Hi, I'm Professor Sridhar Narasimhan of the Scheller College of Business

at Georgia Tech.

In this module, we are going to understand how indicator or

dummy variables are used in regression analysis.

Note, that I'll be using the terms indicator variables and

dummy variables, interchangeably, both mean the same thing.

This module is part of this course on data analytics and business.

There are five lessons in this module.

In the first lesson,

we will introduce a Customer Analytics Dataset.

In the second lesson, we'll see how indicator or

dummy variables are created and used in a regression model.

In the third lesson, we will interpret the coefficients of indicator variables.

In the fourth lesson, we'll see how interaction terms are used and

how do we interpret the coefficient of interaction terms.

And the final lesson is another example to using indicator variables in regression.

So we're going to use a data set that is from a direct marketing company.

So a direct marketing firm has a data set containing information on past

customer behavior actually the amount spent on buying products.

And this is a simulated data set which mimics data from a direct

marketing company.

We are interested in knowing which customer characteristics can predict

AmountSpent, that is the amount our customer spends on buying products.

To answer questions like this, we introduce indicator variables and

interaction terms and try to understand their interpretation in regression models.

So the direct marketing data set has these variables

starting with age, gender and going on to does the customer own a home or not?

Is she or he married, location, salary, number of children,

the type of customer, the number of catalogs sent to this customer and

then AmountSpent, the amount of purchases made by this customer.

So the first few records look like shown here,

and I show this in a table in the next slide.

So this is the structure of the data set, and

the first ten rows of this dataframe look as follows.

So let's explore this dataframe a bit further.

We'd like to understand why some individuals spend more than others.

In particular, we'd like to investigate whether salary

has an influence on AmountSpent.

So how do we get started?

So to keep matters clear, we start with AmountSpent and Salary and

just pick those two variables for the initial analysis.

So we'll do a scatterplot, plot Salary on the x axis and

AmountSpent on the y axis, and we get this scatterplot.

And then we do a simple regression model of AmountSpent

against Salary, and this is the values we get.

And this is scatterplot with the regression line.

So we've introduced custom analytics data set in this lesson to illustrate

indicator variables.

[SOUND]

In this lesson we show how we can create and

use indicator or dummy variables in regression model,

so what should we do about the categorical variable age?

So is this categorical variable important?

Does being middle aged or

old potentially have an effect on amount spent compared to being young?

How can we include the age variable in a regression

model that requires numeric values?

So these are some values of the age variable, so consider that variable,

we want to investigate the effect of age on amount spent,

note that age is a categorical variable.

It has three possible values, young,

middle or old, we need to quantify this variable.

So what we need to do is to create indicator or dummy variables,

since we have three possible values for age,

we need to create two indicator or dummy variables.

The base or reference case, with both dummy variable set to 0, is age = young.

This is the reference group to compare for the other values of the dummy variable.

It's up to the modeler to determine which value of

the categorical variable is used as the base case.

The two dummy variables that we have created are AgeMid and

AgeOld, and AgeMid is set to 1 if the age of a customer is,

if he or she is in the is middled aged.

Zero otherwise, and AgeOld is set to 1 if a customer

is an older person and set to 0 otherwise.

So assigning values to the new indicated variables, so

here is how we're gonna go do it, if you look at the first record,

it is for a customer who's old, so for that customer the AgeMid

callable a value of 0 and age old will have value 1.

The second record is for customer who's middle age for him or

her the AgeMid column A lot of value 1 and the AgeOld is 0.

The third record is of a base case, this is a customer who is young and

for him or her the AgeMid value 0, and the AgeOld value is also set to 0.

So both indicate a variable to set to 0 for someone who is young.

And then we can run the regression model amount spent

equals b0 plus b1 times AgeMid plus b2 times AgeOld.

We'll see that shortly.

Before we go to the next lesson, have a go at this quiz.

With this indicator variables coding scheme, AgeMid and

AgeOld, defined as earlier, can a record in this dataframe

have this value (AgeMid = 0, AgeOld equals = 0)?

The answer is yes, because this record is for

someone who's young, that is the base case.

Can a record in the direct marketing dataframe have

this value, AgeMid = 1, and AgeOld = 1?

The answer is no, every individual has to be inexactly one age category.

So, we've seen how to create indicator dummy variables in this lesson.

[SOUND]

We know how to create indicated variables.

Now, in this lesson, we'll see how we can use them in a linear regression model and

learn how to correctly interpret their coefficients.

So, we've seen this coding scheme, each made an AgeOld.

We run this is regression model, AmountSpent, regressed on AgeMid and

AgeOld.

And then we fit it using the data in direct marketing.

We interpret the regression output for

those records in the direct marketing database.

So we have our own spent regressed on AgeMid and AgeOld.

So which age groups average amount spent is 55.862, is it young, middle or old?

