openSAP
Introduction to Statistics for Data Science

Week 3 Unit 1

00:00:06	Hello, and welcome back to week three of the openSAP course An Introduction to Statistics for Data Science.
00:00:15	In this unit we will look at correlation as a statistical measure.
00:00:21	When two sets of data are strongly linked together, we say they have a high correlation.
00:00:29	Correlation is positive when the values increase together, and correlation is negative when one value
00:00:37	decreases as the other increases. Correlation can have values
00:00:43	ranging from minus one to plus one. One is a perfect positive correlation.
00:00:49	Zero is no correlation, the values don't seem to be linked at all.
00:00:54	And minus one is a perfect negative correlation. In this example, over the past two weeks,
00:01:02	a shop that sells ice creams has measured the temperature each day and the sales values
00:01:09	of the ice creams that were sold. This is plotted on a scatter plot shown in the slide.
00:01:16	You can see that there are higher sales values when the temperature increases,
00:01:22	so there is a correlation between sales and temperature. The Pearson product-moment correlation coefficient
00:01:35	is a measure of the strength and direction of the linear relationship between two variables.
00:01:41	It's referred to as Pearson's correlation or simply as the correlation coefficient
00:01:48	and is calculated using the formula shown on the slide. If you are familiar with entering functions
00:01:56	in Microsoft Excel you could enter the CORREL command, and that will allow you to calculate
00:02:04	the relationship between the variables. If the relationship is non-linear,
00:02:09	then the correlation coefficient does not adequately represent the strength
00:02:15	of the relationship between the variables. In the ice cream example,
00:02:22	temperature is x and ice cream sales is y. Follow these steps to calculate
00:02:29	Pearson's correlation coefficient. The first step is to find the mean of x, and the mean of y.
00:02:36	And then you can subtract the means from every value and follow other steps through
00:02:42	in this calculation that's actually showing on the slide for you.
00:02:49	Finally, you divide all the sums, and this will allow you to calculate the Pearson's correlation coefficient,
00:02:58	which in this example is +0.96, which indicates a very high positive correlation
00:03:04	between temperature and ice cream sales. Pearson's correlation is only relevant
00:03:12	when the relationship is linear. If the data in the example is extended with points
00:03:17	that reflect that sales reduce if temperatures increase further,
00:03:23	then the scatter plot could be as it's shown in the slide. In this example, customers will stop buying ice creams,
00:03:31	maybe because it's so hot they don't go shopping. The correlation coefficient for this data is zero,
00:03:40	indicating there is no correlation. However, the scatter plot shows there is a correlation,
00:03:47	but it's a curve, peaking at around 25 degrees Centigrade. You now understand why it's important to visualize

00:03:55	the data when you are analyzing correlations. These are the scatter plots of Anscombe's quartet.
00:04:04	It's a set of four different pairs of variables created by the English statistician Frank Anscombe in 1973.
00:04:14	These graphs were constructed to demonstrate both the importance of graphing data before analyzing it
00:04:23	and the effect of outliers on statistical properties. The first graph, on the top left,
00:04:30	seems to be distributed normally, and the two variables appear to be correlated.
00:04:37	The second graph, on the top right, is not distributed normally. There is an obvious relationship between the two variables,
00:04:45	but it's not linear. In the third graph, the bottom left,
00:04:50	the linear relationship is perfect, except of course for one outlier
00:04:55	which exerts enough influence to lower the correlation coefficient from one down to 0.816.
00:05:04	In the fourth graph, an outlier produces a high correlation coefficient,
00:05:09	even though the relationship between the two variables is non-linear.
00:05:15	However, the four y variables have the same mean, variance, and correlation, although the distribution
00:05:22	of the variables is very different. These examples indicate that the correlation coefficient,
00:05:30	as a summary statistic, cannot replace visual examination of the data.
00:05:36	You should visualize the data and calculate the correlation coefficient.
00:05:43	So to summarize: Correlation measures the strength
00:05:47	of association between two variables. The most common correlation coefficient is called
00:05:54	the Pearson product-moment correlation coefficient. The sign and the absolute value of a Pearson correlation
00:06:04	coefficient describe the direction and the magnitude of the relationship between two variables.
00:06:13	The value of a correlation coefficient ranges between minus one and plus one.
00:06:18	The greater the absolute value of a correlation coefficient, the stronger the linear relationship.
00:06:27	The strongest linear relationship is indicated by a correlation coefficient of minus one or plus one.
00:06:34	The weakest linear relationship is indicated by a correlation coefficient of zero.
00:06:40	A positive correlation indicates that as one variable increases in value,
00:06:46	the other variable tends to increase as well. A negative correlation indicates
00:06:52	that as one variable increases in value, the other variable tends to reduce in value.
00:06:59	Remember that the Pearson correlation coefficient only measures linear relationships.
00:07:07	A correlation of zero does not mean zero relationship between the two variables;
00:07:13	it means zero linear relationship. It's possible for two variables to have zero linear
00:07:20	relationship and a strong curvilinear relationship at the same time.
00:07:26	So always remember to use scatter plots to visualize the relationship as well
00:07:32	as calculating the correlation coefficient. In the next unit, we will look
00:07:38	at correlation versus causation.

