Week 4

Linear Regression

Video 1: Linear Regression and Hypothesis Testing

Hi. Welcome to the module on linear regression. This is also known as linear module. And the topics we're going to cover during this module include One-and-two-sample t-tests, regression on a single variable from which we'll move to what's called multivariable regression, which means regressing on more than one variable. And we'll talk about how we actually come up with numbers of results when we do these things using the approach of least squares. And we'll also talk about the assumptions we make when we use these approaches, and the properties of the output we get. We'll talk about the idea of confounding and the related idea of collinearity, and we'll wrap it up by talking about non linearity in linear models, as well as what's called interactions. So, the first thing we're going to discuss, as I said was one-and-two-sample t-test. That's going to be the first section of this module. So, we're going to start off by discussing, single distributions. So, statistics is the study of what we call random variables, which just means variables that randomly vary. For instance, looking at the weight of individuals, or their BMI, or their eyesight, all these things are, they vary in the population. So, we like to visualize these things using plots such as the ones you can see right now. So, these are three different ways of showing the same single distribution. The one on the lower left is a histogram, and from that you can see that the highest bar corresponds to values in between -1 and 0. And you can also see the values between 0 and 1 is the next highest bar. So, that suggests that the central tendency of that distribution is around 0. The plot on the upper right shows what some people call Violin plots; they have different names. But it's just the same idea of a histogram, but allowing the spread to go in two different directions. And finally, on the bottom right is a representation of the same distribution using what's called a Boxplot. And so what you see there is a rectangle, and the middle bar going through the rectangle represents what we call the Median, and the upper and lower parts of that rectangle correspond to what we call the 25th and 75th percentile.

So, statistics is very much about modeling. And so modeling means trying to explain something in a simple, easy to understand format. So, the most simple model we can think of is the following. We could write it like you could see in this formula here, where we write, we take a variable Y and we express it as something, we'll call it mu, which represents an average or central tendency, and the error, the distribution around that central tendency. So, this is a the simplest model we could write in statistics, and they'll get more complicated when we get on to writing a linear regression formula and a multivariable regression formula. So, you might ask, what can you do with such a simple model? Well, one thing you can do is ask, what about that parameter mu? Is it actually equal to a particular value? And sometimes we're interested in knowing if the center of a distribution is 0. And the times when that would happen would be, for instance, when you're looking at change, the change in something that's happened. So, if you're running a study in which you give people a new diet, and you want to know if there's a weight loss. So, you want to know if the average change in weight is equal to 0, or is it actually better than 0? They've actually lost weight. So, this is very common in

how science is to conduct what we call paired studies where you have paired measurements like a before and after, a pre and a post.

Or it could be, another type of this variation is called a crossover design, which has some advantages over a pre-post study in which you kind of randomizes when actually people get the intervention, whether it's before the first measurement or before the second measurement. I won't get into the details of why that's better, but sometimes there's benefits to it. But in any kind of these, any of these pre-post or crossover study, we're always interested in knowing, is the average change equal to 0? And for that, we have what's called a paired, or one sample Z or T test. So, there's different types of t-test, one and two samples, and a paired or one- sample is often used. So, the assumptions we make when we conduct such a study, or use such a statistical tool, or that the observations are independent. So, for instance, if you're looking at before and after weights in a bunch of people, being independent probably means, for instance, you're assuming that they're not the same family members. Like, it's not children, like three children from the same family in your dataset, because they might be dependent measures, we said. But to actually implement a paired or one-sample T test, we're going to do this in R in a minute. The actual formula is like this. You don't have to memorize this formula. But the idea is that if you take the average of a bunch of pre post measurements and your hypothesis, or what we call the null hypothesis, is that there is no change at all, then the average should be close to 0. And when it's not close to 0, that's when we reject our null hypothesis. And so this a little more math for the paired or one-sample Z or t-test. Again, you don't have to understand these formulas, but know that when you are running these methods in R, this is what it's doing under the hood. It's just doing some calculations such as this, adding up all the values and dividing essentially by the number of observations, and wondering then if that value is close to 0.

That's what a one-sample t-test does. And so, we sometimes you hear the difference between a z-test and a t-test; in large samples, a Z and t-tests are the same thing. In small samples, the t-test is slightly preferred, although you do end up depending on the assumption of what we call normality. And the assumption of normality means they're the bell shaped distribution. So, the next thing we might do is actually not look at a single distribution, but look at two distributions. So, here you see three visualizations of the same data. You have the red data and the green data. So, you could think of these as, perhaps, the values. The glucose levels in individuals, for instance, people with diabetes, we monitor glucose. So, perhaps, let's say the red values are the glucose values in people that have been on a particular diet. And the green values are people that are on another diet. And so, you're interested in comparing the average glucose values between these two populations. Or it could be anything, these two distributions. It could be weights of these type of animals that got a certain diet, and weights of another type of animal that got another diet. So, we have two distributions here, and what we're interested in doing is comparing the distributions. Usually asking questions such as, is the average between the two distributions equal? So, on the bottom, you have histograms superimposed on each other. On the upper right, you have what we call these violin plots, side by side. And then, the boxplots, this representing the same data but using boxplots, where again, in a boxplot, the middle bar is the median, and the lower and upper parts of the rectangle are the 25th and 75th percentile. The other things that go beyond these whiskers that you see above and beyond, I won't go into their definition. Sometimes I even take those out. But in R by default, you're going to see those.

So, we have two distributions, perhaps, like I was saying, the glucose values and two groups of subjects that got a different diet, or two types of people with hypertension.

And these are the values when some of them got a particular treatment and when some of them didn't get that same treatment. So, you want to know, what is the average? How does it differ between the two groups? And for that, what we do is we end up comparing the means, simply by taking a difference. And like you see on this page here, so you calculate the average in each group and you take the difference. And we do a little more with it, we divide by a certain quantity, and we create what's called a Two-sample t-test. What you see on the top formula there, you see a difference in means, in the numerator. And below it, you see inside the square root sign, something that gets at the variation. So, with two-sample t-test is going to be like a one-sample t-tests, and what you're getting a quantity which, if the null hypothesis is true, you're going to have a value that's close to 0. If the value is far away from 0, it suggests that your null hypothesis is, should be rejected, and that there is a difference in the means between the two populations. So, what we're going to... I'll finally mention here, so for the two-sample t-test, there's two different flavors. One in which assumes equal variances and one which doesn't. If you're interested, I encourage you to look into that by default. The one that's used in R is the safer one to use which is called Welch's, which does not assume anything. Well, it doesn't assume that the variances are equal. So now, having learned about one and two-sample t test, what we're next going to do is going to R and implement those, using the tools in R.

Demo Video 1: Reading Dataset into RStudio for linear Regression

So, what we're going to do to start is reading a data set. So, how I'm doing this is using R, and in fact, I'm using a version of R called RStudio. And furthermore, it's a file that's in a R markdown format. So, the way I've coded it is, has with our markdown, is that, you code in what's called chunks. So, I read... run this initial chunk right here, which is system admin stuff, and then I'm going to go down here, and this is the first chunk of real relevance because we're going to read in the data set. So, this chunk is going to read in the data set, and then it's going to give us some information, such as the dimensions of the data set, number of rows, which is the end sample size, and number of columns, as well as the names of those columns, otherwise known as variables. So, let's run this chunk, and here's some of that output. What's going to be of interest to us in the analysis of running is something called the glucose treatment variable. And so, you can see a breakdown there in how many cases there was glucose treatment. In this case, it's coded as true, 278 times. And how many times it was coded as false, 242. So, overall, this is a data set of 520 subjects, 278 of which received the glucose treatment, and 242 who did not. The next chunk we're going to illustrate the one-sample t-test and central tendency. So, as a reminder, the one-sample t-test is a test that the central tendency, in this case, the mean is equal to zero.

So, in this chunk, the first thing I did was create a new variable, which is simply the difference over follow up of the variable referred to as Hemogblobin A1C. Hemoglobin A1C is a very important variable in diabetes. It's a measure of glucose control, and, so patients often have it measured to get an idea of just how well their glucose is being controlled. High values are bad, the low values are preferred. So, we've created a new variable, which is the difference over a certain time frame, baseline to sometime later in a study. So, this is a study in which patients were measured initially, baseline.

And then, say, six months later, they had their glucose control, Hemogblobin A1C, measured again. So, we're calculating the change during the study. So, in this chunk, what we're doing is we're calculating some summary statistics for that change in Hemogblobin A1C amongst people that actually have the glucose treatment. So, keep that in mind. So, amongst people that got this glucose treatment or glucose lowering treatment, what was the change in Hemogblobin A1C. So, this summary statement's going to do that, and we're going to generate a histogram. And then we're going to test the no hypothesis, that the mean is equal to zero. So, as you can see down here, if we look here in the console window, we get some summary statistics. So, we can say that the Hemogblobin A1C during the course of the study among people who received the glucose lowering treatment, was lowered an average of 0.896. So, I say lowered an average because that's a negative valued mean. And the change in Hemogblobin A1C

range from -5.1, that's a huge drop, up to an increase of 4.0. And the median was -0.9. In other words, the median reduction in Hemogblobin A1C amongst these individuals who received glucose treatment was 0.9. And there you can see a histogram of Hemoglobin A1C. So, each bar reflects just how many people are not that particular lower range, had that change. So, for instance, if you look at this bar, which goes from -2 to -1, so people who dropped anywhere from 1 to 2 units of Hemoglobin A1C, it suggests that that was the most represented value reduction. About 70 people had such a reduction. And now let's go to the one-sample t-test output back to the R console window. And what's it telling us? Well, probably our attention immediately goes to the P value. So, P value is less than 0.0 5 are considered statistically significant. So, this one is actually very significant. It's using scientific notation to represent the P value. And this suggests, for instance, that there's 15 zeros in the P value before you get the first non zero digit, so very significant. So, it's giving us the same mean we saw above -0.896. So, we saw that in the summary statistics. And it's simply asking us, is that -0.89 that average reduction Is that statistically significant from zero? Or is it within random variation? And this is saying very strongly, no, that's not within statistic random variation. This suggests that people really are experiencing reduction in Hemoglobin A1C.

The next chunk we're going to do is now speak about a two-sample t- test for comparing two distributions. So, in this next chunk, we're going to look at the change in Hemoglobin A1C from baseline, when people entered the study, until the end of follow up, when the study ended. And we're going to relate that to whether they got the glucose lowering treatment or not. So, that's the first part here. And so you can see here that amongst people that got their glucose lowering treatment, so it was true they got it, the average reduction was 0.896. We already knew that from the previous chunk. Amongst people that did not get the glucose lowering treatment, the average reduction was almost the same. There was 0.8748. So, not necessarily looking like a big difference in the average reduction in Hemoglobin A1C during the course of the study between people who were receiving the glucose lowering treatment and people who were not. There's a little more separation between the medians, but overall, it doesn't look like necessarily it's a big difference. To get a visual, we can look at the boxplot. So, this boxplot compares two distributions. It compares people, again, who received the glucose lowering treatment, that is on the right here, and people who did not get the glucose lowering treatment. And the boxplot is a visual representing the distribution of those changes. The black line in the middle of the box represents the median, and the upper and lower ends of each box corresponds to the 25th and 75th percentiles, sometimes called the 1st and 3rd core tiles. And from eyeballing these two box plots, these two distributions,

one can say that, yeah, there's really not a big difference between those two distributions. Those box plots overlap a lot. They overlap a lot, suggesting that the two distributions aren't that different. But let's actually go and actually do a two-sample t-test to test that hypothesis. And so this is what this output is, it's a two-sample t-test. And like most times when we're getting statistical output, often our first interest is in looking at the P value. And here we got a P value of 0.8847. So, the important thing here is that the P value is way above 0.05, in other words, is not going below that threshold for statistical significance. So, we would conclude that the change in Hemoglobin A1C during the course of the study, did not differ between individuals who receive treatment and individuals who did not receive that glucose lowering treatment. That's probably end this particular session, and subsequent videos will be speaking about regression for the first time.

