Hi, I'm Sridhar Narasimhan professor in the Scheller College of Business at

Georgia Tech.

I'm teaching this module on nonlinear transformation models,

in this course on data analytics in business.

There are five lessons in this module.

The first one is about why we need a nonlinear models.

The second one is on linear-log models.

The third lesson is on log-linear models.

The fourth lesson is on log-log models, and

the fifth one is on polynomial models.

Here's an example of a nonlinear relationship.

In this case, between the population of your cities and their rank,

the rank is on the x axis, population on the y axis.

As you can see, the top ranked cities have very big populations, but

then the population plateaus off.

We see similar trends in book sales and rank.

Here's another example from the housing data set in the Ecdat Package in R.

This graph seems to indicate Heteroscedasticity.

That is the residuals have non constant variance,

the residuals increase has predicted price increases.

This simple regression model of price regressed on

lot size uses the housing data set.

So here's the model, price is the Y variable, lotsize is the X variable.

As lotsize increases by one unit,

price changes by b1 which is 6.599 units,

holding all other factors constant.

This is the scatter plot for the example in the previous slide and

the regression line obtained by fitting Model A is also shown.

We have previously seen how to do these full

diagnostic plots in R using the plot model function.

So we have this model of price against lotsize.

We need to check if the assumptions for a linear regression model hold.

And the residuals versus fitted values plot indicates that

Model A has Heteroscedasticity, that is non-constant variance.

The QQ-plot also suggests that there is non-linearity.

Hence, we need to start exploring non-linear models.

In this module, we focus on natural log transformations

of the variables because they are easier to interpret.

Note that in R, log computes the natural

log There are several non-linear models.

We are going to discuss three of them.

We start with the most basic, or level level model,

where neither the y nor any of the x variables are log transformed.

This is a model that we have studied previously and know how to integrate

The next model is the Linear-Log Model or Model B.

Here we transformed the x variable by taking its log.

Views log of x in the regression model, y variable is not log transformed.

The third model is the Log-linear model,

we transformed the Y variable by taking its log and

we use log Y as the dependent variable.

In the regression model, the X variable is not transformed in this model.

The fourth model is the Log-log model.

We transform both the X and Y variables by taking their logs.

They use log Y and log X in the regression model.

Note, the log function in R computes the natural logarithm.

And if a variable x has some value equal to zero, or

for some observations its value is equal to zero,

then we use log of x plus one in the transformation.

[MUSIC]

In this lesson, we discuss the linear-log regression model and

how to interpret its output.

So this is the model,

price as a function of log(lotsize).

So we create a new variable in the regression,

which is the natural log of lotsize.

And we run Model B using the housing data set.

So how would you interpret the coefficient of Ln\_lot size?

So, what does the coefficient b₁ = 37660 imply?

Is it A?

Is it B?

Is it C?

Or is it D?

So what's the correct answer?

So the correct answer is D, a one percent increase in

the independent variable increases, or

could decrease the dependent variable by coefficient/100 units.

So, interpreting a linear log model needs to be done very carefully.

It does not make much practical sense to increase

the log of the X variable by one unit.

But increasing X by 1% is almost equivalent to

increasing natural log(X) by 0.01 units.

Hence, a 1% change in X increases natural log(X) by .01 and,

therefore, changes the Y variable by 0.1\*b₁.

Even with this transformation, as you can see in the scatter plot,

there seem to be some higher residuals for higher values of log size.

So we can do the diagnostic plots in our and the QQ plot is problematic

as well as the residuals versus fitted values plot, and the scale location plot.

[MUSIC]

In this lesson, we discuss the log-linear model and interpret its output.

And here's the model, it's the log-linear model,

where the dependent variable is transformed.

So log off price is regressed on lotsize.

And we create a new variable which is a log of price which is

natural log of price.

So how would you interpret the coefficient of lotsize

from running model C on the housing data set?

So what is the coefficient of b1 implied in a log linear model?