Week 3 Unit 2

00:00:06	Hello, and welcome back to week three, unit two, where we're going to consider correlation versus causation.
00:00:15	You've learned in a previous unit that correlation is not causation.
00:00:21	If two variables are highly correlated, it does not mean that one thing causes the other.
00:00:29	There can be many reasons why variables are correlated. However, you do need to think about
00:00:36	the relationship carefully because one variable might cause the other or might not.
00:00:42	If the ice cream shop measures the daily sales of sunglasses sold by an optician
00:00:48	and compares them to daily ice cream sales, then there is a clear high, positive correlation.
00:00:56	But sunglasses do not cause people to buy ice cream. Numerous epidemiological studies showed
00:01:05	that women taking combined hormone replacement therapy, HRT, also had a lower than average incidence
00:01:14	of coronary heart disease, CHD, leading doctors to propose that HRT
00:01:19	was in some way protective against heart disease. However, randomized controlled trials showed
00:01:25	that HRT caused, in fact, a small, but statistically significant increase
00:01:30	in the risk of coronary heart disease. Reanalysis of the data showed that women undertaking HRT
00:01:37	were more likely to be from higher socio- economic groups, ABC1 groups, with better than average diet
00:01:45	and exercise regimens. The use of HRT
00:01:48	and decreased incidence of coronary heart disease were coincident effects of a common cause:
00:01:56	that is, the benefits associated with a higher socio-economic status,
00:02:01	rather than a direct cause and effect, as had been supposed. If two events, A
00:02:09	and B, are correlated, then the relationship between them could be expressed as follows:
00:02:15	A causes B, a direct causation; B causes A, reverse causation;
00:02:21	A and B are consequences of a common cause, but do not cause each other; A causes B
00:02:27	and B causes A, bidirectional causation; A causes C, which causes B, indirect causation;
00:02:35	or, in fact, there is no connection between A and B, and the correlation is a coincidence.
00:02:41	Therefore, knowing that A and B are correlated does not enable you to confirm
00:02:47	the existence or direction of a cause and effect. The response variable is the focus
00:02:56	of a question in a study or experiment. An explanatory variable is one that explains
00:03:04	changes in the response variable. If you suspect that X causes Y,
00:03:10	then there could be a spurious relationship if, in reality, X and Y are both caused by Z.
00:03:17	For example, when examining a city's ice cream sales, it seems that sales are higher when the rate of drowning
00:03:25	in the city swimming pools is highest. However, to say that ice cream sales cause drowning,
00:03:33	or vice versa, implies a spurious relationship between those two variables.
00:03:39	In reality, a heatwave may have caused both. The heatwave is an example of a lurking variable.
00:03:47	A lurking variable is one whose effects on the response variable cannot be distinguished
00:03:53	from one or more of the explanatory variables in the study, and it's not considered in the design of the study.