Video 2: Linear Regression Models

In this section, we're going to discuss linear regression for the first time. So, the key point with linear regression is you have a dependent variable, which is continuous, and you're seeing how it changes with respect to another variable. Think of it as continuous as well. So, when I say continuous distribution, I mean something that can take any value as opposed to categorical value such as race, which have distinct categories, or gender. So, what you see here is a scatter plot. We call these scatter plots, and it's a way of showing how one variable may or may not vary with respect to another. In this case, I've labeled them as Y and X, but you can think of Y as for instance, glucose levels or systolic blood pressure and X as say weight or waist circumference or body mass index. So, we're looking at the association of two variables, sometimes we say the correlation of two variables, and we'll talk a little more about, what's called a Pearson Correlation as we go along. So, the most simple model to relate how two variables might rely another is actually a linear model. That might sound kind of technical, but a linear model is actually a very simple of way of relating two variables. What we see here is actually two different ways of writing the same model. So, the first way I write it that capital E, think of that as statisticians say, expectation but you can think of it as the average. So, how does the average of Y depend on X? Well, it depends on it based on this linear formula. That's what the top equation is saying. The bottom equation is just another way of saying the same thing that Y, the dependent variable, is equal to this linear combination of X, so it's a plus b times X. So, you can remember high school math, like for instance, grade nine or grade ten where your graphing a Y versus an X, and you do things like calculate slopes. So, here we have the parameter b here. You can think of it as the slope. We call it the coefficient, but it relates how Y changes in relation to X. So, if you change X by a certain amount, how much slope do you have making a change in Y? So, that is the linear model. So, this is something an equation you should remember for linear regression or linear modeling.

This is very important to understand this simple linear model. The next question is, how do we go about estimating that slope, that quantity b? Well, the approach we use is something called least squares but the basic idea is that in order to come up for a value for the slope, that coefficient b... This is the variable, the parameter of most interest or coefficient, so statisticians say parameter coefficient, to get at the quantities that were really interested in. And so this parameter b or coefficient b, we're going to come up with a value for it by minimizing this sum of squares, you see. So, we're going

to find the values of b and a that minimize the distance between y, the dependent variable and this linear combination of x, the independent variable. We're going to minimize this distance between the observed values, y and what we think they should be depending based on what we know about the independent variable x. So, we have this sum of squares and we're going to minimize it using, well, getting the least value; that's why we call it least squares. So, you might ask, "Why do we square it?" Well, the reason we square it is because it makes the Math a lot easier. The alternative would be to not square and use just the absolute value.

But that approach ends up being a lot more complicated. It might seem confusing, but the Math gets a lot more complicated if you don't just square it. And we rarely used the approach based on minimizing what's called the least absolute deviations. We almost always used the approach called minimizing the sum of squares or the doing least squares. The approach has been around for 200 years, but it was really not first applied to real actual data and actually biomedical data until the mid 1900s by someone called Francis Galton, who was considered to be one of the first population Geneticists, first biostatisticians in a sense and also happened to be a cousin of Charles Darwin. So, he was actually used the data, Francis Galton, his example the heights of sons and looking to see how it related to the height of their fathers. You want to know, if a father is five inches taller than another father, does that mean the son of the taller father will be five inches taller? And what he noticed when he applied least squares for the first time was that it wasn't quite such a strong relationship. So, if my father is three inches shorter than your father, I'm not necessarily going to be three inches shorter than you on average, I might be an inch or two shorter than you. So, at the same time that he was using this new approach of least squares, he also came across this biostatistics epidemiological concept called regression to the mean, which is a very important concept in epidemiology and biostatistics. And it's basically the idea that things kind of come back to average. On average, so for instance, if I'm feeling sick and I go to the doctor, I have a high temperature, chances are, next time I see the doctor, my chances my temperature, will be down only because, the only reason I saw the doctor in the first place is because I was feeling ill, my temperature was high. Chances are when I see him again, I will be back to normal. That's kind of the regression to mean concept. So, back to least squares and how we go about getting those quantities. So, I was saying how we're going to minimize the sum of squares and it turns out if you do a little calculus, you don't have to.

But for those of you are interested, if you take a derivative of this expression with respect to b and you do some math, you find out that you come up for an answer for this slope, this coefficient of this particular formula. And when we use R to do our least squares, linear regression in the next video, you'll see an example of that. You won't actually see the calculation taking place. It's all done behind the scenes. You just see the final answer. But this is the formula it is using when R runs that linear regression. So, we're going to estimate the parameter b, which is the slope of interest. And by the way, the other parameter a, we need it, but it's usually not of interest to us. We usually don't care about it. It's more like a nuisance quantity that we need to do, but we do not interested so much in its value. So, in my next video, we're going to talk about how we come up with estimates for the precision of estimators.

Video 3: Linear Regression and Precision

So, in this video, we're going to talk about how precise are the numbers, the estimators we call them or estimates that we come up with when we do linear regression. And this is a key point in statistics is that we like not just to give a number, but to give a measurement, a quantization of just how precise we believe that number is. For instance, you get an idea that during any election season, and you hear the results of polls. So, you hear the results of the poll, such and such is leading in the polls, who has a popular rating of like 62% and then they may say something to the effect of, "This poll is accurate within four points 19 times out of 20." You've probably heard that kind of language before and if you haven't actually picked up on it, you might pick up on it next time you hear it. But what they're trying to do is give you a margin of error or standard error. So, statisticians we usually speak of standard error, sometimes in the media, in lay language refer to the margin of error and the margin of error is just actually twice the standard error. So, if the margin of error is plus or minus four points. That means the standard error is plus or minus two points. So, when we run linear regression, we want to know when we get that estimate of the slope. You know, we're going to give you that value of the slope relating the dependent variable to the independent variable. Just how precise is it? And so as part of the R output. We are going to get a standard error and in the way the standard error is calculated, is using a particular formula, which you can see in this slide. You do not have to memorize this, but it's basically saying that the standard error is going to depend on the variation of the dependent variable, but also the variation of the independent variable. So, if there's a lot of variation in the dependent variable, that's numerator, it's going to be big, and it gets reduced to the extent of the X variable ranges a lot, but it also gets reduced to the extent that the sample size and is large.

So, that's one thing in statistics when people say n they usually mean the size of the study when we say the sample size. So, bigger n's means more precise estimates come out of it. That's the key point. And as the sample size increases, there's an inverse relationship to the standard error. It's actually the standard error reduces in inverse proportion to the square root of the sample size. So, besides, from standard errors, another key concept in statistics is confidence intervals. So, and that's just the same ideas that will give you a standard error or margin of error, and then we actually specify the interval around it. So, for instance, if we heard that such and such had an approval rating of 62% and the margin of error was four points, we would say with a confidence interval. And usually almost always when we refer to confidence intervals in statistics, we mean the 95% confidence interval. That's again 95%. It's rare when people speak about a confidence interval without meaning 95% and what that means is that we're 95% confident that the true value is actually in that range. So, in an election season, if you're polling you say 62 plus or minus four points, that means a confidence level goes from 58 to 66. And we're confident of that with 95%. 95% of the time, the actual value, if you were to sample every one of the population, would be in that interval from 58 to 66. So, one thing also to keep in mind is some equivalences. So, I briefly mentioned this idea of A Pearson correlation is a concept in statistics. It's a number that ranges between negative one and positive one, so we can't go lower than negative one or higher than positive one. Negative one means perfect negative correlation. Positive one means perfect positive correlation. So, there's a relationship of regression to Pearson Correlation.

That's just so you know this, that there intimately related Pearson correlations and least squares regression. They are very related concepts. The other equivalence you should know is that, if we were to do a regression where we took a

dependent variable such as, glucose value and we regressed it on a binary variable. So, not a continuous variable, but a binary variable like glucose treatment, yes or no? Well, that's actually equivalent to doing a two sample T test. So, that's some equivalents you should know and going along that, least squares regression is kind of the same thing as other well known statistical approaches it's just a general umbrella for doing these techniques.

Video 4: Multivariable Linear Regression

So, in this video, we're going move on to multivariable linear regression or multivariable linear models. So, what that means is that we're not going to be relating our Dependent variable, called it Y to just one independent variable which we referred to as X before, but to several of those. So, this is the general formula for a multivariable linear regression, is that the average value of this dependent variable Y and for that we substitute whatever example occurs to you at the moment like weights or body mass index or blood pressure or body temperature or any kind of biomarker in a human or for to give you an idea of some data I've been using recently. I've been modeling mortality data in American cities as it depends on government spending each of those cities. So, in those cases I would have a dependent variable, which is the mortality rate in that city. And my independent variables are variables such as, how much the government spends on healthcare? How much it spends on welfare? I also have variables in there for per capita income in that city. Like just how affluent is the city in terms of how much the average income is. So, that's what multivariable models are about. So, we can now start looking at how some variable of interest, the dependent variable depends on not just one, but other variables. And mathematically, this is how we would write it. So, the expected value of Y depends on these other variables, call them X1 through as many as you want by the following linear model. You have the first term, a nuisance term which we call the Intercept. And then, we have these slopes or coefficients, one for each variable. So, this is the linear model, and the reason we like to use a linear model is because it's the simplest way to connect one or more variables to a dependent variable.

A linear expression mathematically, you cannot get any more simple than that. So, you do not have to do this, for any of you who have ever learned any matrix algebra or linear algebra. This might make sense, but if you want to understand the Math, it does start happening at a matrix level. You don't have to know this to use R as we interpret the output. But what actually is going on is a lot of matrix math. And so what's going on is that we create something called a Design Matrix which is basically just all the variables in the dataset lined up side by side. So, you could think of a spreadsheet where the first column might be how much each city spends on healthcare per person in the city. The next column might be how much they spend on welfare per person in the city. The next might be how much they spend on police in the city. The next column might be how much the average income is in the city. So, think of this matrix as a spreadsheet, and this is just the matrix representation of the multivariable linear model. It's a little more abstract than what we saw earlier with the multivariable model, all written out in long style, the matrix style is a more efficient style. But again, it's more abstract and harder to understand. But the reason I put it here is that in order to come up with the numbers, for the estimates for the coefficients, we are using this approach of least squares again. But now it's just the matrix notation. And I don't expect any of your to do the matrix math. If you do, that's great, because you might understand it at a different level. But ultimately what happens is that you come up with a solution which is the following. This is matrix math, but if you ever to say the following to a statistician, you'd probably catch their attention because almost all

statisticians recognize it when you say the following X transpose X inverse X transpose Y. that is the least square solution of a multivariable linear regression in matrix form.

So it's nice, efficient math. You don't have to understand this, but when we run R, this is what it's creating using this formula. So, some other things you have to know when we see the R output, it has to do with the standard errors or the covariances. Just how spread out, or how are precise are the estimators? These are some formulas for how we go about calculating those. If you're interested, please pursue it. But when we start using R the next video, you're not going to have to know these, you can just look at the output and feel comforted that there's a lot of good math going on underneath the hood when you use R. So, the next thing we have to emphasize when we're using linear regression is use of testing. So, testing is a key concept in Statistics. At the beginning of this module, we talked about one and two sample t-test, so when we're doing linear regression, we're also going to be testing or often making use of test. And usually what we refer to these is Wald test. And a Wald test is actually simply calculated. What it does is it takes the estimate, the number from the least squares calculation for the slope, or coefficient, whatever you want to call it. And then it takes that and it simply divides it by its calculation for the standard error. So, if you take the ratio of an estimate to its standard error, you are essentially doing a Wald test and under the null hypothesis, that independent variable has no effect, no association with the dependent variable. in other words, that the true coefficient is equal to 0. Well, the Wald test is going to deliver you a statistic, which is kind of close to 0. If the coefficient is close to 0, that's evidence that the coefficient truly is 0. It's just randomly varying around 0. So, Wald test are the way we typically do that, sometimes we'll use a Z test version of the Wald test.

Sometimes we'll use a F-test or a T-test version of the Wald test. The difference between the two doesn't matter in sample sizes that are large enough. In small samples, there is a difference, and the F-test version is important. But when you run R, it's going to give you the F or a t-test version of that, so you're safe to go. And Wald test can be also done for multiple parameters. So, just as you could take each estimate and divide it by its standard error to get a separate Wald test for that coefficient, you can actually do it for a group of coefficients, to ask if a bunch of coefficients are all equal to 0, kind of get the job done all in once, instead of doing it individually. It turns out that the math of that's a little more complicated and involves matrix math. And, you can see some on the top of this slide here, where it looks at take some, vectors or list of numbers and takes the inverse of a matrix in between and then multiplies the vector again. It's the math is a little more complicated, but this can all be done in R using some simple commands. So, that's all I'm going to say about multivariable linear regression before, we actually move on and do some examples in R. In the next video, we're going to talk about the assumptions and properties of least squares estimators.