The correct answer with AgeMid = 0 and AgeOld = 0,

b0 captures the average amount spent of customers who are young.

That is for customers in the base case.

So amount spent is b0 plus b1 AgeMid + b2 AgeOld.

So what's the average amount spent for someone who is middle aged?

So this individual has AgeMid = 1, AgeOld = 0, so

b0 + b1 captures the average amount spent for folks who are middle aged.

So, that's gives you b0 + b1,

$55.86 + $94.3, $150.169.

So, 94.307 is the increase in amount spent on average for

middle aged customers compared to someone who is young.

Same regression.

What's the average amount spent for someone who is old?

So this individual has AgeMid = 0, AgeOld = 1, so

b0 + b2 captures the average amount spent for customers who are old.

So that's b0 + b2 = 55.862 + $87.35,

you get a total of $143.212.

So, $87.35 is the increase in amount spent

on average for old customers compared to someone who is young.

So we can look at this graphically.

So we have age and amount spent.

So folks who are young, on average,

will spend 55.862.

Folks who are middle aged, they'll spend in additional $94.3.

And then folks who are old compared to the young

will spend an additional $87.35.

Note that you don't explicitly have to create and use dummy variables in R,

in order to use a factor or categorical variable in a linear regression model.

You can use a factor variable directly in R.

And to do that, we just do it as this example shows,

regress amount spent on the categorical variable age.

And we get an output.

But then you'll have to know what is the base case.

So in this case, the R creates new variables AgeMiddle and AgeYoung.

So that base case is old, written amount spent, $143.213.

So what's the average amount spent for folks who are young?

It's 143.2- 87.35 and you get the value of that.

What's the average amount spent of folks who are middle age?

143.2 + 6.95, you get $150.00.

All three groups are the same answers as our coding scheme where

young was the base case.

So R's indicator variable coding can be found by using this command contrasts.

And I will tell you what the coding scheme is.

In this case, R is using a different coding scheme for dummy variables.

If you are new to dummy variables, I recommend defining and

using them in your regression model.

I personally find it more useful to use my own coding scheme.

In the next regression, we're gonna use salary and the dummy variables.

To use salary and the dummy variables.

So amount spent is b0 + b1 \* Salary + b2 \* AgeMid + b3 \* AgeOld.

So what's the average increase in amount spent for a one unit increase in salary?

The answer is $.002.

Graphically, this particular regression model, the output would look like this.

Salary on the X axis, amount spent on the Y axis.

So the folks who are young have this line,

the folks who are middle age have this line, and

the folks who are old have that particular line.

The slope of all the 3 lines is .002.

So here's a quiz.

So, we've used dummy variables and we got this output for this particular model.

So what does this result mean?

Is it A, B, C, or D?

What's the correct answer?

The correct answer is D.

At the same salary level, old customers spend more than young customers.

In this lesson, we have used dummy variables in a regression model and

learned how to interpret the coefficients.

[MUSIC]

In this lesson, we are going to introduce the interaction term and

interpret its coefficient.

So in the data set, location is a categorical variable with a value

equal to Close if the customer lives close to a store that sells similar

merchandise as a direct marketer, and is equal to Far otherwise.

So Far is equal to 1 if the location is far,

if the customer's location is far from a store, it's 0 otherwise.

We want to study the impact of location on the amount spent.

So we say amount spent = b0 + b1 times Salary + b2 times Far,

and we get this table of data aggression output.

We can look at it graphically.

And this is the output that we saw earlier.

So what's the estimated amount spent for a customer who lives far?

So for this customer, is it which of these four values?

So for this customer, since Far is equal to 1, the correct answer is D.

Because amount spent is equal to -20.48

+ 0.002\* Salary + 59.06 \* 1,

which is 38.58 + 0.002\*Salary.

But, AmountSpent is this expression in our regression.

We're assuming customers who live far away from a store that sells

similar products will spend at our direct market firm at the same

rate as customers who live close to a store.

Is this assumption realistic?

So to answer this question,

can we investigate another scenario that the spending rate may be different for

customers who live close and for customers who live far?

So how should we change the regression model?

To do that, we introduce interaction term.

Is spending rate higher for customers who little far away?

So we construct a new variable, call it SalaryFar.

And that's a product of Salary \* the dummy variable Far.

This is called an interaction term.

So now the new regression model is,

AmountSpent is a function of Salary, Far and SalaryFar.

And then when you run this regression, we get amount

spent as b0 + b1Salary + b2Far times b3SalaryFar.

And you get the table with the parameter estimates.

What does this result mean?

So this is the model, this is the output, how would you interpret b3,

the coefficient of SalaryFar?

So b3 is the amount to add to b1 to get the slope for

individuals who live far away.

So in this graph, the slope for

Far is different from the slope for folks who live close.

To test your understanding,

this is the regression model with the interaction term.