00:04:03	A confounding variable is one whose effects on the response variable cannot be distinguished
00:04:09	from one or more of the explanatory variables in the study. The difference between lurking
00:04:18	and confounding variables lies in their inclusion in the study.
00:04:25	In experimental research, spurious relationships can often be identified by controlling for other factors,
00:04:33	including those that have been theoretically identified as possible confounding factors.
00:04:41	For example, consider a scientist trying to determine whether a new drug kills bacteria.
00:04:48	When the drug is applied to a bacterial culture, the bacteria die.
00:04:54	To help rule out the presence of a confounding variable, another culture is subjected to conditions
00:05:01	that are identical to those facing the first culture. However, the drug is not applied to the second culture.
00:05:10	This is called the control. If there is an unseen confounding factor,
00:05:16	the control culture will also die. However, if the control culture does not die,
00:05:22	then the new drug is responsible for killing the bacteria. To summarize: Correlation is a statistical measure
00:05:33	expressed as a number that describes the size and direction of a linear relationship
00:05:40	between two variables. Causation indicates that one event is the result
00:05:46	of the occurrence of another: In other words, there is a causal relationship between the two events.
00:05:54	This is also referred to as cause and effect. For A to cause B, we tend to say that, at a minimum,
00:06:02	A must precede B, the two must covary, vary together, and no competing explanation
00:06:09	can better explain the covariance of A and B. Taken alone, however, these three requirements
00:06:16	cannot prove cause. They are necessary, but they're not sufficient.
00:06:22	Lurking and confounding variables can make it difficult to conclude that it was the explanatory variables alone
00:06:32	that affected the observed changes in the response variable. In the next unit, we will look at scatter plots
00:06:40	and line of best fit.

Week 3 Unit 3

00:00:06	Hello, and welcome back to week three, unit three, where we're going to consider scatter plots
00:00:12	and the line of best fit. You've seen in previous lessons
00:00:19	that you can observe trends in data by creating a scatter plot,
00:00:24	which is a two-dimensional graph of y versus x, where x and y are two of the quantities in your data set.
00:00:34	What you hope to see in a scatter plot is the relationship between x and y,
00:00:39	in which the value of quantity y, which is called the target or dependent variable,
00:00:46	in some way, depends on the value of quantity x, which is called the explanatory or independent variable.
00:00:55	This example plot indicates there's an underlying relationship between x and y.
00:01:04	A scatter plot can be used to show the relationship between two variables.
00:01:09	So in this example, the underlying trend in the data is a straight line,
00:01:15	so the relationship between x and y is linear. You could eyeball the graph,
00:01:21	and with a pencil and a straightedge or ruler, sketch in the line where it appears to fit the data,
00:01:29	estimate its slope and the y-intercept, and be satisfied that your estimate
00:01:35	is close enough to meet your needs. But if you are looking for something
00:01:40	a little bit more accurate, you can obtain the mathematical definition
00:01:45	of the line that most accurately represents the data. This is called line of best fit.
00:01:54	The line of best fit is the best possible straight line that fits the data.
00:02:00	Sometimes the line of best fit is referred to as the trend line.
00:02:05	It's the line where the sum of the squares of the residuals, which are the errors between the individual data values
00:02:14	and the line, is at its minimum. The slope and the y-intercept are two numbers
00:02:21	needed to define the equation of a line, which is y equals mx plus b.
00:02:26	So there are two formulas that you need to define the line of best fit,
00:02:31	for the slope, m, and the intercept, b, and these, of course, are given on the slide.
00:02:37	So let's return to the temperature and ice cream example. The scatter plot enables you to visualize the relationship.
00:02:48	The first step is to calculate the mean of the x, which is x bar,
00:02:52	by adding up all of the x values and dividing by n, the number of observations.
00:02:58	Step two is to calculate the mean of y, which is called y bar.
00:03:02	Again, it's calculated just like you did for x. Step three is to subtract the mean of x
00:03:10	from every x value, and the mean of y from every y value.
00:03:15	Step four is to calculate x minus x bar times y minus y bar, which is the penultimate column
00:03:21	in the table, and x minus x bar squared, the last column in the table,
00:03:27	for each of the different values. Step five is to calculate the slope.
00:03:32	And step six is to calculate the y-intercept. So this will give you the equation
00:03:38	that you're seeing, which is the line of best fit. The trend line and equation of the line
00:03:45	can be added to the scatter plot in Microsoft Excel, as you can see here.
00:03:51	So to summarize: A scatter plot can be used
00:03:55	to show the relationship between two variables. The line of best fit is the line
00:04:01	that describes that relationship between the two variables,
00:04:04	where the sum of the squares of the residual errors between the individual data values

00:04:12	and the line is at its minimum. Therefore, it's the best straight line that fits the data.
00:04:20	The slope, m, and the y-intercept, b, are the two values that are needed to define
00:04:27	the equation of a straight line, y equals mx plus b. So in the next unit, we're going to consider linear regression.