Video 5: Properties and Assumption of Linear Regression

So, in this video, we're going to go over the properties and assumptions of linear regression. So, first thing I'm going to start with is the assumptions. So, there's really four or five main assumptions that underlie linear regression and I'm going to go over each one of these. So, briefly the Model is right, Independence, Exogeneity, Homogeneity of the variance, and Normality or a Large sample. So, the first assumption is that your model is correct. So, what does that

mean? Well, if you're writing down a linear model, you're basically you're relying on the fact that the true phenomena you're trying to explain is approximately linear. It's efficient that it's approximately. If it's far from linear, well, then your model is going to give you answers which might lead you a little astray. So, that's the basic idea. But keep in mind that, we could make the model more complicated, but then it gets harder to interpret. So, a linear model is nice and simple. but with that nice and simple interpretation comes the reliance on the fact that you're assuming that the true phenomenon is approximately linear. The second assumption we make is independence of the records or think of it is like if you have your data in a spreadsheet and each row is a different observation or like, for instance, patient as often the case and health data science. what that means is that the rows for each patient is independent of the other. So, reasons to question that would be, for instance, if you're getting data from family members. So, you have a dataset and a 100 people that happens to be from 20 different families. Well, you might not rely on the assumption of independence because some of the results from people in the same family we might worry or more associate it have more in common than they do with other people in other families.

And the reason that becomes a problem is that if you have lack of independence, it creates problems when we're trying to come up with the standard errors. So, standard errors can be wrong, if you have big violations of this independence assumption. The other assumption we're making is Exogeneity. This is a word from Economics. The word we typically use in Statistics is confounding. Or in epidemiology, we use the term confounding as well. In a linear model, you're taking a dependent variable, and you're writing it as a function, as a linear combination of independent variables and then some noise. The noise is the error. Also called the Error Term is the variation you can't explain with the independent variables. So, we're making the assumption that those the noise is independent of your Covariance. And this is a rather abstract notion which might become a little more less abstract. When we talk about confounding in upcoming module. Another assumption we make is Homogeneity. And here we're assuming that the noise term, like the variation that cannot be explained by the independent variables, we're assuming that doesn't really change from observation, observation to observation or from for instance, person to person, that the variance does not change between people. So,metimes there might be reasons to not believe that. So, actually, to give you an example, which I mentioned in the previous module, So, in the dataset, I'm looking at mortality across different American cities. It's not hard to appreciate that when I come up with a mortality calculation for a city, say, like New York or Houston or Chicago or Los Angeles, which are all huge cities. It's mortality rate, it's going to be more precise than a calculation made for a smaller city like Burlington, Vermont. So, what's mortality rate?

It's kind of the less precise estimates. It's going to vary more. So, when I use multivariable linear regression for that, I have to account for this lack of Homogeneity of the variance or the precision of the dependent variable across the observations. But often we don't have to worry about that. In the final assumption that we need to address is the Normality of the residuals. So, whenever you hear the word Normality, just think bell-shape distribution. Sometimes people refer to it as Gaussian, so this assumption is needed if you have a small sample. So, if you're doing linear regression with just like 20 or 30 or 40 observations, you're going to be relying on this assumption of normality. But as your sample sizes get bigger, you don't really need this assumption. So, keep that in mind. If you have a small sample size, you're going to rely on the assumption of normality. You have big sample size. This assumption not needed, just

not relevant anymore. And by the way, the reason it's not relevant is because of the beauty of something called the Central Limit theorem. The central limit theorem is something that says that certain estimators get or certain statistics get more and more bell shaped. And that's what we can depend on in big sample sizes. So, with all those assumptions, well, then we'll have the properties. Some of those assumptions are not met. These properties might not hold, but of those assumptions are met, we're going to have the following properties. Number one. We're going to come up with estimators that are unbiased. So, unbiased means that they're on target. So, you can think of like using an arrow like your arrow is hitting the target on average, That's what this is getting out, as opposed to being, hitting far to the left or far to the right. if you have all the assumptions met, then you can be confident that your estimators, your coefficients, your slopes.

They're on target. You can also be confident that the confidence intervals that we're going to be driving. So, when we make an interval around the estimate, and call it a 95% confidence interval that it is going to be capturing, the true value on target as opposed to being off a lot of the time. So, they're kind of related concepts, unbiasedness of the estimators and unbiasedness of the confidence intervals. So, when we use a 95% confidence interval, we're counting on the fact that it contains the true value 95% of the time. Well, if some of the assumptions that we discussed or not met it might not actually be containing the true value 95% of the time, it might be 10% or it might be 99.9%. It depends. So, that's something to keep in mind unless you have those assumptions being met. The unbiasedness of the estimators and the 95% confidence of confidence interval might not be holding. Also, there's this idea of unbiasedness supplying to test. So, when we construct the test, and we looked at the p-value, we're usually looking to see if that p-values is less than 0.05 That's our usual cut off. Under the Null hypothesis and of the assumptions are met that p-value should be less than 0.05 by chance alone just 5% of the time. And that's why we use a 5% cut off. Is that by chance alone we're accepting that it's going to be off 5% of the time, but if some of those assumptions are not met, it actually might be more than 5% of the time. But if the assumptions are met when we use the p-values from linear regression, we can be confident that about 5% of the time they're going to be wrong. Whereas if the assumptions are not met, it might be more than 5% of the time. And the final property we have is something called Efficiency. That just means that linear regression is coming up with the best estimators possible.

That's all it say, you can be confident that least squares estimation is the right tool for the job, and there's lots of math to show it. So, now, having talked about the assumptions and properties, What we're going to be doing in a subsequent video is actually utilizing linear regression in R and interpreting the output. Talking about the estimators and the standard errors and their confidence intervals and the test making use of the p-values. That's coming up in the next video.

Demo Video 2: Performing Linear Regression in RStudio

So, we're back to using R. So, what we're going to do in this session is some regressions, or linear regressions. So, as you can see, I have RStudio open and I have an R Markdown file, which means I have my code in these chunks. So, in this first chunk, you can see I'm going to do the following. We're going to make a scatter plot of what hemoglobin A1c at

baseline looks like in relation to hemoglobin A1C at follow up. So, that would be the first part of this chunk. And then we're going to calculate the correlation between those same two variables. And then we're going to do our first linear regression. So, I'm going to do all those three things, a scatter plot, a correlation, and a simple linear regression. That's what we're doing right now. And so, look at the plot first. So, this is a scatter plot, of how hemoglobin A1c at the end of the study, that's the vertical axis, looks like in relation to hemoglobin A1C at baseline. And what you can see there is a real strong association between those two variables, which is, you know, not surprising, because this is hemoglobin A1C measured in the same individual at two different time points. And though they are correlated. So, if your hemoglobin A1C was high at the beginning, chances are it's going to be high at the end. The scatter plot suggests a strong association between the two. And so what we could do is, number one, calculate a correlation, that's what's right here, in this particular box. You can see the correlation calculated between the two variables, and what you see there is that the correlation is equal to 0.73.

And it also comes along with a test, that the two variables are related. So, the p-value there is suggesting, again very significant p-value with many many zeros. So, to remind you, a p-value less than 0.05 suggests statistical significance. Here it's actually has many zeros. It's much lower than 0.05. So, a strong correlation between hemoglobin A1C at the beginning and hemoglobin A1C at the end. In other words, your glucose control at the beginning is definitely correlated with your glucose control at the end of study. So, that's the correlation. The other way of doing this, which is mathematically equivalent but the output's going look different, is the linear regression. So, this is going to be the first linear regression or first statistical linear model we run, and here you can see the output. So, when you run R, you get this particular output. So, right there it says that the top exactly what we did. It's a linear model relating hemoglobin A1c at follow up to hemoglobin A1c at baseline. You can ignore the residuals and just go down here to where it gets the coefficient. And, so, what this is telling us, this part of the output is what we want, this particular line here where it says hemoglobin A1c at baseline. You can ignore where it says the intercept. That row is of less interest to us. This is the role of particular interest. So, in relating hemoglobin A1c at follow up to hemoglobin A1c at baseline, this is telling us that for every increase in one unit of hemoglobin A1c at baseline, we can expect hemoglobin A1c at follow up to go up by 0.74 units. And it's telling us that this is a very statistically significant model.

In other words, that hemoglobin A1c at baseline definitely predicts hemoglobin A1c at follow up. Just as the Pearson correlation above told us at 0.73. Pearson correlation or correlation is very much the same thing as doing a linear regression like this. Sometimes people call it simple linear regression, where you just have one variable related to the other. That's the same as doing an examination of correlation. So, now we're going to move on to linear regression, that's a little more complicated. We're going to start throwing in some more variables. So, in this line right here, what we're going to do, is that we're going to relate hemoglobin A1c at follow up to not just hemoglobin A1c at baseline, but also the gender of the individual, whether they were female or not. So, we'll run that. And then, at the same time, in the same chunk, not that we have to, we're actually going to run a multivariable model that's even bigger than that. So, my idea here is, just so you know, that when you want multivariable models, you can put as many variables in on the right as you wish. It's always just going to be one on the left. You can move on and putting more than one on the left, but on the right, when you run linear models, or statistical models in general, you can put any number of variables in on the

right-hand side. So, let's just run these two examples of multivariable linear models, and we say multivariable because there's more than one variable on the right-hand side, in each of these. And let's see what the results are. So, in relating hemoglobin A1c at follow up to hemoglobin A1c at baseline, and gender, we get the following results for this multivariable model. So, here's where our interest is going to lie, in these two rows.

One row for hemoglobin A1c at baseline, and one row for sex. And what this is telling us is, in terms of statistical significance, as we already identified hemoglobin A1c at baseline is a strong predictor than hemoglobin A1c at follow up. Sex, on the other hand, is not related. So, sex of the patient is not predicting their hemoglobin A1c at follow up, but hemoglobin A1c at baseline definitely is. The one thing to keep in mind, any time you interpret a multivariable linear model is that when we interpret the coefficients, we interpret them as follows. So, this is what we'd say about how hemoglobin A1c at baseline relates to hemoglobin A1c at follow up. For every increase in hemoglobin A1c at baseline of one unit, we can expect hemoglobin A1c at follow up to go up by 0.74 units. And the next part's important, controlling for the other covariance. So, notice I'm adding this additional language, controlling for, or you could say, adjusting for. But the important point here is that when we put other variables in the model, we're controlling for each other. We could also say the same about gender or sex here. So, the difference between females and males, controlling for hemoglobin A1c at baseline, is that females on average have a 0.05 unit higher hemoglobin A1c at follow up, which does not happen to be statistically significant. So, this is important, because sometimes, or often, you might have a covariant, which is an important predictor of something, but when you control for other covariant it's no longer important. So, I'm going to go down now to just down a little bit farther, and now with the even bigger multivariable regression model.

So, in this multivariable regression model, we have not just hemoglobin A1c at baseline, and sex, but also age, and BMI, and systolic blood pressure. So, they're all in the model at the same time. And if we go down here, this is where interest is going to be, in particular, in the p-value column. And for the estimates, as well as their standard errors. So, important point, statistics is very much about quantifying uncertainty, and that's what standard errors do. So, in terms of what's important, based on the output here, well, it's like we've been learning. Hemoglobin A1c at baseline is a strong predictor of hemoglobin A1c at follow up. A very significant p-value, much less than 0.05. Everything else in the model, not important. So, we can say about sex, and age, and BMI, and systolic blood pressure, that each of these is not statistically significant. They do not explain variation in hemoglobin A1c at follow up, at least when we control for each other. So, my point there being that if we put some of these covariants in by themselves, they might be important. But the fact is that we when we put the model in, the only one that stands out as being important now is hemoglobin A1c at baseline. Everything else is no longer explaining variation in hemoglobin A1c at follow up, if the other variables are in the model. So, that's how I'll conclude our introduction to linear models and multivariable linear models in R. And in some subsequent R demos, I'll talk about use of categorical variables, and collinearity, and confounding.

Video 6: Linear Regression and Analysis of Variance (ANOVA)

So, in this video, we are going to talk about the relationship between linear regression and something called ANOVA. ANOVA is an acronym for analysis of variance. So, I am just going to make the connection between those, just like I make a connection between regression and two sample T Test and regression and Pearson Correlations. So, ANOVA is yet another, statistical concept which fits into the regression or linear model umbrella. So, one way to appreciate the value of ANOVA is actually by talking about categorical variables. So, an example of a categorical variable is race, for instance, like white, black, Asian, native American or Pacific Islander. Sex, typically male, female, we could have more categories. So, there is categorical variables often arise in statistics. Sometimes is they arise as ordinary variables. all So, for instance, you might have rankings like is low, moderate, high. Those are often the kind of variables that appear in datasets. So, if we are modeling the dependence of dependent variable, Y on a categorical variable X So, we have a dependent variable Y but just think on the right our X is no longer continue it's actually this categorical variable. So, you no longer necessarily visualizing a scatter plot. You might be going back to thinking about a bunch of box plots. So, we have say, we are going to model how, systolic blood pressure depends on race. So, you might have a distribution for whites distribution for blacks blood pressure, a distribution for Asians, a distribution for Native American and Pacific Islanders. so but what actually goes on when we run linear regression is that in the background, how we actually capture or and code that categorical variable is we create a bunch of what is called Indicator Variables.