If the salary of a customer who lives close increases by $10,000,

what's the predicted increase in AmountSpent for that customer?

So for this customer, that is the baseline case,

Far = 0, thus the relevant slope is b1 = 0.002.

Hence, the increase in AmountSpent on average for a customer who lives close,

increases by $0.02\*10000, or $20.

Same question but for a customer who lives far away,

what would happen if his salary increases by $10,000?

What's the predicted increase in amount spent for that customer?

So for this customer, Far = 1, hence the relevant slope is b1 + b3, which is 0.003.

Hence the increase in AmountSpent on average for

an individual who lives far away is $30.

If you had a categorical variable with M Values,

then you will need to construct and use M-1 indicator variables.

Be careful when using and interpreting the value of the coefficients

of the dummy variables and for any interaction terms.

Remember the base case applies to the group where all

indicator variables are set to 0.

And all other cases have to be interpreted with reference to the base case.

So in this lesson, we saw the concept of interaction terms and

learned how to interpret their coefficients.

[SOUND]

In this lesson, we'll look at another example that uses indicator variables.

So this is a publicly available data set, and

we're focusing on the LA market for this particular lesson.

And the link is provided, we're looking at rentals in the Los Angeles area.

The data was collected by us on May 2nd, 2017.

We discarded listings with price greater than $1,000 and

also listings that had missing values for bed, baths, ratings.

So there are several variables, there's price, reviews, beds, baths.

Capacity is how many folks can be accommodated in a property, and so on.

And these other variables that you see here.

So the questions that we have is to figure out if

a property owner lists his property in the LA market and

wants to get a fair price for his property or her property.

Then it is essential to understand the key factors that influence the price for

his or her property.

So he or she may want to understand what variables influence listing price.

Is there a relation relationship between capacity and price?

Just the type of rental, is it shared or private?

Is the entire house change this relationship?

So we do some data wrangling with the data.

We convert, do all these things that I've shown out here so

that we can use the data in our regression model.

See now, first, simple regression,

we're trying to figure out does price vary by room capacity?

And we get this particular output and

now your model would suggest that an individual added

in our rental adds $38 to the price in the data set.

So if your property can accommodate one more individual,

you can charge 38 more dollars.

So this is what the scatter plot with a regression line would look like.

So we're going to create some dummy variables for,

The type of property, whether it is private or entire home.

If it's a private room, or an entire home, or

apartment which is listed in the rental.

There are three types of rooms, shared, private, and entire home.

So shared is the base case, and then we define two dummy variables,

House and Private.

So, again, as I mentioned, the base, or reference,

case with both dummy variables set to 0 is Room type = Shared.

We'll use this as the reference group to compare for

the other values for the dummy variable.

So how does price vary by room type?

So that's a regression model, price = b0 + b1 × Private + b2 x House,

this only has dummy variables and we get these values.

So the question, which room's types average price is 37?

Answer is shared room, what's the average price of a private room?

So add $35 to 37, what's the average price of an entire house?

Add $133 to 37, so let's take a look at this graphically.

So a shared room on average you get a price of $ 37 a night.

A private room, you add $35 to the shared room price.

And for a house, you add $133 to the shared room's price.

So do a second regression, now introducing

capacity and keeping the dummy variables.

So if a property can accommodate more people, what can you get?

And this is a regression, price is a function of capacity and

the two dummy variables.

So what's the average increase in price for

each additional individual a property can accommodate?

So the answer here is $29.29.

Graphically, let's look at the output of this regression model,

and this is what it looks like.

The shared room is bottom-most line,

then you have private room, and then you have the house.

And you see the offsets for the private room is $30,

and house, you get an additional $75.

Slope of all three lines is 29.292.

We then, in our final model,

add an interaction term, which is P\_Cap and H\_Cap.

So private dummy multiplied by capacity, and

H\_Cap is housing dummy multiplied by capacity.

And these are the interaction terms, P\_Cap and H\_Cap, and

then we run the new regression.

So price regression capacity, private individual,

room, house, P\_Cap, and H\_Cap.

So run the regression and we get this table, how would you interpret b4 and b5?

The coefficients of P\_Cap and H\_Cap respectively.

So b4 is the amount to add to b1 to get the slope for a private room.

b5 is the amount to add to b1 to get the slope for a house.

Statistically, capacity and house are not very different from 0,

so let's take a look at this graphically.

So shared room is shown in black,

a private room is shown in red, and a house, shown in green.

All three have different intercepts and different slopes.

So in this module, we had several lessons,

we saw Customer Analytics Dataset, which we used.

And then in the second lesson,

we used a set of indicator variables.

And then we saw how these variables could be interpreted,

used in a regression model.

And then in the fourth lesson, we saw interaction terms and

how to interpret their coefficients.

And then in the last lesson, we saw another example

of using indicator variables, thank you.

[SOUND]