Hello, I am Professor Sridhar Narasimhan of the Scheller College of Business at

Georgia Tech. I'll be teaching this module on linear regression in this course

on data analytics in business.

[BLANK\_AUDIO]

There are several lessons in this module.

The first lesson is on the various steps involved in doing regression analysis,

which is a very iterative process. The next lesson is

an example real estate data set that we will use extensively in this module.

Lesson C is on the commonly used notation in regression analysis.

Lesson D is on how R squared and adjusted R squared are defined.

Lesson E is how to use R to do simple or one predictor regression.

Lessons F and G extend that to multiple variable regression and

show how R squared and adjusted R squared are calculated.

Finally, lesson H is about using regression to do prediction.

[BLANK\_AUDIO]

So there are various steps in regression analysis and

regression is an iterative process.

Step one is knowing what problem you are addressing.

The second step is to have a clear understanding of the type of

analytics that you plan to do. It's either diagnostic,

that is understanding why, predictive, what'll happen or prescriptive,

what to do to make something happen? This in turn affects step number three,

which is the selection of the relevant variables. Step four is determining V,

you're going to get the data. Internally within the organisation,

externally, are you going to purchase data, run experiments etc.

The next step is the choice of the fitting method. In this module,

we use ordinary least squares. In step six,

there is the issue of selecting software and

using it to fit a model using the data that we have. And

then we get estimates of the parameter. Step seven is about diagnostics.

Step eight is to reflect on the model and improve it if necessary.

Finally, the model is ready for its intended use. Nope,

that building a regression model is an iterative process as you

follow the steps in the previous slides, you may have to add

constant variables and add interaction terms. As the modeler you

have to be satisfied that you have taken adequate care while developing your model.

[BLANK\_AUDIO]

Nearly every business function is a candidate for regression analysis.

I list a few examples to illustrate the broad variety of applications.

The first example has to do with pricing of use cars. And

how it is related to it's mileage, it's age, condition, etc.

The second example has to do with the effect of advertising on sales.

The code example addresses an operations issue. The fourth

example is from e-commerce. What is the contribution of a product's ratings and

price on the likelihood of that product being added to a shopping cart?

The fifth example is how to set starting salaries for

new employees based on their work experience and years of education.

The sixth example is setting the sales price of a house using it's size,

the number of bedrooms and it's location in a city.

The second last example is from finance, will a customer default on their loan?

And how is that associated with his or her credit balance, income,

age. The last example has to do with customer churn and

how that is impacted by our customers length of contract and age.

[BLANK\_AUDIO]

Before we go to the next lesson, try take a moment to answer this quiz.

Could a variable, say price be either a dependent or an independent variable?

Answer true, this depends on the purpose of your model. See where price

appears in examples number one, and number four in the previous slide.

Next question, a variable that takes binary values for example,

pass or fail, or true or false cannot be a dependent variable.

[BLANK\_AUDIO]

The answer is false. We do use 0/1 dependent variables in logistic regression

in a different module. Example number seven in

the previous slide is one instance where we have

to use binary dependent variables. [SOUND]

In this lesson, we're going to introduce a real estate data set that

we will use extensively iIn this lesson. Assume that you need to sell your house,

you want to predict the listing price based on how

other houses are listed in the market. How would you approach this task?

[BLANK\_AUDIO]

A typical approach is to ask realtors. And realtors often will use comparable.

That's recent sales of houses in your neighborhood,

usually three to five houses and somehow come up with a suggested sale price.

However, you want to be more analytical in your approach.

You have access to recent actual home sales in your city.

You'd like to know what are the impacts of factors such as lot size,

number of bedrooms, number of bathrooms, ec on the price. Could you

use linear regression to help you get a better estimate of the listing price?

[BLANK\_AUDIO]

Conveniently for us, there is a dataset in the R package Ecdat.

This dataset has a cross section of 546 home

prices In the city of Windsor in Canada. So

this data set is a sample of the real estate transactions in one city, and

it's called the housing data frame. It's a cross section,

as I mentioned of home prices in one city in Canada. Alternatively,

if you could collect house prices from websites or scrape them from the web,

that could also be a good source of data for something like this.

[BLANK\_AUDIO]

So looking at this housing dataset, we see that there are a whole bunch of variables.