So, what is actually going to happen is that well, usually we make a choice about what is the referent category and in R, the referent category usually becomes the first category. So, if your race variable is coded as 1,2,3,4 which happens to be white, black, Asian, Native American, or Pacific Islander. Category one which in this case is whites will become the reference group. If it was coded differently, like one was Asians and two was blacks and three was whites and four was Native Americans and Pacific Islanders R would still take that category one. But this time it is Asians and now it becomes a reference group. So, but just keep in mind that and are you do have the power to change the reference group. so But what is going to go on in a linear model is that in order to do the linear regression and do this least squares, it is going to create an indicator variable for each group except the reference group. So, for instance, whites will be the reference group. Blacks will be getting indicator variable, which means the following you are going to have the variable in the dataset and you do not have to do this R does this in the background it is going to create a variable of ones and zeros, one if the person is black and zero, if they are not black and then for Asians, it will create its own indicator variable, which will be one if that person is Asian and zero otherwise, And for the fourth category, which is call it other at this point, it will be one if they are of other race and zero otherwise so that would mean in your dataset, a person who was black would end up being have a one where it says where the variable is black and zero where it says Asian and zero where it says where the variable is other. So, indicator variables is how, linear models takes care of categorical variables.

But the important point for you as a user using R is to keep in mind what the reference group is because everything is going to be contrast that are compared to the reference group. So, just to tell you what analysis of variances if you are not familiar, But what analysis of variance does what it means? Basically, it is just a generalization of two sample T test to more than two samples. So, whatever you appreciate about two sample T-test, just think that ANOVA is taking the number two and making it bigger. So, it might be three samples or four samples, or how many samples. That is what

analysis of variance is, but nicely enough, the linear model is a way of doing ANOVA without nova actually typing ANOVA into R you can use the LM command and kind of get done what you would do with analysis of variance. so when you do ANOVA, you know, one thing you could do would be just to, like, take a difference. A pair wise difference of all the different groups like group one versus group two group one versus group three, group two versus group three. You could do all those possible combinations. That is the charm of the ANOVA. It gives you a single P value, as opposed to doing a bunch of two sample T-test, and that is why people use it. You get a single P value, not just many. And so a generalization of ANOVA called Analysis of Covariance. So, analysis of covariance is basically ANOVA in which you add one more covariant. And so we have already done this, so in our example, where we had a linear model with hemoglobin A1C at baseline and gender in the model. So, I wrote in the R code I wrote hemoglobin A1C, Hb A1C at baseline and female. Whether a person is female or not, you can think of that being an analysis of covariance where you have a categorical variable which is being female or not, and you have a continuous covariance.

Which is the hemoglobin A1C at baseline. So, analysis of covariance to those who refer to it basically means a linear model where you have a single categorical variable and a single continuous variable. I only mention it not that I think you have to ever call it that itself because I just call it the linear model. But some people will talk about analysis of covariance and you just have to know that they are speaking about a linear model where you have a categorical variable and a continuous variable. It is a special case of regression or a linear model. So, in our next video, we are going to be talking about confounding and Collinearity, two really important concepts in biostatistics and Epidemiology as well as economics. And anytime linear regression is used.

Video 7: Cofounding and Collinearity

In this video, we're going to learn about the concepts of confounding in the related idea of collinearity. Often, in the sciences terminology could be very important. So, but confounding and endogeneity are two names from different sciences for the same thing. So again, confounding is what we usually say in the health sciences, like Biostatisticians and Epidemiologists. But economist and those in the social sciences usually refer to this concept as Endogeneity. So, if you're talking to an Epidemiologist or Biostatistician, you use the term confounding. If you're talking to an Economist, you would say endogeneity, and depending and if they speak the other language, you may or may not confuse them. But the best way to understand it, is a graphic like this. Sometimes this typo graphics are called a DAG. D A G. It's an acronym for D A G is for a Directed Acyclic Graph, and what that means is you have these nodes labeled X, Y, and Z. You have arrows and the arrows only go in one direction. So, that's what the directed graph means, you have arrows. And the other thing, acyclic means there's no loops. So, in this particular acyclic graph I've drawn, you see one arrow points from Z to X, one points from Z to Y and one points from X to Y. You don't have any looping around where the arrows will go pointing, looping around. The key point I want to use here about this particular DAG or Direct Acyclic Graph is that it really helps us to understand confounding. So, the confounding issues the following. We are interested in understanding how this variable X affects or is associated, how it relates to this variable Y? So, Y is the dependent variable, X is independent.

We want to know how X affects Y. That's our key question. So, it might be in the Health Science if you want to know how the dosage of this drug X affects this health outcome Y, for instance, weight or blood pressure. But what we're worried about is a possible confounder Z. A confounder is this other variable that has impacts on both the exposure of interest X and the dependent variable Y. So, there's a confounder that affects both of them. And if you have that, you have problems. So, what confounding does, it confounds. How you interpret the association between X and Y. If you're studying the effect of intake of wine, like how much a person drinks, how many servings of wine a person has in a week and how it influences some cardiovascular endpoint, such as like their blood pressure, their cholesterol level or whether actually they have a heart attack in the next five years. So, if you're studying that, there's been a lot of study of how alcohol intake affects health outcomes, some of which suggest it's negative and some suggest, maybe it's okay. But always in moderation, if it's going to be okay. That's definitely one of the strong findings. But something that might really confound that association is income. So, people with higher incomes, they just tend to drink more wine and they also happen to have better health outcomes for other reasons besides, from drinking wine, no doubt. So, income of the person would be a confounding variable. If you're studying the influence of wine intake on health outcomes. So, that's a favorite example. And what's it going to do. it's going to lead you to find an association and maybe doesn't exist. So, the person with higher income, they have better health outcomes.

They just happen to drink more wine, too. So, when you look at the association of wine intake and health outcomes, you're going to see a correlation which maybe isn't actually due to the fact that wine is, wine intake is good for you. It just happens to be that the wine drinkers have higher incomes, and that has other health, has well documented health benefits for many other reasons because they live in safer neighborhoods, with less pollution, etc., We are going to do that in R and I'll show you how to take care of that and in R the solution is pretty simple. You put the confounding variable in your linear model, and that mostly will take care of the confounding problem. So, you have a linear model where you have the health outcome, like maybe their cholesterol level, and you're regressing it on wine intake. But because you're worried that income is a confounding variable, you're going to put that in the model as well. So, I'll talk about that in a future when we get to it, using R, how to interpret the output from a multi variable model where you include the confounder. If you have more than one confounder, put all of them in as well, and that mostly will take care of your confounding issue. The only limitation that we have to face is that, what if you're, for instance, going back to the wine example, what if you're studying the effect of wine on health outcomes, but your dataset doesn't have any information on income levels, then you can't control for it. And so in that case, when you get your output, you're just, you're going to have confounded results that, when you interpret it, it's going to be a huge limitation when you write it up to share with other researchers and the media for instance. So, that's confounding. Collinearity is a related topic. So, two or more variables are collinear if they're all linear combination of each other.

So, for instance, like a very simple example. If you had in your dataset temperature measured in Celsius as well as temperature measured in Fahrenheit. So, you have two different rows and you can't put them both in the model, that might seem like obvious. But if you put both in the model, it's going to create issues, because they're collinear. So, I was talking about coding rays. If you have, if you've coded rays and actually created the indicator variables yourself like you actually created an indicator variable for white, an indicator variable for black, an indicator variable for others, say, we're

combining everyone that's non, black, and white. If you put all three covariants in a model, you're going to get mathematical problems like your linear regression is going to break down and run properly. So, what you have to do in that case is that you have to leave out one of those three categorical indicator variables. Because the problem is that if you put in all three, they're collinear. If you know two of them, you know, the third one. That's what we mean by collinearity. So, collinearity is a, can be a mathematical problem. It can also be more of a epidemiological or statistical problem in the following sense. So, maybe you don't have two variables which are perfectly correlated. But you have two variables which are highly correlated. So, for instance, if you're modeling the impact of weight on some health outcomes, so measuring how weight affects whether people have a stroke in the next 10 years. If you put in your dataset as independent variables body mass index, weight, and waist circumference, they're all highly correlated. So, if you end up putting all three in the model, you might get away with it. You might get results, but interpreting those coefficients is going to be really tricky. So, whereas you can do it, usually we abstain by putting more than one highly collinear variable in a model, because it really makes it hard to interpret the effects of variables.

So, to give you another example, suppose you're looking at some, you're looking at health outcomes. So, for instance, whether people are going to have a heart attack or die in the next 10 years and you're looking at the effect of education. So, suppose you've measured education in more than one way. You measured like years of education, you actually asked people how many years have you been educated like ranging, maybe, from eight if they just had an elementary school education or less than that, or up to like 16 or 20 if they have, like multiple graduate degrees. But you also ask people what was their highest graduate degrees. So you have, like years of education. Plus, like whether their highest degree was about high school or Bachelor's degree, Master's degree, PhD or MD. If you put both of these in a model, you can, and sometimes it will be okay to do this. But if you put both in the model, it makes really, it makes a really tricky to interpret the coefficients you get for those. So, for instance, if you put both in the model years of education and highest degree obtained, so if you look at the coefficient for whether a person had a high school education or not, when we have a multi variable model, we always have to interpret that, like holding all other variables constant and so holding, how can you hold other variables constant which are highly collinear, and that's where it lays in the problem. So, anytime we interpret coefficients from all the multi-variable model, the proper interpretation is: This coefficient corresponds to how much you can expect the dependent variable to change. If you change this independent variable by one unit, keeping all the other variable, all the other variables constant. And if you have two highly collinear variables, it's really hard to make that interpretation work because you can't hold years of education constant while increasing from a high school degree to a University degree. There in lies the problem, which makes it tricky to interpret multi variable linear regression when you have collinear or highly collinear, correlated variables in the model. So, the next thing we're going to do is run R, some multi variable models with confounding and collinearity.

Demo Video 3 Name: Multivariable Linear Regression in RStudio

So, what we're going to do in this demo is... It shows multivariable linear modeling, also known as multivariable linear regression. And we're going to start now by showing what happens when you have categorical variables in the linear regression. So, that's what this chunk is doing. This chunk is going to first create a new variable, which is going to be a categorization of BMI or Body mass index. Body mass index is a measure of weight, body mass, level of obesity. In the

American population, it probably averages around 27, 28, maybe 29. And we're going to categorize it into three groups - less than 30, 30 to 35, and 35 and above. And then we're going to use that in our multivariable linear model. So, you can see here, in this chunk, after creating this categorized BMI variable, we're going to run a linear model with that relates hemoglobin A1c at the end of the study to hemoglobin A1c at the beginning of the study, sex and age, and this BMI categorized variable and systolic blood pressure. So, this is the output of the chunk here and let's look at it. So, how does BMI relate to hemoglobin A1c at follow up controlling for everything else in the model. So, here is the output telling us how BMI categorize relates to hemoglobin A1c at follow up. It's in these two rows right here. Our interest would quickly go to the P Value column as this one right here, and you see two P values way above 0.05. So, what immediate reaction might be if you go over? I don't care about BMI categorization because it's not important. That would be a legitimate interpretation that BMI, at least categorized right here, does not explain variation in hemoglobin A1c at follow up. That's the dependent variable. But, we have to be careful when we interpret the results of multivariable linear models because we always have to do it as follows: So, what is the effect of BMI categorized into these three categories on the dependent variable controlling for.

