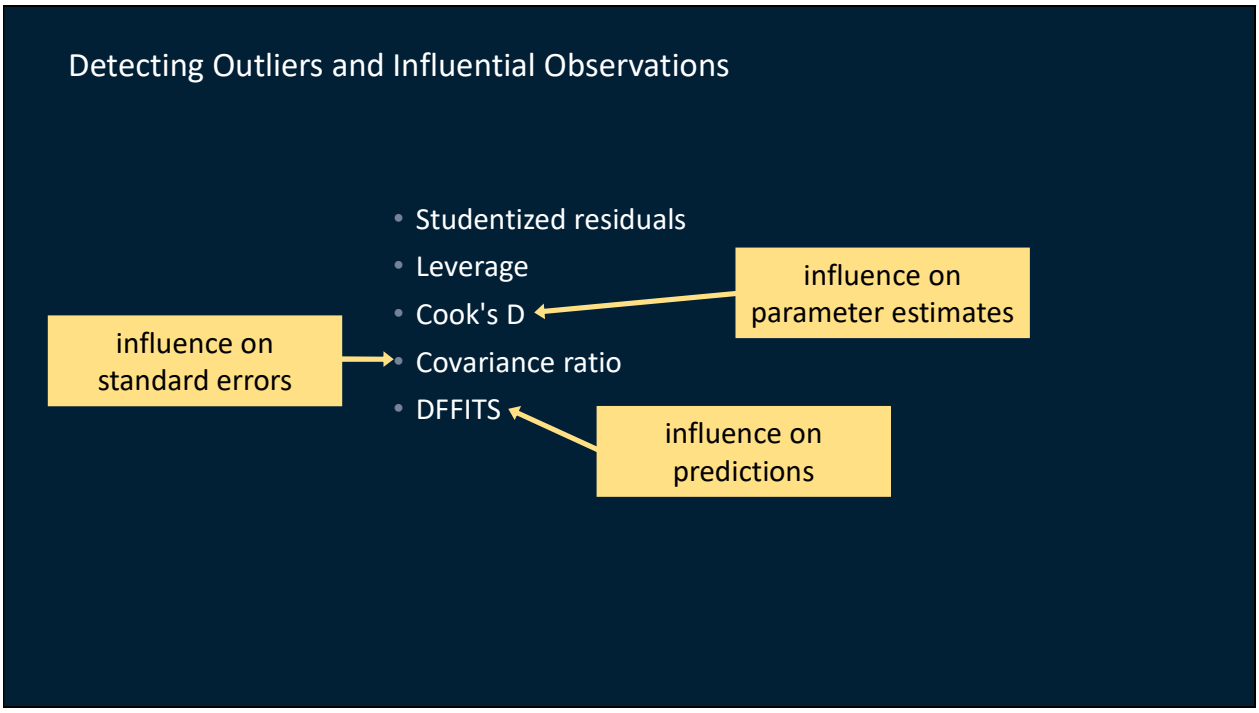
How do we decide whether an observation is influential? Most of the influence statistics take the same approach. They take your data set and calculate an influence statistic of interest. This statistic could measure any of several different parts of the model. Then an observation is deleted, and the 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.