Week 3 Unit 4

00:00:06	Hello, and welcome back to week three, unit four, where we're going to be considering linear regression.
00:00:15	Linear regression is an approach to modeling the linear relationship
00:00:19	between a target variable, also referred to as the dependent variable,
00:00:25	and one or more explanatory variables, known as independent variables.
00:00:30	The case of one explanatory variable is called simple linear regression.
00:00:37	For more than one explanatory variable, the process is called multiple linear regression.
00:00:46	Most applications of linear regression fall into one of the following two broad categories.
00:00:54	Firstly, prediction or forecasting. Linear regression can be used to fit a predictive model
00:01:01	to an observed set of data values of the target and explanatory variables.
00:01:08	After developing the model, when additional values of the explanatory variables
00:01:13	are collected without an accompanying target value, then the model can be used to make a prediction
00:01:21	of the target values. Secondly, explaining variation in the target variables
00:01:27	that can be attributed to variation in the explanatory variables.
00:01:35	Linear regression analysis can be applied to quantify the strength of the relationship
00:01:40	between the targets and the explanatory variables. It can be used to determine if some explanatory variables
00:01:50	have no linear relationship with the target variable. Linear regression models are often
00:02:00	fitted using the least squares approach, although they may also be fitted in other ways.
00:02:09	The best fit from a least squares perspective is to minimize the sum of squared residuals,
00:02:16	when a residual is the difference between an observed value, and the fitted value provided by a model:
00:02:24	in other words, the error. In the ice cream example,
00:02:28	we can calculate the residual, and this is shown in the table.
00:02:33	Ordinary least squares OLS is a type of linear least squares method
00:02:38	for estimating the unknown parameters in a linear regression model.
00:02:44	OLS chooses the parameters of a linear function of a set of explanatory variables
00:02:50	by the principal of least squares. That's minimizing the sum of the squares of the differences
00:02:58	between the observed target or dependent variables, that is, the values of the variables being predicted,
00:03:06	and those predicted by the linear function itself. We can use linear regression to forecast target values.
00:03:17	The relationship between the target and explanatory variables
00:03:20	is modeled using linear predictive functions. The unknown model parameters in the regression equation
00:03:29	are estimated from the data. In the example, the linear regression model
00:03:36	is sales in dollars equals 29.82, temperature and degrees, 145.71.
00:03:42	Therefore, if the temperature tomorrow is forecast to be 20 degrees,
00:03:46	then the store can expect to sell 20.82 times 20 is 145.71. Over $450 worth of ice cream.
00:03:56	Least squares works by making the total of the square of the residuals as small as possible.
00:04:04	The straight line – that is, the regression line – minimizes the sum of squared residuals.
00:04:11	In the graph on the right, you can also see how an unusually high or low value,
00:04:16	an outlier, will have a high influence on the model, as it will pull the regression line towards it.
00:04:28	Standard linear regression models with standard estimation techniques

00:04:33	make a number of assumptions about the predictor variables, the response variables, and their relationship.
00:04:44	The main assumptions are usually listed as follows. Firstly, there is a linear relationship
00:04:49	between the explanatory and target variables. Secondly, there is no or low multicollinearity.
00:04:57	This is correlation between the explanatory variables. There's also no auto-correlation.
00:05:04	This occurs when the residuals are not independent from each other.
00:05:09	In other words, when the value of y(x+1) is not independent from the value of y(x),
00:05:18	the data is homoscedastic, meaning the residuals are equal across the regression line.
00:05:24	Numerous extensions and advances to the basic regression approach
00:05:29	have been developed that allow each of these assumptions to be relaxed – in other words, reduced to a weaker form –
00:05:37	and in some cases, eliminated entirely. Please note that there are many articles
00:05:44	relating to this subject on the internet. And some list a number of additional assumptions.
00:05:54	Linear regression requires the relationship between the explanatory and target variables
00:05:59	to be linear. This assumption addresses the functional form
00:06:05	of the model. The regression model is linear when all terms
00:06:10	in the model are either the constant or a parameter multiplied by an explanatory variable.
00:06:18	You build the model equation only by adding the terms together.
00:06:23	This rule constrains the model to one type shown in the slide.
00:06:29	In the equation, Y is the target variable you're trying to predict.
00:06:35	The X, X1 to Xk, are the explanatory variables you're using
00:06:39	to predict the target. The betas, β1 to βk are the coefficients
00:06:44	or multipliers that describe the size of the effect the explanatory variables are having on the target Y.
00:06:52	These are the parameters that ordinary least squares process estimates.
00:06:58	β0 is the value Y is predicted to have when all the explanatory variables are equal to zero,
00:07:05	the Y intercept. Epsilon is the random error.
00:07:09	The linearity assumption can be tested with scatter plots. You can see there are three simple examples
00:07:17	showing a strong linear relationship, a weaker linear relationship
00:07:21	and no linear relationship at all between the target Y variable
00:07:25	and the explanatory X variable. Also please be aware that linear models
00:07:32	can model curvature by including non-linear explanatory variables,
00:07:38	such as polynomials and exponential functions. Multicollinearity occurs when the explanatory variables
00:07:49	are highly correlated with each other. If multicollinearity is found in the data,
00:07:57	the simplest way to address the problem is to remove one of the correlated variables.
00:08:03	However, there are a range of more sophisticated techniques available.
00:08:13	Linear regression analysis requires that there is little or no autocorrelation in the residuals.
00:08:20	This means that the error terms must be uncorrelated so that one observation of the error term
00:08:27	should not predict the next. Autocorrelation occurs when the error terms
00:08:34	are not independent from each other. In other words, when the value of y(x+1)
00:08:41	is not independent from the value of y(x). For instance, if the error for one observation is positive,
00:08:50	and that increases the probability that the following error is positive,
00:08:55	then there is a positive correlation. You can assess if this assumption is violated