So, that's important to keep in mind is that when we use a multivariable linear model, the coefficients we get have to be interpreted using this language of controlling for or adjusting for. So, hemoglobin A1c at follow up controlling for hemoglobin A1c at baseline and sex and age and systolic blood pressure, it's not significant. How do we interpret those coefficients over here? Well, this coefficient right here, we would not really be interested in interpreting it because it's not significant. But, it's saying that people in the BMI category of 30 to 35 compared to people in the lowest BMI category have hemoglobin A1c's of 0.015 units lower. So, that's the important thing to keep in mind is that when you have a categorical variable, it already has say three categories. Well, you're going to get two coefficients, not three, because one of the categories becomes the referent category. In R typically by default, the referent category is the lowest level and the lowest level here is the BMI less than 30. So, this is comparing people in the middle category, the people in the lowest level of BMI, and this coefficient right here, -0.126, is comparing people with BMI's of 35 and above to people in the lowest BMI category. So, it's saying that the dependent variable on them is -0.026 lower on controlling for hemoglobin A1c at baseline and sex and age and systolic blood pressure. Just below that, we have what's called the ANOVA version or representation of the model. ANOVA is an acronym for Analysis of Variance. And in this version of the representation of linear model, really what we're getting is a single P value.

So, whereas BMI has three categories and when we do a linear model, we get two coefficients and two P values. When we do this to ANOVA version, we just get one P value. So, it packs those two coefficients into one piece of information and gives us a single P value. Answering the question does Hemoglobin A1c at follow up, the dependent variable, does it vary with respect to the categorized BMI? And what it's telling us is, well, it's not significant, just like we kind of knew already by looking at those two coefficients above. So, essentially, these two rows get collapsed into one row down here and get this as a single P value. That's what the ANOVA representation is doing. So, the next chunk I want to run is illustrating some of the ideas of collinearity. So, collinearity is the idea that if two variables are correlated, for instance BMI and hemoglobin A1c at baseline, they're associated. I might not have shown that, but they're associated and so it

turns out that if you put two correlated variables into a model, they're going to compete to explain variation in the dependent variable. So, let me illustrate that. So, right here. How does BMI relate to hemoglobin A1c at follow up.

So, can you predict hemoglobin A1c at follow up using BMI? Well, that's what this first part of the output here is telling us and the answer is a resounding yes. As BMI rises, hemoglobin A1c at follow up also rises very significant P value. So, this tells us that the P value has nine zeros before you get to the first nine zero decimal. And it's suggesting that for every unit increase in BMI of one unit, hemoglobin A1c at follow up is going to go up by 0.20 units. So, that's very interesting, very strong correlation or connection between BMI and hemoglobin A1c at the end of the study. But is BMI important in predicting hemoglobin A1c at the end of the study? If we actually control for hemoglobin A1c at baseline, that's what this linear model is addressing. So, this linear model is addressing the question. Does BMI predict glucose control at the end of the study controlling for glucose control at baseline? That's what this output is telling us. You're answering. And the answer is, if you just look right here, does BMI predict hemoglobin A1c at follow up controlling for hemoglobin A1c at baseline? And, to get the answer, we can go to the final column, which is the P value and in this case, what we're getting is no. So, that corresponds to a P value if we put it in regular notation of 0.81. So, the P value is much higher than 0.05 And so we would conclude that, you know, BMI is not related to hemoglobin A1c at the end of the study if indeed we control for where they started from with hemoglobin A1c. So, it's important point to get is that if two variables are correlated and each predicts its dependent variable, if you put them both in the model, one of them might no longer be important.

Video 8: Nonlinearity and Interactions

This final video is on nonlinearity and interactions. So, if you're modeling the joint effect of two variables on the dependent variable, it's not necessary to stick the linearity. So, one way we might break from the linearity is actually creating what we call an interaction of the two variables. So, for instance, if you have, if you to use the example we've been using in R, if we're trying to predict glucose control three months from now and we have glucose control today, baseline and we have gender, we might ask, is there an interaction of the two? Well, not only does being female have an effect on your glucose control three months from now but does being female affect how your glucose control today affects your glucose control three months from now? That's where interactions are about. There are going a little more higher order. So, one way to make a nonlinear model is to take two covariates and create a product of the two and that product of the two we call an interaction. So, in this slide, you see we write, the dependent variable Y as the intercept term, I say A sub zero here. And then we have the first variable X and a coefficient for it. And the second variable, the second covariate of interest X two and a covariate for it, and then we have an interaction term and an altogether different coefficient for it. So, the interaction term gets its own coefficient. And so, despite the fact, this is still a linear model, I like to emphasize this. This is still a linear model but we're actually accommodating a nonlinear effect or an interaction. So, that's something I like to emphasize is that with, despite the name linear models, you can deal with nonlinearity. It might sound like a contradiction, an oxymoron, but indeed that's the case.

So, often in the machine learning community you'll criticize linear models as being nonlinear. Well, in fact, you can accommodate nonlinearity with linear models. So, it's an important point I like to make. So, these interactions, sometimes they're highly important, as I was getting at when we're interpreting results. So, for instance, I was making the example that we could put in a term there for the interaction of gender or sex or being female or not, with glucose control at baseline and see, and we would end up using that coefficient and whether it's significant or not to say whether sex modifies the effect of glucose control at baseline on glucose control at follow up. And so it might be other examples. You might put, if we had a race variable, we might ask, does race interact with glucose control? And this could also refer to the treatment effect. So, if we're looking at the treatment effect, we could put in an interaction of treatment and sex and ask, does sex modify the effect of treatment? So, if the treatment lowers glucose control or improved glucose control by this amount in males, is the effect in females even better? This is a very common thing to do when reporting the results of clinical trials, you know. Typically, what you'll do is report the effect of the intervention. For instance, in this COVID era, vaccines, they're going to be a lot of those vaccine trials being done in the next six months and no doubt when they report the results, they'll break them down by gender or by race. And so each time they do that, they'll actually probably do at some point in a model put, for instance, vaccine and gender or vaccine and race and interaction of those in the model and test to see if it's significant. And if it's significant, it will be able to tell them, oh the effect of the vaccine is different between males and females or different between blacks and whites or blacks and Asians, etc.,

So, that's how interactions can be taking advantage of in linear models or any kind of model, such as logistic regression models as we'll get to in the next module. And this is all the other name for this is effect modification. So, effect modification addresses the question, does a certain categorical variable or even sometimes continuous variable. Does it modify the effect of another variable? And effect modification is all addressed using interaction so they go hand in hand. People are speaking about effect modification probably means they're talking about the use of an interaction term in a model or they should have done that at some point. So, the next form of nonlinearity that's important is actually to go beyond the effect of a continuous covariate. So, for instance, so you know, I asked you to remember like, high school in drawing, you know, slopes on the board. But if you remember, you know, linear equation. You know Y equal A plus BX, but you can actually move beyond that to quadratic effects. So, it is possible within a linear model to study a quadratic effect. So, again, it might seem like a bit of an oxymoron. But you can study quadratic fits, or shapes, reforms. People have different names for this idea within a linear model. And for instance, all you have to do is create a new variable, which is a quadratic variable. So, let's talk about glucose control at baseline versus glucose control at follow up. So, we already established that glucose control at baseline has an effect on glucose control at follow up. When we actually do this scatter plot, we could see that it was followed a straight line, but we could return and try to fit a quadratic model to it. And I would have to do is take hemoglobin A1c at baseline, create a square term, put that in the model, and end up estimating a coefficient for it.

And if that's coefficient was non zero or significantly different from non zero, we conclude that it has a quadratic effect or departure from linearity. So, that's your other take home point I want you to take. Linear models can address nonlinearity, it's an important point. The final point I'll make about in this session about nonlinearity and interactions is transformations. So, sometimes what you want to do is transform your dependent variable. So, we've been using in

our examples, human hemoglobin A1 c wants to get follow up. And it has, like, a nice bell shaped distribution. It turns out that there's really no reason to transform it, but sometimes you're going to come across, dependent variables, which are highly skewed. An example of this is in economics. When people are studying, often they're studying monetary outcomes, which are dollars like the cost of healthcare in a year, for instance. What you often find with variables like that as well as with a lot of biomarkers we use in medicine. A lot of like measurements that we make such as biomarkers like cholesterol. They often have these skewed distributions, and when you have a skewed distribution like that, what you might want to do is take a log transform.

There's other transforms you can take like a square root or a cube root. But a log transform is by far the most popular type of transform, which is only possible if you have any zeros in your dependent variable. You can't take a log because the log of zero mathematically is negative infinity, so it creates a problem. But if you don't have any zeros, certainly consider taking a log. And the nice thing about taking log transformations is that then when you interpret the coefficients, they end up having a nice interpretation as well and it's all comes down to a percentage scale. So, in that case, if you've done a log transformation, if you take coefficient and exponentiate it, that means that you can say, make a conclusion of the following that when you increase your independent variable by one unit, the percentage increase in your dependent variable is such and such. So, using a log transformations means you can make interpretations in terms of percentage changes, and that's a nice applications of log transforms. So, I hope you enjoyed this module on linear regression. The next module is going to cover logistic regression. So, we'll cover a lot of the same ideas but what's going to change is that no longer do we have a continuous dependent variable. We have a dichotomies dependent variable.

Demo Video 4: Correlation- Regression diabetes in Python

Hello. Today we're going to do correlation with linear regression for Diabetes Treatment Dataset. As usual, we're going to import the packages that we're going to need. I might import some others further down. Then we're going to get the data frame. Then we're going to use the diabetes, we're going to convert this diabetes treatment CSV file into a data frame called df. Simplicity. Going to look at the head of it. And so here's the Glucose Treatment, the Baseline, Hemoglobin A1c Follow up, Female, null says bool, Age, BMI, and systolic blood pressure. Going to look at the information here. It just tells us the types, the floats, and the bools number, no missing data. That's all very good. Going to a quick statistical describe here. Okay, now we're going to plot the hemoglobin A1c follow up versus the hemoglobin A 1c baseline for each patient. This is patient by patient, of course. Then we're going to look at some correlations between the follow up and the baseline. And then we're going to use, linear regression to look at the correlations between those things, and also some other features. We're going to add in some features as we go along on. This is why I'm going to use the method below. But before that, let's plot the scatter plot here. We'll see this nice, positive correlation here. If it's perfect, it would be a plus one. It's not quite perfect. If it was, if it is perfectly uncorrelated, negatively correlated, it would be a minus one. The R-value will be minus one. It'll be this way. And if it's perfectly uncorrelated, it would be, it's scattered all over the place. And that would be a zero. So, plus one for positive R-value, perfectly correlated, positively correlated. Minus one for a perfectly negative correlated, and zero for no correlation at all. And here we see, this is the independent variables, this is the dependent variable. That's the way we set it up. And again, I just, what I did here is I just got the data frame and just ask for those columns. And that was one feature against another.

Now we're going to look at the correlations, but let's put some more numbers, some numbers on this rather than just a visual. We're going to do with more visuals, but we'll also put some numbers in them. So, let's just remind ourselves about the columns. Here they are. Then I'm going to, this is a nice, quick way to do this. I'm going to pass these features in. So, I'm going to make a list out of these columns or these features, and just call it features. And then I'm going to pass that into the data frame as a list, and it creates another little list, which I called a hemoglobin A1c_df. And that allows, so I'm going to run this first, I may have to run that. Here you have to run, make sure you run them in sequence. And that should give me this little data frame here. Because now you see, what I can do is actually run that as a pair plot. I'll just have a quick look at it. Which is everything you want to do as a machine learning data scientist. You want to do this sort of thing. So here, each feature is compared to each other. When the feature is compared to itself, it has a little plot, little histogram plot here. And you can see they're more or less normal distributions. This one's got a bit of a skew on it. This one looks quite good. This time we're going to do, we'll look at the correlation between them instead of visually like that with a heat map. So, I'm just going to do a Matplotlib type plot. Do the figure size. But then with a heat map, so I put that into plotted, actually, but to get the heat map, I'm going to use the Seaborn library. The alias is SNS for that. And we're going to pass it in the data frame, the small data frame I created, plus the correlation method. And we're going to also annotation to be true. So, we're going to get this out. Just be careful here. You can easily mix this up being negative down here because it's black. But actually, there's no zero in here. So, this is actually a positive, entirely positive scale. So, and you can select different palettes. So, if you don't like these colors, you can, I'll show you how to do that in a moment.