Is it A?

Is it B?

Is it C?

Or is it D?

So what's the correct answer?

The correct answer is B, the dependent variable

changes by 100 times coefficient percent for

a one unit increase in the independent variable,

while all other variables in the model are held constant.

So increasing x by one unit will increase the natural log of y by b1 units.

So with x equals lotsize, in this particular case,

this particular model is the same as y eqals e to the b0 + b1x,

Hence dy/dx is given by the b1 times y or dy over y = b1 times dx.

You multiply both sides by 100,

we get 100 times dy over y = 100 times b1 times dx.

Note that 100 times dy over y is the percentage change in y.

If dx = 1, then this one unit change in x leads to

100 times b1 percentage change in y.

Note, this interpretation works when b0 + b1x is very small.

The exact percentage change in y is given by the expression e

to the power of b1- 1 times 100 for a one unit change in x.

We can look at the scatter plot of log of price versus

lotsize and all the fit show the regression line.

And then we can look at the diagnostic plots, and we still have some concerns.

For example, the residuals versus fitted value shows heteros capacity.

[MUSIC]

In this lesson we'll discuss the Log-Log model and interpret its output.

We see the results of fitting the Log-Log model, both the x and

y variables are log transformed as you can see in model D.

Both the left hand side and the right hand

side variables are log transformed.

And the question is how do you interpret the coefficient of large size?

So this is a model, so the value of b1 equals 0.54,

what does it imply, does it imply A, B, C, or D?

So what's the correct answer?

So the correct answer is A, the dependent variable changes by b1% for

a 1% increase in the independent variable while all

other variables in the model are held constant.

So interpreting the Log-Log model.

So increasing the log of x by 0.01 leads to

increasing natural log of y by b1 times 0.01.

Sorry, increasing natural log of x by 0.01 is

almost equal to increasing x by 1% which implies changing y by b 1%.

In a regression setting, we'd interpret elasticity as

a percentage in y, the dependent variable when x,

the independent variable increases by 1%.

Hence, b1 captures elasticity in the Log-Log model.

So let's look at the scatter plot, the scatter plot and

the fitted regression seem to suggest that this particular model is most appropriate.

And the diagnostic plots are very encouraging from

the perspective of not violating the regression assumptions.

So when we compare the models, we see the four models here,

the Level-Level, Linear-Log, Log-Linear, and Log-Log.

And you can see the R-Squared and Adjusted R-Squared values.

So, which model would you choose?

So there are several reasons for doing log transformations.

One is to achieve a more linear relationship.

To make a distribution more normal.

To make the variance more constant.

To get a better fit in the model, that is increasing the R squared.

And here's a cheat sheet to understand log transformations.

So we have x and log x and y and log y.

And I've listed the four models,

A, B, C, and D and how to interpret

the coefficient of b1 for these four models.

[MUSIC]

In this final lesson, we'll look at polynomial models.

So in this example, we're going to introduce a quadratic model.

So price is regressed on lotsize and the square of lotsize.

So I've created a new variable called lots\_square,

which is the square of lotsize.

And then I fit the model, price is a function of lotsize and lotsize squared.

And we get the result shown in this table.

So how would you interpret the coefficient b1?

So b1 is 14.81, what does it imply?

It is A, B, C, or D?

What's the correct answer?

The correct answer is D, none of the above.

Interpreting these models, polynomial models, are complicated.

The coefficients b1 and b2 cannot be interpreted individually,

because when lotsize is increased by 1 unit, it is not possible or

meaningful to hold lotsize squared constant.

So a quadratic model does not allow for

an isolated interpretation of coefficients, since d(price) or

d(lotsize) is given by the expression there, and that expression has lotsize.

This means that the slope or

an impact of 1 unit increase in x is not a constant.

It changes at every point of the quadratic curve if you plot y versus x.

So in this module, we've had several lessons on

different nonlinear models that are used in regression.

Thank you.

[SOUND]