00:09:02	by graphing the residuals in the order that the data were collected,
00:09:07	you hope to see a randomness in the plot. While a scatter plot allows you
00:09:13	to check for autocorrelation, you can also test the linear regression model
00:09:18	for autocorrelation with the Derbin Watson test. The variance of the errors should be consistent
00:09:29	for all observations. This means that the variance does not change
00:09:34	for each observation or for a range of observations. The scatter plot is a good way to check
00:09:43	whether the data are homoscedastic, which simply means that the residuals are equal
00:09:49	across the regression. You can check this assumption by plotting the residuals
00:09:56	against the fitted values. Heteroscedastic appears as a cone shape,
00:10:02	where the spread of the residuals increases in one direction.
00:10:06	The scatter plot on the right shows an example of data, though not homoscedastic,
00:10:13	they're heteroscedastic. Please also note that when assumption three,
00:10:18	no auto correlation that is, and four, homoscedastic, are both true,
00:10:22	statisticians say that the error term is independent and identically distributed.
00:10:29	They refer to this as spherical errors. Although there was no assumption about the distribution
00:10:38	of the explanatory variables, it is good practice to examine and understand the data
00:10:45	before you build a model. You can check the distribution visually,
00:10:50	for example with a histogram, a box plot, or a Q-Q plot. A Q-Q, Quantile-Quantile, plot is a plot
00:10:58	of the quantiles of two distributions against each other, or a plot based on estimates of the quantiles.
00:11:06	The pattern of points in the plot is used to compare two distributions.
00:11:11	So to test for normality, you could use it to compare the distribution
00:11:15	of an explanatory variable to the theoretical normal distribution.
00:11:22	If the two distributions being compared are similar, the points in the Q-Q plot
00:11:28	will approximately lie on the line y = x. However, if the distributions are linearly related,
00:11:36	the points in the plot will approximately lie on a line. But not necessarily on the line y = x.
00:11:45	To summarize: Linear regression is an approach to modeling the linear relationship
00:11:51	between a target variable and one or more explanatory variables.
00:11:58	Simple linear regression has one explanatory variable, and multiple linear regression
00:12:03	has more than one explanatory variable. In summary, your linear regression model
00:12:11	should produce residuals that have a mean of zero, have a constant variance and are not correlated
00:12:19	with themselves or other variables. If these assumptions are true,
00:12:25	then the ordinary least squares regression procedure will create the best possible estimates.
00:12:34	In the next unit, we will consider how to interpret the results from a regression.