So, here we see the correlation between the follow up and the baseline. Hemoglobin A1c is 0.73. As we saw before, it's a quick way of looking at it. I'm going to get a table at now. So if I actually, what I can do is create another little data frame called the correlation data frame, and I can do that by just calling the hemoglobin A1c data frame and getting the method, but this time passing the parameter, the method I want, the method to be pairs in correlation. There's other correlations you can use, but this is a good one, and then I'm going to display that. And then I'm also going to ask for the heat map again with the correlation df. So I'm passing at this. I'm going to pass that in. I'm going to ask for the cmap because SNS diverging palette knows there's a dot there. And if you change these numbers, you'll see that the colors will change, and you want some contrast in your map as well. So, let's run that quickly. So, you get this nice little table here. And if you have a lot of features, it would do this as well. And so you can easily look at the correlation in a table form. And again we see that between the baseline and the follow up, 0.73. And we get this nice contrast visual as well, the heat map. Now I want to do this, this was just another way with the statistical package, but this time, I'm going to use a scIpy package. I think I imported it further up, but I want to emphasize this. So import scIpy. I'm going to get the stats package from that and then run the Pearson Correlation, pearsonr, it says on that. And I'm going to input X and Y, the input and the output. And I'm going to create these little data frames out of just passing in df, the actual baseline and the follow up. So, the input is the baseline, and the output is the follow up. And then this actual, if we run this, it returns two values. One is the R-value, correlation R-value. And one is the P-value here. And then we could print those out.

Let's have a quick look like, run that, I have to make sure. And so here's the correlation, 0.7311. We've seen this before, lots of decimal points here. I don't think we really need this. But the P-value, as we saw before is almost zero, look at that, it's 10 to the minus 88. You can't get too much closer to zero in that. But, so, it's well below the 0.05 value that we consider for a 95% confidence interval. And that means that it's, since zero is well below 0.05, it means that this is a very statistical event. A strong statistical significance. And so this means this is a very good correlation and a very good value. So, that's one way to get that out very quickly. Now there's another way to do this using linear regression, but not the linear regression I'd normally use, but one that's in the stat package, a very simple one. There's just, I'd only used for a couple of features, the two features we're considering. And so this is the scIpy.stats.linregress model, and I'm going to pass X and Y again from up here. This is the X and Y. We're going to pass in again. And here we're going to print out the result of the slope, the R-value, the standard error, and the P-value. Remember, the slope is actually the regression coefficient. It's the number that goes in there. I'll talk about this a little more. Although I'll talk about that, on the next time. Okay, I'm going to leave that discussion just for a bit longer. Talk about this, but let's run this and see what happens. So, there's the regression coefficient, 0.74. Here's the R-value, 0.73. They are very different things, by the way, don't get them mixed up. There is some sort of relationship between them, of course, but the values, but not the values. So, just be careful and I'll explain that a little bit more. And here's the P-value. So, that's another way of doing it using linear regression instead of Pearson Correlation coefficient. We use the best fit line. It's trying to fit the best fit line again there, but within the stats package.

This is another one. And now we're going to look at, looking at correlations using a simple model of linear regression. And here, we're going to, again, set up the features as the baseline. I'm doing it this way because we're going to be adding in some features in a moment, and then the label like this, and then what we're going to do is create the X and Y again using the features and the label, those lists were passing in. and we're going to import now the stats.api package or library, I should say. And we're going to use the alias SN. So, let's go ahead and run that. There's more information here for the stats models, and the regression OLS model, the linear regression OLS model we can use. OLS just stands for ordinary least squares fit. It's a very simple one. I wouldn't use it. Usually I use the Scikit-Learn linear regression model where you can tweak the parameters more and it's a, and you can split the data in the way you want it to train and the testing data. But for this, it's sufficient for what we're doing right now. We'll talk about the other stuff later on in another video. So, here what I'm asking it for, this add constant, by the way, on the model SN is just asking it to consider an intercept. The interceptors, they intercept of the best fit line to the Y axis. All right, the intercept. Just Y equals MX plus C. Y is the axis. X is the axis. And then M is the gradient, and the C at the end, the MX plus C is the intercept between the best fit line and the Y axis. So, I'm passing in X again from up here, and we're going to do the ordinary least squares fit. We're going to call the model, just regcorr, call it what you like. You can call it reg, or LM, or anything like that. And then we're going to do the fit. So, I'm going to use the model name and then we're going to do a fit on it, this is where it fits the line. And then we're going to call that stats, and then we're going to print stats summary.

We're going to get a lot of information out here and it's going to be, I'll explain it, there's quite a bit, but let me just go through and run this. So, here's the information. So, OLS Regression Results. And first of all, the R-squared, is actually the square of the correlation coefficient, just to let you know. R-squared really tells you how well your model's doing.

And your model, that really means how well your features explaining your output or predicting your output. That's really what R-square does. But the, and it's called the Correlation of determination. And, but our R-correlation or R-value is the square root of that, if you want that. And the reason I'm telling you that is because down here is the information, down here is the information I really want. This information here. And the constant is the intercept, so I'm not really interested in that. It has one parameter. The other parameter of a linear regression of a straight line is the regression coefficient, which is the slope. Alright, so here's the slope of the line. Here's the intercept of the line, and it's about the same value as before, 0.7426. And if you notice, it's different, too. If I took the square to this, I'd get that 0.73 number. I did actually check that. I think it's 0.7311. It's different to that, of course and the reason it is different because it's the regression coefficient. So again, if you had that line, if you had a line, Y = three X, then, in that case, your regression coefficient would be three. That would be the value of the slope, it's the number before the X.

However, if that, if we look at the data for Y = three X, and it was perfectly correlated, it could be no more, the R value could be no more than plus one. And so, in that case, the R value or the correlation coefficient would be close to plus one, but the regression coefficient would be three, a rather different value. They are related in some way, but value wise, they're not. So, you have to be very careful. The P value here, it's over here. This tells you that this is a significant result for the constant, for the intercept. So, that's a good value. And this also tells you, this number here is P value for the regression coefficient, tells you that the slope or the regressed line for the between the baseline, the hemoglobin A1C baseline, and the hemoglobin A1C follow up is a significant result as well. Now, this means basically, to show us the slope, that if you go along the X axis, which is the baseline, hemoglobin A1C baseline by 0.7426 units, you would get one unit up on the follow up. All right? And that's basically, that's basically the definition of the slope, isn't it really. So it's just the, this is the slope if you want to think of it like that. Right. So, that's all that data. So, the one we're really interested in is just this bit here. But this is also very useful information. Later on, we'll be looking at this, and this stuff down here as well, because that gives us a lot of information about the model. I'm going to put in another feature here. The female, the gender. Well, we call it female, true or false. But you can't put that easily into a regression model, certainly not the ones we're using. So, we need to convert it from a bull into a number like 01. So, when the female is true, where we can turn that into a one, and when the female is false, we'll turn that into a zero. There's ways of doing that with things called dummy variables. Bit of overkill for us here, we'll talk about this another time. But you could just call this method or this function on that particular column, and now convert it into another column.

So, I'm asking it to convert it into another column in that data frame called the gender. And if I run that, I'll see. There it is. There's the gender here. So, when the female true 1 1, and then we had a false here zero, this one here. So, we see that since we're working okay. A very quick way to change those. And now we've got that, what we're going to do, we're going to add the gender into the features, and then we're going to keep the label to follow up the same. We're going to just have the follow up as the output. And then we're going to put the features and the label into another data frame, basically X and Y. And then we're going to pass that into again, do the same thing as before, into the ordinary least squares. Regression model here. Again, it's okay doing correlations and that on doing this way, but I actually wouldn't do it, do a machine learning model with this, this particular regression OLS. And then we'll run this and see what we get. Now we see that we've got the gender curve regression coefficient as well. And again, some people will actually use

these regression coefficients as denoting the importance of that feature in predicting the output. There are some things you have to do, like normalize your data, you should really normalize your data on that to do that. And so it's a rough way of seeing the feature importance. But there are some pitfalls, so you have to just be a little careful. For this data, it said it's fine, but it does give us an idea that this certainly is more important. The regression curves which is more important or the regression. So, I should say the HbA1C.Baseline is a more important feature in predicting the follow up output for the gender because this is a very low number here. And we can see that the correlation is not significant between the gender and the follow up, hemoglobin A1C follow up, because the P value is well above 0.05, so it's not significant in denoting and predicting the output value of the follow up. I'll talk about that a little more in a moment, but I just want to, we're going to add this Multivariate Features now using a simple Linear Regression model. We're going to use the same one, the OLS, simplicity and speed.

So, let's look ahead again, just to make sure. Now, we're going to look at, we're going to put all of these in. We're going to put the baseline, the gender, age, BMI, and systolic blood pressure. And the output is going to be the follow up. We're going to put that features and labels again. Get X and Y, put that in. Let's run this quickly. Now what we'll see here, actually, that these numbers change slightly. That might just be the way they actually do it. There might be some, you know, sampling error and that sort of thing. But there could well be a change here with the regression coefficient. Certainly we see that this is still very significant, in particular because of the P value, we can double check that, that it has a regression coefficient that's quite large, which means, which indicates that a lot of the variation in the output is caused by the variation in the baseline. That's another way of putting it. And, but we see the other ones are quite low, and their P values are quite high. They're above 0.5, so they're not significant in determining the output particularly. But we have to be a little careful here as well, because things could change when you go from just comparing the baseline to the follow up. It might have a different regression coefficient because now you're adjusting basically, you're adjusting.

Remember, you're doing a best fit data with this regression model. You're adjusting that best fit line, so you're trying to fit all of these features. And that will, that means you have to take account of all the other features. And so there will be some change if you were to run just the gender. So this one, if we ran the systolic blood pressure against the hemoglobin A1C follow up, we might find that it was significant, here it say it's not significant because 0.257 is above 0.05. But if we just ran it against the follow up, just one on one, we might find it is. However, when we adjust that best fit line for all the other features, it might make it seem like it's not a significant as it truly is. So, that you have to be a little careful there as well. There are methods of finding how significant actually these are, there are some different methods. But looking at this now, we can see, generally speaking, that hemoglobin A1C is better at predicting the output or the hemoglobin A1C follow up than any of the other features. And according to this statistics, none of the others are very important and they're not highly correlated or they're not correlated significantly with the hemoglobin A1C follow up output either. Just one thing here. It does say this condition is very large down here, but this, it might be pointing to the fact that there are some, there's some strong collinearity between some of the variables. We have to be careful about that as well.

Demo Video 5: Hypothesis Testing in Python

Hello Welcome to using t-tests for testing the significance of the glucose treatment in diabetes. As usual, we're going to import some libraries as we go along here and notice I'm going to be importing this researchpy library as the alias rp and then the usual ones here, as well. We're going to be using the researchpy right at the end. I'm going to go and get the diabetes treatment as a .csv file read it into the data frame; going to call it df. Going to look at the head. There's the head of it. Notice that we have the glucose treatment for the patients either true or false, as of all here. We're going to be looking at the baseline, the follow up as well. We're not really going to be using the other information here. Let's look at the information, though, in general. So, I've got 520 entries for each column. But the thing we're going to need, we're going to be looking at the change because the patients are going to come in and there's a baseline for the HbA1c. There's going to be a baseline taken and then, sometime time later, there's going to be a follow up, and they're going to take the reading of the hemoglobin A1c. HbA1c gives you a measure of the possibility of diabetes and you want to keep an eye on that. So, you want to keep eye on the change. The lower the HbA1c, the better. And if it goes a negative change would be good because you're lowering it. So, it'll be good to know look at the negative changes as well. I'm going to get that change by getting the data frame; getting this column out of the data frames, subtracting it or subtracting the baseline from that, from the follow up. You could do this other way around, doesn't really matter. But you have to know what sense is better than the other, negative or positive. Let's run that and look at the head. So now, I added this column, the HbA1c to to this data frame here.

But just remember, we're looking at all the patients at the moment: those that had the glucose treatment and those that didn't. So, we have to do something about that because we want to know, we want to test to see if the change in the HbA1c is significantly different between patients that received the glucose treatment and The mean of the population in this case is for the change. We want to compare it to see if there's no change; compare the actual values to no change. The mean of the population for the change would be zero, in this case. Before doing that, we have to create a new data frame, which is going to be the data frame for the glucose treatment; we're going to call it df_gt. I'm going to put those values into the data of frame, only if the glucose treatment values are true. Then we can look at the head of that. There's the change and notice that glucose treatment, they're all true; and everything else correlates with that across. So, everything still in synchronization there. Let's just do a quick describe on this change, as well. Remember, this is just for the patients who had the glucose treatment; just looking at that change. There it is. So, by the way, we can see the count here is 278 out of 520, actually had the treatment. The mean of the change for this population, the treatment population, is -0.89 that seems to indicate it'd to go down; it's a negative number. Then we can have a look at the quartiles: 25%, 50%, and 75%, as well. That gives us just some idea to look at the data there, the maximum and the minimum, of course. This is the one we're really interested in here: How much difference is there between this mean and the population mean? We could also plot this. It's important before we do a t-test, by the way, it's very important to plot the data and look at the distribution. Because the problem is if you do a distribution, if you use the distribution that's not the normal distribution, you have to be careful which t-test you use or which statistical test to use; or you should normalize it or standardize it. I should say really. So, let's have a look at that. Let's just, by the way, here's the link to the seaborn plots, histogram plot. I'm going to do distribution plot. Same thing, really. And I'm going to set a few things up there.