Some of quantitative and there are about six of them that are quantitative and

six of them that are categorical. So categorical variables

are also called factors in R. Note that 1 is no and

2 is yes is how the factor variables are coded in R

[BLANK\_AUDIO]

Taking a look at these first ten records in housing,

you'll see that the first row of data has a house whose price is $42,000 and

a lot size of 5850 Square feet. It has three bedrooms, one bathroom,

and is two stories tall. The house has a driveway. It does not have a recroom.

It has a full basement. It has no gas heated hot water system.

It has no air conditioning. Has a one car garage. And

is not in a preferred neighborhood of the city.

[BLANK\_AUDIO]

One important thing to do before you start doing regression is to do

exploratory data analysis and histograms help with that.

So it always helps to do histograms of some of the key variables.

So this is the histogram of prices which is skewed to the right.

The median value of home sold is $62,000. And

the main value is 68,122.

We've also done the histogram of the distribution of lotsizes, and

this is also skewed to the right. The median value here is 4600 square feet,

and the mean value is 5,150 square feet.

It's always useful to do a correlation metrics of the key variables, prior to

building a linear regression model. As you can see, the correlation between price and

lotsize is interesting, and will help develop a simple regression model.

This matrix also shows the density functions along the diagonal and

scatter plots of the two variables in the lower triangle.

[BLANK\_AUDIO]

In this scatter plot, the y-axis denotes home prices and

the x-axis shows the lotsize. A simple linear regression line

seems a reasonable place to start doing R regression analysis. So please

take a look at this quiz before you go to the next lesson in this module. The mean

of a variable that has a right-skewed distribution is smaller than the median.

[BLANK\_AUDIO]

Answer is false.

The correlation coefficient can capture the strength of both linear and

non-linear relationships. The answer is false.

[MUSIC]

In this lesson we'll go over the commonly used notation in regression analysis.

Note that explanatory variables are also called independent variables or

features. Dependent variables are also called response variables.

[BLANK\_AUDIO]

So our first notation is the observations of record in a data set and

this is a sample of the population. The x

variables refer to the n observations of the p explanatory variables.

And the n observations of the dependent variable are denoted by y.

And the mean of the dependent variable is y bar. And

the mean value of each of the x kth explanatory variables is the xk-bar.

[BLANK\_AUDIO]

Continuing with the notation, we have the betas as the parameters of the regression

line, if you have the entire population. Most often,

we don't have the entire population, we have a sample. And the estimates of

the beta parameters obtained by fitting the regression line are called the b's,

b0 through bp. Please note that

a lot of textbook will use beta instead of b for the estimated regression parameters.

The error term for the ith observation in the population is epsilon i.

The error term for the ith observation in sample is ei. The estimated value of y for

the ith observation is called y hat i. This is obtained by fitting the regression

line at xi. Simple linear regression is also called bivariate regression,

since there's only the y variable and one x variable involved. So,

for this Housing dataset, we observe the data, which is a sample, and

we want to build a model for the population, which is a valid relation.

And as described earlier, epsilon and i are the independent and

identically distributed random variables. Which are normally disturbed at mean zero

and standard deviation sigma. However, we don't have the values of beta or

sigma, so we need to estimate them based on the sample in Housing dataset.

So using the sample we're going to build

a model Yi = b0 + b1 times X1 plus ei.

[BLANK\_AUDIO]

So here's the regression line, which is shown in this diagram.

The slope of the regression line is beta 1 and the intercept

is beta 0. Again, this is the case if you have the data for the entire population.

The errors are assumed to be normally distributed with mean of zero and

standard deviation sigma across all the values of x.

This constant sigma is also called homoscedasticity.

Notice the y variable is price and the x variable is lotsize.

The estimates of slope and intercept depend on the sample being used.

So, when we are trying to regress price on lotsize with sample A,

we may get this regression line, and with sample B,

we might get a different regression line. So,

there are different points drawn in sample B.

The idea is that if we draw 100 such samples, then 95 of them

should contain the population slope within the confidence interval.

So how do we draw these regression lines? So

we're gonna use ordinary least squares to fit this line.

Estimated value of y equals b0 plus b1 times x.

So we want to determine the slope b1 and the intercept b0.