Week 3 Unit 5

00:00:06	Hello, and welcome back to the fifth and final unit of week three, where we're going to consider
00:00:13	how to interpret the results of a linear regression. Regression analysis is used to produce an equation
00:00:21	that will predict a target variable using one or more explanatory variables.
00:00:28	This equation has the form shown in the slide, where Y is the target variable you are trying to predict.
00:00:37	The Xs, X1 to Xk, are the explanatory variables you are using to predict the target.
00:00:44	And the betas, β1 to βk are the coefficients that describe the size of the effect
00:00:50	the explanatory variables are having on the target variable Y.
00:00:55	These are the parameters that the ordinary least squares process estimates.
00:01:01	β0 is the value Y is predicted to have when all the explanatory variables are equal to zero,
00:01:08	it's the Y intercept. Epsilon is the random error.
00:01:15	This is a simple example using the R package in RStudio. You can install this free-to-use software if you want to
00:01:24	and explore the tremendous functionality that's available. There are many user guides you can download,
00:01:33	so you can start to use the package. We'll make the data available for you.
00:01:40	The data shows the price obtained for a number of properties.
00:01:45	This is called the target variable or the dependent variable.
00:01:50	The data also includes the number of bedrooms, bathrooms, the size of the living area in square feet,
00:01:58	and of the overall lot size, and the number of floors in the house.
00:02:03	These are the explanatory variables or the independent variables.
00:02:07	Initially, you could run a summary to look at the range, the quartiles, median, and mean of the variables.
00:02:16	This example data has been provided for you if you would like to repeat this exercise.
00:02:22	There are links to download RStudio in the slide. The data frame is called house_train.
00:02:31	The data summary uses the R command "summary". The number of rows and columns
00:02:37	is given by the R command "dim". This initial analysis can often indicate
00:02:43	interesting features in the data. In this example, there is a similarity in the data summary
00:02:49	for the explanatory variables "bedrooms" and "bathrooms", and this will require further investigation.
00:03:01	You might also want to visualize the data. This is a simple visualization using the R command plot.
00:03:09	Here you can see the basic relationship between each of the variables in the data.
00:03:16	You will notice that there is a strong correlation between the number of bedrooms and bathrooms.
00:03:26	This might indicate that there is a problem in the data, and this should trigger some further analysis
00:03:34	so that you can discover what's happening in the data. You can fit a multiple linear regression model
00:03:42	with price as the target variable, and the other variables as the explanatory variables.
00:03:50	The R command is given here in the slide. This is the output
00:03:56	when you build the linear regression model. The coefficient terms can be represented
00:04:03	in the equation shown in the slide. Note that there is no coefficient for bathrooms X2,

00:04:12 you'll explore why this is later in this lesson. You will need to test

00:04:18 whether the explanatory variables in the model collectively have an effect on the target variable.

00:04:26 To do this, you test what is called the null hypothesis, which is represented as H0,

00:04:33 against what is called the alternative hypothesis H1. This is a simple test.

00:04:40 If there is a significant linear relationship between the explanatory variables, the Xs,

00:04:47 and the target variable Y, then the slope will not equal zero.

00:04:52 The null hypothesis states that the slope is equal to zero, and the alternative hypothesis states

00:04:59 that the slope is not equal to zero. When you test the null hypothesis

00:05:05 it simply means you are assessing the probability that there is no relationship

00:05:11 between the explanatory variables and the target variable. You then either accept or reject this null hypothesis.

00:05:21 Therefore, the hypotheses read like this: H0 is where beta one equals beta two equals zero.

00:05:29 And H1 is where at least one beta value is not zero. The F-statistic is the test statistic

00:05:37 used to decide whether the model as a whole has statistically significant predictive capability.

00:05:44 The F-statistic is the ratio of the mean regression sum of squares

00:05:49 divided by the mean error sum of squares. Its value will range from zero up to a large number.

00:05:58 And the value represents the probability that the null hypothesis for the full model is true:

00:06:04 in other words, that all of the regression coefficients are zero. In general, if your calculated F-statistic

00:06:12 of your regression model is larger than the threshold critical F-statistic

00:06:18 given by F-distribution tables or calculated in your statistical software,

00:06:25 you can reject the null hypothesis. The p-value is determined by the F-statistic

00:06:32 and is the probability your results could have happened by chance.

00:06:37 If the p-value is less than the alpha level, commonly set at 0.05, you can reject the null hypothesis.

00:06:46 In our example, the output shows that F equals with a very small P value,

00:06:53 and that indicates you can reject the null hypothesis that the explanatory variables collectively

00:06:59 have no effect on the target variable. If you can reject this null hypothesis,

00:07:07 you continue by testing whether the individual regression coefficients are significant

00:07:14 while controlling for the other variables in the model. The p-value for each individual variable

00:07:22 tests the null hypothesis that the coefficient is equal to zero:

00:07:27 that is, there is no effect. A low p-value less than 0.05 indicates

00:07:33 that you can reject the null hypothesis. So basically, the explanatory variable

00:07:39 that has a low p-value is likely to be a meaningful addition to your model

00:07:45 because changes in the variable's value are related to changes in the target variable.