Let's run that and see. So, we're looking here at histogram plot. We've got 10 bins and we're looking at the change of the HbA1c for the patients who had the glucose treatment against the frequency of that. For this value here, about -1, there's almost 60 patients that had that. Notice, though, that the change of the zero change. So, this normal distribution is slightly skewed to one side because it seems that all the patients did move down a bit, which is a good thing if we're looking at the change because, remember, we're subtracting the follow up from baseline. That looks good and it looks like a normal distribution. We can always check for a normal distribution, by the way, by using a stats.shapiro. Shapiro is a measure of the normality of a distribution. For this particular glucose treatment data frame, for the change, I can run a shapiro test on it. This gives a P-value of 0.911 which is above 0.05. Remember, if it's above 0.5 from 95% confidence interval, that implies there's no significance. So, this is saying there's no significance between this data distribution and a normal distribution. Basically, that's what it's saying. And we're not gonna, this is the T statistic for that. We're not going to worry about that for this purpose, but this is the P value that comes out to that. Okay, so that's good.

It's a normal distribution. Now we're going to, although we gave it, sort of, looked to it and it looked okay, we really want to know if there is a difference significant difference, between the mean of this sampled population and zero, the mean of zero. We're going to use from the statistical package from Python, we're going to do the t-test; we're going to do a one sample t-test. That's this here. And we're going to do it on the data frame for the glucose test and we're going to pass it in the HbA1c change. We're going to compare it to the, what's called, this is a parameter that's passed in, popmean, population mean. We're going to set that to zero. So, let's just go ahead and run that. Let's see what happens. Again, this statistic is T statistic. Here, we see that the P value is very low, actually, because it's in scientific notation here, -18. That's a very low number. It's much less than 0.05. Now, anything below 0.05 for 95% confidence interval, it tells you that there is some significance. There's a significant result here that's not just due to random fluctuation of sampling. There's something else going on. There is does seem to be some change in the sample population here. That so, for this one, because there is a significant change, we reject the null hypothesis, there's no difference and we accept the alternative that the change here, the HbA1c change mean, is significantly different from the population mean that we set to zero. Just to clarify that. Now, If we want to test to see if the change in the HbA1c is a significantly different between patients that received the glucose treatment and those that did not receive it, we have to do a similar process.

We're going to use a slightly different test. But what we're going to have to do, first of all, we're going to have to create another small data frame that has the glucose treatment treatment set to false for those who did not receive it, in other words. Here we've got df_Nogt glucose treatment and we're going to look at the head of that. Run that. Here we see they're all false, the patients now we're dealing with, and here we see their changes. We're going to do a quick describe on this. There's 242 patients but look at this mean it's -0.874, I think the other one was -0.896, I think it was, for the mean there. So, It doesn't look like there's a big difference between the means of these two populations, the ones who had the treatment and didn't for their HbA1c change. Data here, and notice that we are using the data frame. Just be careful here. I'm using this df_Nogt glucose treatment. So, here is the frequency against the HbA1c change for the people who didn't have any treatment and there did seem to be a slight change as well. This is the zero and this is, you know, for those patients who didn't have the treatment. It shifted little bit, as well. But let's see, let's just do those two describes together. Let's look at the glucose treatment patients, the date for that and for the no glucose. So, let's just do the two

here. Then, I'm going to run this one. Just get it in screen here quickly. So, yeah that means zero, this is the people who, the patients who had the treatment the mean is that and for people who didn't the mean of the change, remember? Look at the mean of the change.

It's that. So, there don't seem to be too much difference between them in terms of their means for the change. But let's see the significant difference between those two populations now. Let's have a look at it first. So, what I'm going to do first is doing a visualization, the difference in change between treatment and no treatment. Here's a nice link here to the Seaborn pydata.org documentation. They explain different ways of doing the boxplot, whisker.boxplot. We're going to use sns.boxplot. We're going to pass it the glucose treatment. We're going to pass it as the Y axis the hemoglobin A1C changed. Notice that now, instead of I'm going to pass it the original data frame containing both groups, the patient group. I'm sorry, the patients who received the glucose treatment and those who did not. I'm going to order now by true and false for those because if you remember, it was given determined, the bull determined where they received or not, true and false. So, let's just run that. And here is the boxplot for that. Here's the change here, and by comparing the, whether this is the glucose treatment the patients got the treatment or they didn't. This line determines the mean of each one. And around it, of course is the quartiles. These were probably outliers here because they're outside of these Q three, Q four quartiles. So, yes, well, there looks like there experience sort of overlap between those two, doesn't it? There's not much difference. So, looking at it visually, we can see there's not a great deal of difference. certainly the true one seemed to patients who had the treatment seemed to drop a little more. There change was more negative than the ones who didn't have the treatment? But there's other factors, of course.

Maybe these people excised, I'm not sure. But then there's always other factors you have to be careful to take into consideration, which we will do later on as well. Let's do a test on that and the test I'm going to use. Now, If we were just comparing before and after for each patient in the group, just before and after for each patient for the same patient, we could use a paired T test, but because we're looking at the change for each of the groups within themselves, the ones who had the treatment and the ones that didn't, we can't use a pair test there. We have to use a two sample test. When you run that in the Stats TT test, you have to call it an independent two sample test. So, there is a, you have to be a little careful, which T test to use. So, here I'm going to use the stats T test independent and here's the link to the documentation of that. And I'm going to pass in from the data frame of the glucose treated patients, the hemoglobin A1c change. And also we're going to compare it to the data frame for the non glucose treatment patients. We are going to pass it the change as well. And let's just run that test. Now the P value here, we won't talk about this one here, but let's talk about the P value. The P value here is much greater than 0.5 so there is no significant result here. So, since there's no significant difference between the means of the two populations for those treated and those not treated. We will accept the null hypothesis which states that. Now there is another way to do this. I thought I'd introduce you to it. The research pie. You have to install it first of all, so you can go to the Anaconda prompt and you could just do a Pip install researchpy.

If you have some problems, you could do a pip three install researchpy. But mine work just for the pip install researchpy, and it'll install it for you. Just best to use the Anaconda prompt because then it installs it into the Jupyter package. You don't have to do this, but I thought I just introduced it. And here's the documentation, very very, good for the documentation. It does a lot of things as you can see. So, I'm going to do just the TT test. Now the researchpy, which is the alias research by a set up further up, you can depends how you set this TT test up here. I'm just using it as two sample independent test here basically. So, let me run that and show you what you get out of this. So, again I'm passing it. The df_gt, the glucose tested the glucose treatment and the non glucose treatment using same change seem, comparing the two populations there. We get out a series form for this and you can see the data. If you like looking at it like that. Let's just look, if you this actual test here returns two Python object. One that describes the information one that gives you the results. So, I'm going to ask for it for the describe. And so now I'm going to run this and again going to pass those same things. But now I'm going to look at the objects and this will give me a sort of data frame format. So, here is the data frame format, and the nice thing is that you can compare here is the hemoglobin A1c change for the glucose treatment patients and this is the one for the non glucose treatment patients down here. And it compares the two different things side by side here, which is very nice. And you can look at all those things you notice we use in the 95% confidence interval. Now, if I look at the results for that, it puts it like this, which I like. This is the T statistics, which we spoke about before. It's just another metric.

Here's the P value, which we're interested in and 0.8842 again, not significant there. This T value, actually the coast of zero it also implies it's not a significant difference between the means of the population so of the change, the population or the sample populations of the change. That's what I mean by that. So, here is the two sided T test for the P value. For that. There you go. Which is what we've got before but it's using the different research packet. So, this gives you more information. There's a lot if you look at the link I gave you. There's all sorts of other information you can glean from the data with these tests as well. Next video we're going to be looking the regression of this linear regression of this data. Thank you.

Demo Video 6: Performing linear Regression in Python

Hello. In this video, we are going to be looking at Linear Regression, ANOVA, and Collinearity for the Diabetes Dataset. And we are going to be comparing all the features, the input to the hemoglobin A1c follow up as well. We are going to be doing that, actually, separately, and then altogether. And discussing how this pertains to collinearity. So, let's get those libraries, and we are going to do, doing the usual, We are going to go and get the diabetes treatment that's there using the V CSV method. We can look at the information here, as usual, do a describe, and next Run. Just note here that the maximum for the BMI is about 40. I'm going to be using that information in one of the examples further down. Now we are going to be using BMI. What we are going to do in this video is look at how to actually deal with categorical variables as well. And we are going to be using BMI for this, but BMI is an important aspect of diabetes because it relates to the obesity of somebody, and obesity is related to diabetes. So, we should see some relationship between BMI and the HbA1c baseline follow up. We are going to be talking about these categories in relation to collinearity with other features that we are also inputting as well. But before I do that, I have to sort something out, I had a problem with using the dot here, and so I am going to look for one of the libraries down below. So, I'm going to change it to underscore here,

and then just put another column in with that, use that instead. So, I just created these two columns on the end, these two on the end here. I just wanted to. And then we are going to change this Boolean here into a type int, so we can use it easily in our linear regression. There it is. Changed over.

All right, now, before we proceed, we have to actually change the BMI information into categorical variables so we can play around with them. There is two ways to do that in Python, there's pandas.qcut, and pandas.cut. They do have some differences, that's why I reference them so you can look at their subtle differences. But just to explain it quickly, if I look at the pandas series here form, and I looked at a range of numbers, 1 to 12, and I am going to chop those up into three parts. And then I could look at the value counts of that on the back end. So, this is a quick example. And so it's chopped it into this ranges, shows you the range, and it shows you how many in each, 4, 4, 4. So, let us apply that to the BMI. So, we are going to use in that particular column in the data frame, the diabetes data frame, we are going to chop it into three, then look at the output of that using the pd.qcut here. And so we see it's not quite symmetric because it's defining it by three, but it's more or less chopped up to three equal parts. And then we can assign that output to another column. We are going to call that column BMI category, and we'll print that out, have a look at it. There it is. We see the BMI category. And, I'm also going to show you how to do it with a cut, and the reason I'm doing this as well, so I'm going to put it into specific ranges as well. But you could do that using the cut method here, they call it bins. Sometimes this is called binning actually, three bins. And so I'm going to pass these ranges. Zero to 30, 30 to 35, and 35 to 40, as a list into the, for the bins parameter in the method cut, as cat bins, and then run that. And we should see that we have three different ranges here, and as we can see here in the head, and they are consistent with the R video ones.

Now we are going to do some maneuver tests on the categorical BMI to see how we handle categorical features. And another course stands for analysis of variance. And it has several tests in, like R squared and the F statistic, and other things, different methods within and over table, they're going to get printed out eventually. There is many ways to do this in Python, but I have picked one that is very similar, actually to the R method of doing it. I'll read this out afterwards because I want to show you. We are going to just do the summary as we have done before. Notice the format, now I'm doing the stats=ols and I'm using a very similar format to the R thing here. I'm going to pass all of these features in, including the BMI category. Also going to use, you know, data's going to be the df, the data frame, and we are going to do a fit. We are going to do a linear regression fit, fit the best fit line to that. And then we are going to get a summary out here. So, if we look at this now, let's read through it. In this summary of the follow up against all other features, we see that the P-values for the BMI categories have values greater than 0.05 which implies no significance. So, we can go down there, here and see this. If we look, I'll just move this up a little bit here, if we look at these categories here and notice, first of all, I should say, it's only going to be two, and the reason is, when you do it this way of categorizing a particular feature, you'll always going to have one less for the actual regression coefficients here for the row. Because the first one, the actual zero to 30 range, the lowest one, is used as a reference. So, it's always referenced. It's similar to dummy variables, you usually have a dummy reference, if you're familiar with that. And, so, you won't see it, you'll only see the two. And this coefficient is saying that, well, it's not terribly significant.