[BLANK\_AUDIO]

So what do we do? There are various steps. First, we need to compute y-bar,

which is the mean value of y. So that's what regression does first.

[BLANK\_AUDIO]

And then, we want to figure out the slope and the intersect.

So, b0 is the intercept and b1 is the slope.

So, what does ordinary least squares regression actually do?

Given this type of data. So how is this line fitted in OLS?

So consider a point xi as shown in this figure.

For this point we have the observed value of yi.

We need to find a regression line, which is y hat equals b0 plus b1x.

Obviously there are lots of lines that can be used to fit a regression line.

However in OLS, we're going to fit a line that minimizes something

called the sum of squared errors, or SSE. But,

to get there, we need to understand what error means in this context.

At this value of xi, we have a predicted value of y hat i.

Now we need to define total deviation, okay?

So total deviation is the difference between the observed value yi and

the mean value of y, shown in blue out here.

[BLANK\_AUDIO]

The explained deviation is the predicted value of Y. Minus the mean value of y.

[BLANK\_AUDIO]

That's shown in red out here.

[BLANK\_AUDIO]

The error term or the residual is the difference between the predicted value and

the actual value for the i hat observation.

[BLANK\_AUDIO]

That's shown in green out here. So ordinary least

squares minimize the sum of the square of these errors across all the observations.

[BLANK\_AUDIO]

So before we go on to the next lesson, have a go at this quiz.

The total deviation at observation (xi,

yi) is yi minus y bar. The answer is true.

[BLANK\_AUDIO]

In ordinary least squares, the estimates of slope and

intercept do not depend on the sample being used.

[BLANK\_AUDIO]

The answer is false.

[SOUND]

[BLANK\_AUDIO]

In this lesson, we define R squared and adjusted R squared. These

are one set of measures used to judge the suitability of a regression model.

As I said earlier, OLS determines the regression line that minimizes

the sum of squared errors. So if you look at that green expression,

it's going to square it, and sum it across all the observations. And

the line that has the minimum such sum is the line that is fit in OLS.

So summing the deviations, we get this expression. The total sum of squares,

that is the difference between each observation's y value and

the average of y value, square it, and sum it, over all observations.

SST is equal to the sum of squared errors, SSE,

plus the sum of squares regression, which is SSR.

[BLANK\_AUDIO]

We defined R squared and adjusted R squared, and

these are based on those expressions in the previous slide.

R squared never decreases as you keep adding variables. R squared is also

called the coefficient of determination. It's a measure of the overall strength

of the relationship between the dependent variable Y and the independent variables.

So R squared is given by this expression out here,

and it's the explained deviation over the total deviation.

We also have R squared telling us how much of

the variation in Y has been explained by our linear regression model.

Adjusted R squared ensures that you never account for the inflation in the number

of predictors since R squared never decreases as you keep adding variables,

we use adjusted R squared. Especially in business applications,

we want to mainly use predictors that are useful. So adjusted R squared

adds a penalty for the number of independent variables, p.

So adjusted R squared equals this expression out here. And

here we're gonna provide three examples to better understand what

R squared actually means, and what is it actually capturing.

[BLANK\_AUDIO]

In the first example, where R squared = 1,

knowing the value of X, we can know all the variation in Y.

So the predicted value equals the actual value, and

is on the regression line. So sum of squared errors equals zero.

In the second example where R squared equals zero, X accounts for

none of the variation in Y. In the third example,

R squared equals .75. Most of the Y variation can

be explained from predicting Y from the X values.

[BLANK\_AUDIO]

So here's a quiz before we go on to the next lesson.

[BLANK\_AUDIO]

So r squared equals zero implies that x values account for

all the variation in the y values. Answer, false.

R squared equals zero implies that X values account for none of the variation

in the Y values. Next question, R squared can take any value,

from negative infinity to positive infinity. The answer?

False, it can take on values between 0 and 1.

[MUSIC]

In this lesson, we do simple regression.

That is there's only one predictor variable. And we're gonna be using R.

[BLANK\_AUDIO]

We're using the same housing data set that we saw earlier. And

the function in R is called LM, and we're gonna use LM to fit a linear model

using ordinary least squares. The LM, the R text is shown here.

And when you run the linear model in R,

you get these results shown here in the slide. And

the minimum value through the maximum value of the residuals

are shown first in green, highlighted in this particular slide.