00:07:51 Conversely, a larger insignificant p-value suggests that changes in the explanatory variable

00:07:59 are not associated with changes in the target. You use the coefficient p-values

00:08:06 to determine which terms to keep in the regression model. The standard errors of the coefficients

00:08:16 are the estimated standard deviations of the errors in estimating them.

00:08:22 The larger the standard error of the coefficient estimate, the less precise the measurement of the coefficient.

00:08:31 The t value is the coefficient estimate divided by its standard error.

00:08:38 In the output shown in the slide, the first column gives the estimated value

00:08:44 for each coefficient. The second column gives the standard error,

00:08:49	and the third column gives the t value. The output compares the t-statistic of the variable
00:08:57	with values in the student's t-distribution to determine the p-value.
00:09:02	The student's t-distribution describes how the mean of a sample with a certain number of observations
00:09:10	is expected to behave. R-squared is a statistical measure
00:09:16	of how close the data are to the fitted regression. It's also known as the coefficient of determination
00:09:24	or the coefficient of multiple determination for multiple regression.
00:09:29	R-squared is the percentage of the target variable variation that is explained by a linear model.
00:09:38	It's equal to the explained variation divided by the total variation.
00:09:43	It's always between 0 and 100% if there's an intercept value.
00:09:49	0% indicates that the model explains none of the variability of the target data around its mean.
00:09:56	100% indicates that the model explains all the variability of the target data around its mean.
00:10:04	However, there are two problems with R- squared. Firstly, every time you add an explanatory variable
00:10:11	to a model, the R-squared increases, even if due to chance alone.
00:10:16	It never decreases. Therefore, a model with more terms
00:10:21	may appear to have a better fit simply because it has more terms.
00:10:26	Secondly, if a model has too many explanatory variables, it begins to model the random noise in the data.
00:10:34	This is known as overfitting the model, and it produces misleadingly high R-squared values,
00:10:41	but the model is basically unable to make accurate predictions.
00:10:47	The adjusted R-squared is a modified version of R-squared that has been adjusted
00:10:53	for the number of explanatory variables in the model. The adjusted R-squared increases only
00:11:02	if the new term improves the model more than would be expected by chance.
00:11:08	It decreases when an explanatory variable improves the model by less than expected by chance.
00:11:16	Note that the adjusted R-squared can be negative, and it is always lower than the R-squared.
00:11:26	Regression coefficients represent the mean change in the target variable for one unit of change
00:11:32	in the explanatory variable, while holding other predictors in the model constant.
00:11:39	This statistical control that regression provides is important because it isolates the role
00:11:47	of one variable from all of the others in the model. For example, the equation shows
00:11:53	that the coefficient for the explanatory variable square foot of living is 3.281e squared.
00:12:01	The coefficient indicates that for every additional square foot in living space,
00:12:07	you can expect the target price to increase by an average of 3.281e squared dollars.
00:12:16	You may have noticed that there is an error identified in the output:
00:12:21	One coefficient is not defined because of singularities. This occurs because two of the explanatory variables
00:12:29	are perfectly correlated: bedrooms and bathrooms. So the coefficient cannot be specified.
00:12:37	You can see this when you produce the correlation matrix using the R command given here.
00:12:46	Therefore, the explanatory variable "bathrooms" is omitted from the model.
00:12:54	So to summarize: Multiple linear regression is used to describe data
00:13:01	and to explain the relationship between one target variable
00:13:05	and two or more explanatory variables. The analysis requires you to analyze
00:13:13	the correlation and directionality of the data, train the model, and then evaluate

00:13:20	the validity and usefulness of the model. This unit introduced you to some of the results
00:13:27	that are generated that will help you evaluate your model. You must assess these very carefully
00:13:34	so that you can be sure that your model is valid. Remember that it's also very important
00:13:42	that you check and confirm that the basic assumptions for linear regression
00:13:48	that were discussed in the previous unit hold true. With this, I'd like to close the third week.
00:13:56	I hope you enjoyed these units of this course, and we are happy to get in touch with you
00:14:02	in our discussion forum if you have any content-related questions.
00:14:08	Now, we will wish you all the best for the weekly assignment and see you next week,
00:14:14	where we will introduce probability and Bayes' Theorem.

www.sap.com/contactsap

THE BEST RUN