There is not much change in the slope of it. Now, you can see if you draw a line, the slope of it would be, it would reference to the follow up data, the hemoglobin A1C follow up. It would have a very significant line. For one unit of this, the hemoglobin follow up would go down by that amount. And similarly, for this one, for one unit of the BMI category, along X axis, the follow up will just go down by that amount, so it doesn't seem very significant. We can definitely look at the P-values for these, and these two P-values here 0.938, 0.630 are way above 0.05. So, that means, they're not significant. There is no statistical significance in there in predicting the behavior of the HbA1C follow up. Another way of saying that is that they do not explain significantly the variation in the hemoglobin A1C follow up output data, or predicted levels. Conversely, we see that here the HbA1c baseline has quite a strong, you know, coefficient value. It's, you know, the slope, in other words, a slope, if you draw a graph of the Hemoglobin A1c baseline against the hemoglobin, A1c follow up with all the other features involved, you still get a very quite a slope on that graph, it seems to, so it seems to implying that, the some of the variation in the hemoglobin A1c baseline is actually explaining the variation in the hemoglobin A1c follow up output of the model. And we can see that with a P-value, we can see it's significant in explaining the, or predicting the output because it's P-value is just about zero here. Now, just to be clear here, we have to make sure we understand that when we talk about, when we're using the BMI categories here, we are actually controlling for all of these other features here.

We're controlling for all these other features. We're adjusting for all these other features. And so, you know, they are all sort of competing in some way to explain the variation. And we're going to talk about this a liitle bit later on with the Collinearity, that if there is some co variation between some of these features, they'll compete to explain that. And so we find here, actually, that as we'll see in a moment, BMI category does actually come into play when it's all by itself. But here, when it's controlling for, or adjusting for the other features that have been input to the model, these features here, we find that it doesn't seem to be very significant at all, which seems to imply that there is some sort of Collinearity going on between these features, which we'll talk about more in a moment. So, you have to be, when you're dealing with these particular features in a regression model like this, you have to be very aware of this information, and you have to analyze it very carefully. They therefore do not significantly predict the HbA1c follow up level, or output when controlling for, or adjusting for the other input features. In this case, prediction means that these categories do not explain the variation in HbA1c follow up level. Note, and I did mention this about the, there'll only be two of the coefficients, regression coefficients because of the reference one. All right, so, let's look now what happens when we just print out the ANOVA test here. So, this is a way of doing an ANOVA test, or using the statistics package from Python. We just use this method here, anova_lm, which stands for linear model. We pass it the stats that we're going to get out of this ordinary linear, ordinary least squares regression, and then we'll print it out.

Type 2 just means it cuts down on the output of the different columns. I'm cutting out some of the columns we don't need. You could leave that out if you want to. And here we see it again, so this is just another way of doing it, this is the Anova table that you get out for a linear regression, and it is doing the P-value, the F-value, sum of the squares, that sort of thing. We're interested, particularly, in the P-value here. So, for the BMI category, which is all lumped together, by the way, when you do the Anova, it's all lumped together. And we're just going to get one P-value for it, it's not going to separate out those categories and, sort of, lumping them together. And we see here, that when we're looking at the P-

value for the significance for explaining the output, or predicting the output. When it's controlling for all the other, or adjusting for all the other features, didn't quite get that there. It's actually got a P-value of that comes out to 0.862, alright, it's just ten to the minus 1. So again, it's not a statistically significant in explaining the output of the model. Once again here, however, this one is, this is almost zero, which is the Hemoglobin A1C baseline. And then, these are not, none of these are high, it's mainly this when they're all controlling for each other basically. Now that brings us on to collinearity, because now we're going to discuss this in more detail. When we actually look at the BMI category by itself, against the... in predicting the Hemoglobin A1C follow up. So collinearity, as professor McKenzie said, "If you put two correlated variables into a model, they will compete to explain variation in the model output." That's it in a nutshell basically, and so you have to be very aware of that. There is a test, a several test for this as well, that you can use.

So, previously, we saw that when we combined all the other features with BMI, including Hemoglobin A1C baseline, the BMI did not significantly predict the output to explain the variance in the output. However, let's see what happens when we just use BMI by itself as a predictor of the Hemoglobin A1C follow up levels. Here we can see that the regression coefficient is sizable and the P-value is very significant, implying that BMI is of significant important in predicting the Hemoglobin A1C follow up. And I should put after this, when it's not controlling for anything else. So here, we're just comparing, or we're looking at the relationship between BMI and, as an input to the model, and the Hemoglobin A1C follow up, as the output; just that feature. And we're going to do the same, we're going to look at the summary here. And I'm not going to look at the top stuff, look at this again. And we see here for P-value its just about zero. So, when the BMI is not being used, controlling the other features in conjunction with the other, in controlling or adjusting with other features, the actual P-value is extremely low, which means that it is significant in explaining variance in the output, or predicting the output, the Hemoglobin A1C follow up levels. So this means, that perhaps, that there is something that's collinear, that's covarying, with the BMI as well. That's an indication of that sort of thing here. And we can look at that in the thing that is most likely to be doing that, of course, is actually the baseline. We'll talk about that in a second, but we can look at this. Supposing now the BMI, when we're looking at it, is controlling for now the Hemoglobin A1C baseline here. What do we get?

Let's look at that summary. And certainly here, when we look at this now, we see once again, when we're controlling for... when BMI is controlling for... adjusting for the Hemoglobin A1C baseline, we notice its P-value is no longer significant, because the Hemoglobin A1C baseline is competing with it to explain the output. And it's really, it's highly significant explaining the output, where it sort of pushed out the BMI as being a main, important feature in explaining the output. And so there's definitely some collinearity going on there. So, in conclusion, down here, while BMI related by itself to Hemoglobin A1C follow up is significant in predicting output, its P-value then, if you remember in the previous table was almost zero, if controlled from where the patient's started, i.e., the baseline, it is not significantly related to Hemoglobin A1C follow up. Effectively two input features Hemoglobin A1C baseline and BMI are collinear. And, so that's another thing you have to be careful of. There was actually a test in here, I think it's the Durbin-Watson test, but I'd have to look into that again, give a better indication test for that more severely. So, that's actually important. Collinearity is an important thing, particularly if you're looking at a particular feature and how it's related to the output, you have to be aware that there is possible competition from other collinear features in there. I put this in here if

everyone to also run the Anova test on it. Now, I want to introduce you quickly to another method of doing linear regression, in Python, which is a little more thorough in the sense from a machine learning point of view.

It's a very useful tool to have, and I'll explain why in a moment. We're going to get the same data frame. I did it a different way with another data frame, the stuff I took out with other data, but here's the diabetes treatment data. We're going to put it into the data head. We're going to do the usual thing with the female, true false data, put into agenda, form, going to do that. And then, instead of doing the ordinarily squares... Well actually, let's look at the columns, let me just look at that, check that, see what we want. What I'm going to do here, I'm going to do something very simple. I'm going to call the features the Hemoglobin A1C baseline. And the target, the output, sometimes here called the label, whatever you want to call it, as the Hemoglobin A1C follow up, which is we've already done this sort of thing before. We're using the OLS, simple linear regression model. And I've put this here because you might want to go back and try this as well, with it. I'm putting all the different features, and see what sort of results you get. I'm just going to do this because I want to graph it. If you go back and you run all the features, you take this one out and then use this one. Just remember, you want be able to graph it, because you're going to be really graph a two dimensional line fit. So, I've done it just with the input being the baseline and the output being the follow up. So, let me run that quickly. Then I'm taking the features and creating another little data frame called X, and another data, to put the target into it, creating another little data frame called Y. And, so run those. Then I'm going to go to Sci-kit Learn. Now Sci-kit, is the basic name of it all, but the abbreviated version when you get in the library is sklearn. And this is really the machine learning package for Python; it's very powerful, does a lot of things. And here we're going to go and get a linear model, and the linear model we're going to use is linear regression, et cetera.

Now we're going to import that. We're also going to do something, we're going to do a model selection here on sklearn. This is the thing that will split the data into a training part and a testing part. Generally speaking, machine learning, not so much in statistics, but in machine learning, because the way machine learning works, you want to actually use some data for training. But you don't want to use the data you've trained the model on to test it, because what can happen, you can get what's called overfitting, and that's not good, it gets false outputs for your model. It makes you think your model's better than it is, and then you're going to get a lot of variance on the actual results of your model. That's the whole subject in itself. Then I'm going to look at some metrics and measuring how good our model is. I'm going to use the mean_squared_error, and r2_score. We've come across both of these briefly before, talk about that a bit more, so let's run that. So this is where I split, I'm going to take the X and Y, and I'm going to split it for X and Y into, an x train part for training, a y train for training, and then an x test for testing, and a y test for testing. And you'll see that basically all this does, it separates the inputs and outputs of the model into... If I don't tell it otherwise, into a 50-50 split of test data for the inputs and outputs, and training data for the inputs and outputs, for the train. And then, also, data for the testing, for the inputs and output as well. And the inputs being the Xs and the outputs being the Ys, and so it splits it up. I can tell it, often you split it more like 75- 25, 75% for training, 25% for testing, but we'll just keep it as it defaults to here.

I think it is 50-50, I can not actually remember. So, run that and then what I am going to do, I am going to call my linear regression model. I am just going to call it linrgr here in Linear regression, simplicity. And then once I have produced this model, I am going to do a fit on it. I am going to pass in the x training data and the y training, that is the input training data and the output training data. This is the stuff you are trying to predict, right. This is the target. And then we will run that, and it gives you some information. I have not set too many other variables in there are too many parameters, I should say. Then what we are going to do is going to score it. So, now we have done a best fit line and I am going to see how well this best fit line is actually doing. I mean, score it. This basically tells me roughly this is really the R squared value of the score, When you do the dot score on this and it is telling me that, of course I am using the test value is not, I did the training. I did the training up here, but now I am passing it the test data or to test it with different data has not been trained on. This is telling me that how much of the input is explaining the variance in the output, how well is your model actually modeling the data, basically, and it is giving me about 0.5 value here. You should do a thing called cross validation with this, basically, but we will talk about that another time. I am going to do this in a different way and look at this in a different way. There is many ways to do these things in Python, so, I am going to look at how are my model is predicting the outputs its going to produce early predictive output, so I am going to run my model basically, with the test data, I am going to call the method predict on it, which is going to create a whole list of predictions for the input data.

We will use that later on this y predict these are the predictions of my model for the outputs, given the input test data, then I am going to look at the linear regression coefficients for all the different input variables. I only suddenly picked one here, but you can run it with all, the others, if you like. Then I am going to look at what is called the coefficient of determination, which is the square of the correlation coefficient or the R value. That is another measure that really measures the how well, this that is the same as the score up here. Basically, there are subtle difference between the dot score and the R squared. Here, we are going to pass into the R square. We are going to look at how well do their model compare, So, this is the actual ground truth, y test data. The actual data from the data set and this is what for the outputs and this is what in comparison we can use that our models predicted and we will do exactly the same thing. We are going to do that for the R squared, and we are going to do it for the mean squared errors well down here. The mean squared error, just compares data for data the how well the y test does against the prediction outputs here. Okay, and it is just another measure of we do squares you do not get negatives. It is just another measure of how much error is between our output and what the actual ground truth data is saying. So, let us run that we get some output there. So again, we saw the regression coefficient when we compared the baseline haemoglobin to the follow up haemoglobin. It is about the same as we got before, In our previous analysis, 0.74 and the coefficient of determination R squared basically tells us how much you know you can interpret it as being how much one particular feature caused the output to be the way it is. But, more to strict way of saying is how much of the variance in that particular feature caused the variation in the output.

The mean squared error rather is as well, and that is just one against the other is that is quite a fair that it is not a terribly high value, but it is not exactly low seemed better. So, we are saying, basically, with this is saying that yes, the Haemoglobin A1c baseline does explain a lot about the Haemoglobin, A1c follow up levels. However, it is not explaining

all of it because it is only 0.5. It is about 50% of explanation. There is about 50% explanation for the follow up being the levels they are or it is caused by other features like BMI, age, possibly gender and systolic blood pressure. So, we are only comparing one feature in there. So if you run it, you look at this data again you should get a different you probably get a very different value here I will leave that up to you to do. I just want to plot this and the reason I picked those just two features so I can plot it and so I am going to import seaborn as sns the alias, as we have done before, matplotlib, pyplot actually some extensions and the A list is going to be plt and going to plot that in line. I put this in this way, In case you want to mess around with this, you could do this look at BMI against follow up, but I am actually going to do the baseline against follow up. And I am going to also draw, if you notice I am passing it the x test data. And so this is the ground truth. x input test data for the haemoglobin A1c baseline. But I am going to put the line that is going along output prediction, which is basically the best fit line in this model. This is what our models producing this sort of line here. So, let us see what we get when we run this. And there it is, and that is that was what our...this is the data. This is the ground truth data of we got from our data set haemoglobin baseline follow up. This line is actually our model prediction here. Thank you.