[BLANK\_AUDIO]

The output of running that regression model in R,

we get the values of the coefficients,

the intercept is the value of b0 and

is equal to 34140. The coefficient of lotsize is b1 and

its value is equal to 6.599. And then the standard errors,

think of them as the standard deviation of their respective perimeter estimates.

[BLANK\_AUDIO]

So how do we think of these other values in the output?

The T-value is obtained by dividing the coefficient estimate by the standard

error. And the null hypothesis is that the parameter is zero. And the P value

is the probability of finding a T value of this size if the null hypothesis is true.

In this case, given the very small values of the P values,

we reject the null hypothesis for each of the two parameters, D0 and B1.

The H statistic, shown here in this slide,

is the probability that B1 equals 0 and has its value is

219.1 with the degrees of freedom shown there.

[BLANK\_AUDIO]

So how do you interpret these coefficients?

[BLANK\_AUDIO]

So these are the values we saw of the coefficients. So

what's the meaning of b0 = 34,140? So

these coefficients have to be interpreted with this idea of ceteris paribus.

That is, meaning all else is the same or is held constant. So

the value of b0 is the intercept of the regression line with the y-axis

when the lotsize is zero. Not very useful for us. b1 is 6.599,

and essentially that means an increase of 1,000

square feet is associated with an increase in the sale

price of a house by $6,599, keeping all else constant.

[BLANK\_AUDIO]

The regression output from the LM model also gives us a nova table,

as shown, and we can use that to validate the R squared and

adjusted R squared calculations. So

the some of squares of regression is 1.1156 and

e to the eleven, and the values of SSE is 2.77 times e to the eleven,

and so on. Right? And SST is the sum of SSR plus SSE.

We have seen the definition of R squared earlier, and the definition of adjusted R

squared. We'll next show how these values are derived from their definitions.

So the regression output R squared and adjusted R squared.

Let's see how those are derived. So we know the values for SSR, SSE,

and you can total them up to get SST, and the R squared,

multiple R squared value that you get by running the LM model in R is 0.2871.

And the adjusted R squared is 0.2858. So to compute R squared,

we apply the formula R squared equals SSR over SST and you get a value equals

0.2871, which is what we saw in the regression output from R.

Note in the simple regression, square root of R squared is 0.536.

Which is the correlation coefficient between price and

lotsize. Adjusted R square is a slighty different formula, and

you get a value equals, plug in the right numbers, and you get 0.2858.

Which is the value we saw in the output of the regression.

[BLANK\_AUDIO]

The one way to test if the regression model

is statistically significant is to use the F test.

The F statistic is the ratio of the means regressions sum of squares

divided by the mean error sum of squares and it's given by the formulas here.

Its value will range from 0 to an arbitrarily large number. And

in this particular case, the F statistic is 219.1, and

this is what R had produced, also. So in this case, H0 is rejected.

[BLANK\_AUDIO]

In this lesson, we extend the analysis to do multiple regression using R.

[BLANK\_AUDIO]

We have multiple linear regression, in this case with p explanatory variables.

So we'll have b0, which is the intercept, and then b1 through bp. And

these are all estimates of beta zeroes through beta p. And

we have to predict for the Wyatt at an observation xi.

We have the expression out there if you knew the b's, and

the residual is given that, yi minus y hat. So again,

our goal is to choose those b0, b1 through bp parameters to

minimize the sum of squared errors given by the expression here.

Using the Housing Dataset in the Ecdat package,

let's add bedrooms to the analysis. So

the first record has three bedrooms, the second has two bedrooms and so on.

We need to find out if the number of bedrooms impacts the price of

a house. Earlier, we had seen the scatter plot on the left,

does the scatter plot on the right makes sense? Yeah, to some extent.

Since as the number of bedrooms increases, the house price should typically increase.

So here's the regression output for multiple regression,

where price is regressed on lot size and bedrooms, using the same housing dataset.

The null hypothesis or H0 is a parameter zero,

and H1 is the parameter is non-zero. So this is the output that you get for

the intercept, lot size and bedrooms coefficients. And

these are the standard errors of these three coefficients shown here. And

then the T values and the P values. The coefficients of lot size and

bedrooms are statistically significant. And then the S statistic which is

the last line of this artwork is the probability that b1 equals b2 equals zero.