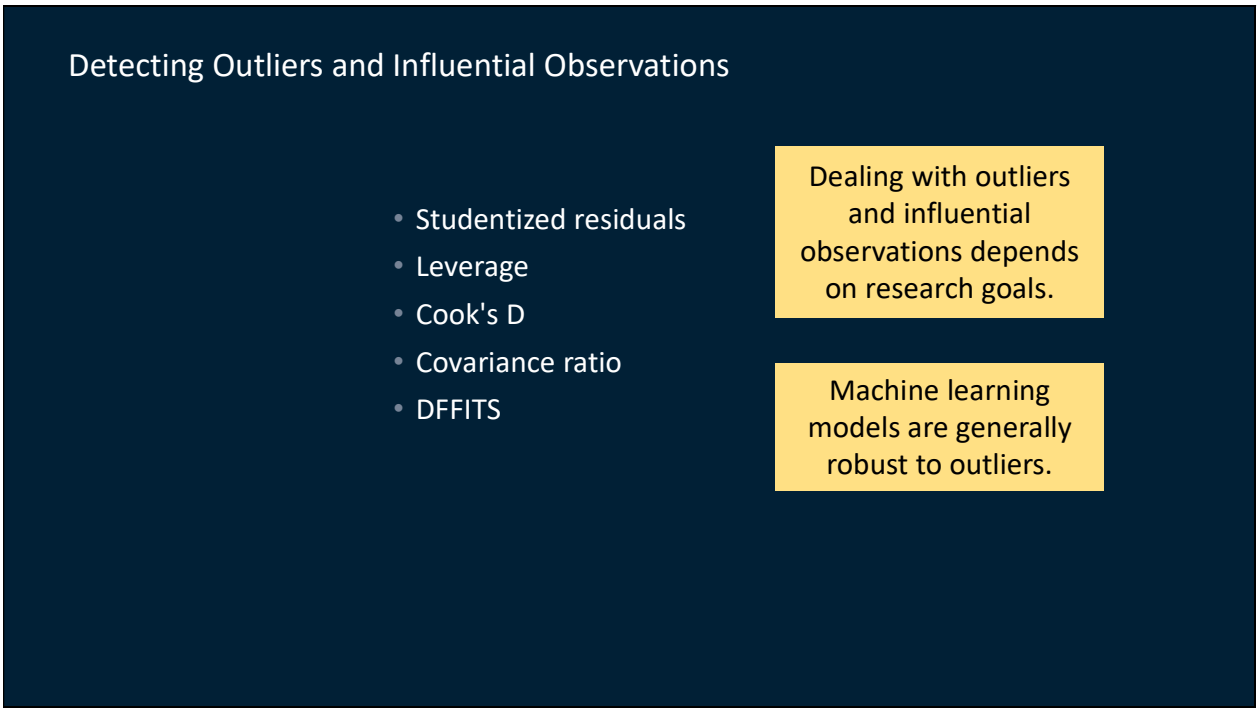
Cook's D, covariance ratios, and DFFITS help find observations that influence various aspects of the model.



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."

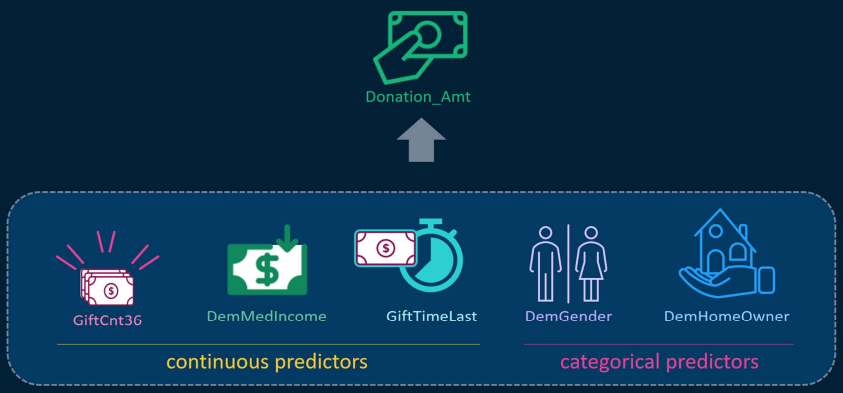
Another approach to determining how big is big enough is to look for observations with large values of the influence statistics that are well separated from other values. Several advanced types of regression models such as generalized linear models and mixed models have no suggested cutoffs for influence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



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



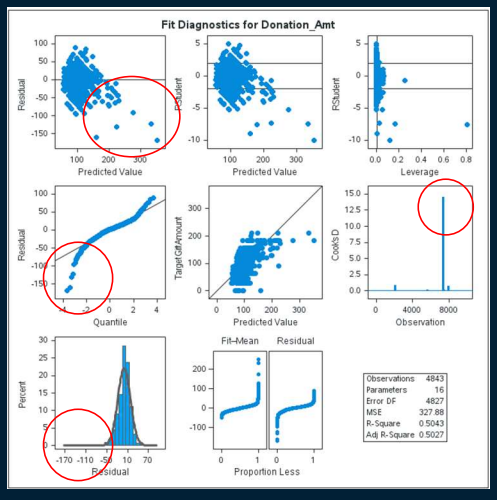
- Were the regression assumption met?
- Is there collinearity among predictors?
- Are there influential observations or outliers?

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.

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.