That is what this low parameters are equal to zero. So how do you interpret

the coefficients in this multiple regression says,

these are values that we got. So what does b0 equals 5613?

This is the intercept of the regression line with the y-axis,

when all x's are zero. Again, this is not very useful. b1 = 6.053,

so an increase of 1,000 square foot in lot size is

associated with an increase in the sale price of a house by $6,053,

keeping all else constant. b2 = 10570,

an additional bedroom is associated with an increase of the sale price of

a house by $10,570, keeping all else constant.

[BLANK\_AUDIO]

[SOUND]

In this lesson we are examining the changes to R squared and

adjusted R squared in multiple regression.

[BLANK\_AUDIO]

So again this is the output of the regression,

this is the sum of squares from the Ennova table values SSR and

SSE and SST. SST is always the same, but the SSR and

SSE values have changed. And we get F-statistic values, okay?

And then we have the regression output, R squared and adjusted R squared.

And I've shown the computations for those values. Which is R squared is .3707.

Adjusted R squared is .3679.

From the values of SSR SSA and SST.

So we can do an F test of the overall significance of the model,

that is the b1 and b2 parameters equal to zero jointly is the null hypothesis.

And F-statistic has a value of 159.6. And

the value of probability of F is the probability

that the null hypothesis, the regression model is true. So in this particular case,

since the F-statistic is 159.6 and

the p value is so small, the null hypothesis is rejected.

So we have a model that is statistically significant.

Comparing Simple vs Multiple Regression. For the Simple Regression

we got R-squared of 0.28 and adjusted R-squared of 0.2858.

For the Multiple Regression we got Multiple r squared of .37 and

adjusted R squared of .3679. As you add variables,

r squared will not decrease. Comparing the two models,

N is the number of observations, P is a number of variables. The bigger model,

which has more variables, let's call those number of variables P2. The smaller

model which has fewer variables, let's call that a model with P1 variables.

Then we can use an F-test to see whether model two gives us significantly better

fit to the data. And we have this F statistic here.

[BLANK\_AUDIO]

Please check these quiz questions before going on to the next lesson. In general,

adding more variables decreases the overall R-Square value of the multiple

regression. Answer, false. In the regression output shown below,

a p-value of 2 and

10 to the power -16 means that there's not much evidence for

the coefficient of lotsize to be different from 0. The answer is false.

[BLANK\_AUDIO]

[MUSIC]

In this lesson, we're gonna look at common problems seen in regression.

There are some assumptions in linear regression.

First is, that is the expected value of Y at each value of X,

approximate to a straight line. There's an assumption about the errors.

The error terms for each observation are independently and identically distributed,

or iid normal random variables, each with the mean of zero and

constant variance. That is, we have homoscedasticity. And

we also have some assumptions about predictors. In multiple regression,

the predictor variables are assumed to be linearly independent of one another.

So when we fit linear regression model, we have to be on the lookout for

some common problems that could occur. These problems include,

one, non-linearity of the relationship between the response and

the predictor variable. Two, correlation of error terms.

Three, non-constant variants of error terms. Four,

the presence of outliers in the data. Five, high leverage points.

And finally, multi-colinearity or colinearity.

So is the relationship nonlinear? Is the YX relationship linear?

Or the residuals plot versus fitted plot have no patterns? So

we can check the scatter plots of Y versus each X variable. Is it linear?

Another plot to use is the residuals plot vs fitted values plot,

especially useful in multiple regression. And

in that, we want to see no patterns.

So is the relationship non-linear? If there are concerns,

then can you model non-linear relationships with higher order terms?

So instead of using X, can you use X-squared, for example? Can

you use variance reduction transformation, such as taking a log off a variable?

That will give a better linear fit. Are there outliers or

certain sections of the observations that seem to drive this non-linearity?

Are there any important variables that you left out from your model.

That is, the age or gender of an individual.

So this called the omitted variable problem. Or

maybe was there systematic bias when collecting the data and

you need to redesign the data collection? So

checking residuals helps discover useful insights about your model and

data. So we're going to look at this scatter plot of the response variable

versus the explanatory variable. This is useful to do before you fit a model.

So you can plot Y versus X to identify any pattern. For example, in this case,

the scatter plot below suggests using X squared rather than X as

a predictor. You could also use the residuals versus fitted values plot.

So this plot should be examined after the model is fitted.

The figure on the left-hand side, Case 1, seems to be what we should aim for

in a residuals versus fitted values plot. Case 2 exhibits non-linearities.

The second item to look for in problems in regression is correlation of error terms.

An important assumption in OLS is that error terms are uncorrelated.

If they are not, then we have what's called autocorrelation. So

knowing the value of ei should not have any influence on the magnitude or

size of ei + 1. This property is used to

estimate the standard errors of the parameters of the model.

If there is correlation in the error terms, we have these.

The estimated standard errors will underestimate the true standard errors.

Confidence and prediction intervals will be narrower than they should be.

And p values will be lower than they should be. And we may have a sense

of confidence in the model that is not warranted.

So the Durbin-Watson test is used to detect autocorrelation in a linear model.

Heteroskedasticity, or non-constant error variance,

is something that we should always check. The assumption is that

the spread of the response around the regression line is the same at all levels

of the explanatory variable. That is, we have constant variance or

we have homoscedasticity. You may have non-constant error present,

for example, if the errors increase in size with their fitted values.

So you can detect this with the residuals versus the fitted values plot.

If non-constant error is present, then your hypothesis tests and

confidence intervals can be misleading. If there is heteroskedasticity,

then transformation of the Y variable may be called for,

for example, taking the log of Y or the reciprocal of Y,

etc. If there is heteroscedasticity, then test to reject H zero and

also calculations of confidence intervals can be misleading.

Here's an example of heteroskedasticity. The predicted price is

on the x-axis and the residuals are on the y-axis.

And the residual value seem to be increasing with predicted price.

So here's a quiz, if the scatter plot of Y vs X shows a nonlinear pattern,

then we should not change our linear regression model. Answer, false.

Autocorrelation is the correlation between each of the ei variables.

Answer, true. Heteroskedasticity means having constant error variance.

Answer, false.

[SOUND]

We're continuing with our theme of common problems and

fixes in fitting linear regressions. Outlies,

outlier is a point whose yi value is far from it's predicted yi hat value

One way to visualize outliers is to plot residuals against predicted value of y,

or even better, standardized residuals.

Outliers could occur because of incorrect data recording or

because the phenomenon could very well be non-linear.

So don't assume that an outlier observation should be removed.

It may signal a model deficiency, for example, a missing predictor.

There's also the issue of high leverage points. So in simple regression, look for

observations that have a predictor value outside the normal range of observations.

A point has high leverage if it's deletion by itself or

with 2 or 3 other points causes notable changes in the model.

With many predictors in a model, one could have an observation

that is within the range of each predictor's value but still be unusual.

So here's an example. We have these points, and

one point which seems to fall within the ranges of these

two variables here. This observation at (1.5,

1) is within the range of (-2 to 2) for X1 and

(-2 to 2) for X2. So

one statistic to identify influential points is the Cook's Distance Ci

that measures the difference between the regression coefficients

obtained from the full data and from deleting observation i.

A rule of thumb is to identify points with Ci>1 and label them as highly influential.

There's also a function called plot(model) in R, which is extremely useful,

and you have to do that every time after a regression model is fitted.

This comes with four built-in plots, residual versus fitted,

you can use that to check if residuals have nonlinear patterns. Normal Q-Q,

check if residuals are normally distributed. Scale-location,

check if standardized or square root of standardized residuals are spread equally

along the range of fitted values. And then residuals vs leverage,

that is, use to find influential points, that is any with Ci>1.

So for the simple regression that we did earlier,

I'm going to call that model a.lm where I regressed

price on lotsize, and then I plot a.lm.

If I plot a lot a.lm, I get these four plots,

and all four plots indicate that the model that we

had has some problem, okay? Outliers and influential points, so

in fitting a model to a given body of data,

we'd like to ensure that the fit is not overly determined by one or

few observations, aka outliers. So there are actually two types of outliers,

Y, which is a response outlier, and X is a predictor, outlier or

leverage point. So an outlier has the potential to be influential or

labeled as an influential data point if it unduly influences the regression analysis.

The next few slides will define these two types of outliers on

how to classify the outlier as an influential point.

So outliers in the response y variable,

as stated earlier, have a value that is far from its predicted y hat,

y value. So you'll have a large standardized residual.

So typically, points that have standardized residuals larger than two or

three standard deviations away from the mean are called outliers. If removal

of the outlier causes substantial changes to the regression analysis,

then it is an influential outlier. So detecting these outliers,

one way to visualize and identify them is to plot residuals

against predicted values. So why do outliers occur?

Outliers could occur because of incorrect data recording,

because of real anomalous events measured correctly,

or because the phenomena itself could very well be nonlinear.

Do not assume that an outlier observation should automatically be removed.

It may signal a deficiency in your model, that is,

you may have a missing predictor. In the X variable,

you could have leverage points, and extreme X values where X

is a predictor variable or could be a high leverage point. So

the data point xi will be unusually out of range of the other predictor x values.

It may not have a large standardized residual, but

it can affect regression results. So identify leverage points by index plots,

dot plots, box plots or Cook's distance.

And if the leverage plot is flagged via Cook's distance, then it's also

an influential point and this has substantial influence on the fitted model.

So why does this exist? It requires a case-by-case data analysis.

And often it's best to analyze the leverage point by creating model with and

without the data point to see the effect it has on the fitted line.

This method also applies to both Y and X influential points.

So here's an example, an outlier is an influential point

if it causes substantial changes in the fitted model.

If its deletion causes large changes, then the data

point has undue influence. And detection, with several variables,

it's very difficult to detect influential points graphically.

Hence, we use Cook's distance, and we've seen that earlier,

how Cook's distance is shown. And the rule of thumb is to

identify points with Ci>1 as highly influential.

We'll next discuss the multicollinearity between the independent variables.

This creates problems in fitting regression models.

So I'm going to be using an example with two regressions from

a different data set. It's called the auto data set in R.

And one of them will exhibit symptoms associated with multicollinearity.

So I have Reg1 and Reg2. And

Reg1 has miles per gallons regressed on cylinders.

And you get a value for coefficient of cylinders, -3.55. And it's significant.

Reg2 has mpg regressed on cylinders displacement and

weight, and I get a coefficient of cylinders of -0.26.

And in Reg2, you notice cylinders coefficient is no longer statistically

significant and its magnitude is also changed. What could be the reason?

Such a change in a parameter estimate could indicate the presence of

multicollinearity in Reg2. So multicollinearity exists when two or

more of the explanatory variables are more or

less linearly related. So to detect multicollinearity,

one approach is to use what are called VIF, or variance inflation factors.

So we regress the predictor variable Xj against all other predictor variables and

name the resulting R squared as Rj squared. So we define VIFj

as this expression 1/(1-Rj squared) and

do this for all the p predictor variables.

If Xj has a strong linear relationship to the other X variables,

then Rj squared is close to 1 and VIF will be large.

Typically values of VIF> 5, signify presence of multicollinearity.

So we use of VIF function on the predictors of the Reg2 model and

we obtain VIF values of 10.5,

15.8 and 7.8, respectively.

These high VIF values indicate multicollinearity problems.

And we see that in the correlation matrix also.

Consequences of multicollinearity, you've already seen how VIF is calculated.

Consequences are that OLS estimated parameters may have large variances and

covariances, thus making precise estimation difficult.

The confidence intervals of the estimated parameters tends to be bigger,

hence you may not be able to reject H0, or null hypothesis can be rejected.

Regression coefficients have the wrong sign, or

regression coefficients are not significantly different from 0 or

the R squared is high. And

adding an explanatory variable changes other variable's coefficients.

So what's the solution if you have multicollinearity? So

one option is to pick one variable if two variables measure the same thing.

The other option is to use techniques such as principal components analysis or

factor analysis to create more useful variables. So in this module,

we've covered several topics related to linear regression.

We looked at the steps involved. We looked at this real estate data set.

We looked at the notation used in linear regression,

understood the concept of R squared and adjusted R squared.

Saw how to do simple regression on one predictor variable using R.

We then extended that to multiple regression. Then we saw how R squared and

adjusted R squared are computed when we do multiple regression.

Then we looked at common problems and fixes in linear regression. Thank you.

[MUSIC]