**Week 2 OLS**

**Pages 1-2**

I am Dr Chao-Yo Cheng. Welcome to Advanced Topics in Quantitative Social Research.

In this video, I will provide a practical overview of multiple linear regression, building on what you have learned so far.

I will start by use the bivariate regression to go over the core concepts and mathematical properties of the linear regression model. I am also going to explain how as quantitative social researchers we can view the data analytical objectives of linear regression differently and how these different perspectives will accordingly influence our understanding and reading of the results we get from conducting linear regression analysis. Our review of the linear regression model will then briefly cover the key assumptions we need as well as the main principles of statistical inferences when it comes to linear regression.

Next, I will discuss the underlying rationale for multiple linear regression. Why has multiple or multivariate linear regression become the most common way of understanding the social world among quantitative social researchers? What issues do we have to take into consideration when we try to build a multiple linear regression model and carry out the analysis? I will end this part by discussing some basic principles when we design and implement our linear regression analysis.

The most important key takeaway from this lecture is to recognize all statistical models are only approximations of the reality. There is no perfect statistical model although we do have some principles that can guide our decisions such as how we choose between different variables to include in our model and which model estimator we should and want to use in our analysis. And yet, for quant social researchers, we will still need to make a lot of decisions and very often these decisions may not have any compelling reason or foundation. After all, as George Box put, “all models are wrong, but some of them are useful.” Box is one of the most important statisticians of all time and this line has become one of the most important mottos for all quantitative researchers.

**Page 3-6**

Page 3

Regression analysis is a very common statistical technique because it allows us to describe and estimate the relationship between different factors or variables in the social world or the human society.

The linear regression model is one of the regression techniques we can use, but it is perhaps the most widely used model estimator because it provides us with a very straightforward and intuitive framework to achieve the analytical objectives I just mentioned – that is, linear regression allows us to use a linear function to present the statistical relationship between X and Y.

Consider the linear function I put here. We will talk about multiple linear regression in a minute but let us focus on bivariate linear regression for now because the basic ideas can be extended to multiple linear regression. Bivariate linear regression only has one predictor or independent variable. In the linear function, we can use X to represent the predictor and Y to represent the outcome or dependent variable of interest. Some researchers like psychologists sometimes also call Y the response variable of X. For this linear function, we can say Y is a function of X because we can use the value of X to infer or derive the corresponding value of Y.

Page 4

After we learn the basics such as what X and Y refer to, we now can go over other parts in the linear function or the linear model more carefully.

When we carry out linear regression analysis, the idea is to use one or more predictors or independent variables to predict the value of Y. Once we use our data to estimate the intercept and slope, then we can use the data we have to predict Y.

On the right-hand side, you can there is an error term at the end, which is to capture the random noise that cannot be captured given our predictors or current model specification. For instance, if we want to come up with a model to predict someone’s income, a person’s luck may can only be included using the error term unless we can come up with a systematic and replicable measure of luck. Formally speaking, the error or disturbance is defined as the difference between the actual and predicted outcome.

Putting the error term aside, in the same function, you can see the model is defined by the intercept and slope in addition to the predictor and outcome variable. On the one hand, the intercept of a linear function will tell us the corresponding value of Y when we set the value of X at zero. The slope, on the other hand, indicates the corresponding change in the value of Y when we increase the value of X by 1 (say from 0 to 1). You will see this more clearly in the next slide.

Page 5

Here you can see what a linear regression model looks like when we only consider one predictor or one X. This model uses beta zero and one to represent the intercept and slope, respectively. The figure is very helpful because it clearly shows where the intercept is located. Once again, the intercept is the value of our dependent variable when we set the value of our predictor at zero. The slope, or beta one, shows the corresponding changes in Y over changes in X. When the denominator is 1, then the slope again tells us how much our dependent variable will change when we increase the value of our predictor by one unit.

Page 6

Here is a more realistic example of a bivariate linear regression. What we have here is a scatter plot containing both the predictor (on the horizontal axis) and the dependent variable (on the vertical axis).

The predictor in this case is voters’ perception of competence for politicians from the Democratic Party in the United States.

The competence score is constructed by the researchers, ranging between 0 and 1, with higher value suggesting higher competence. The dependent variable is how much percentage of votes Democratic candidates can win relative to the vote share of their Republican competitors. Each small circle in the scatter plot represents an observation – and you should be able to see that the unit of observation in the dataset is a Democratic candidate in an election.

Suppose now we carry out linear regression analysis, using the data we have, and derive the estimated coefficients for the intercept and the slope of the regression line. Looking at the regression line, given that the slope is positive, it is apparent that Democratic politicians can win more votes relative to their Republican rivals when the voters consider them more competent.

This scatter plot also shows where the intercept is located. Since the intercept is negative, it tells us when voters consider Democratic politicians having no competence at all, they will lose votes against Republican politicians. In the same scatter plot, you can see there is a small hat above alpha -- this is the common notation researchers use to represent a statistical metric estimated based on the data. The correct way to read alpha hat will be the “estimated” intercept.

Likewise, the scatter plot also uses one of the data points or observations to remind us of the definition of a residual. You can also see small hats being placed by predicted Y and residual epsilon. Residual can also be understood as the estimated error or estimated disturbance. Some researchers have a more nuanced understanding of the difference between error and residual, and I will explain this more carefully in our class.

Before I move on, here is one interesting mathematical fact of regression line you may have noticed: The mean values of the predictor and the dependent variable will always fall on the regression line.

**Page 7**

Again, linear regression is a very convenient technique for us to infer the statistical relationship between the outcome and the predictors. To summarize our discussion so far, the main objective of using linear regression is twofold: First, linear regression is a statistical technique that allows us to use a linear function to represent the correlation between our predictors and outcome variable. This may be a bit too obvious, but this is a very important point to keep in mind. Second, and this is where things become more interesting -- the least squares linear regression will allow us to estimate the values of intercept and slopes with some desirable properties.

I will discuss these properties in a minute. But before I do that, we need to discuss some additional nuances when it comes to linear regression.

You may have heard researchers use different names to call X and Y. So far, we have used “predictor” to call X, but doing so actually implies an assumption about why we need linear regression and what we use linear regression for.

First, also in line with those who invented the basic techniques of linear regression about hundred years ago, we can use the linear regression model for the purpose of forecasting. Using the same case we have just gone through in the previous slide -- it is tempting that we can use voters’ perception of politicians’ competence to predict how these politicians will perform in the elections. Of course, this is a very unrealistic scenario because a politician’s vote share, or how many votes a politician can get in one election, is by no means just a function of a single factor or variable. Sometimes elections may just be influenced by random disturbance that we do not know and other factors we really cannot measure tangibly. But in any case, if you believe the main objective of linear regression is forecasting, then X can be understood as the predictor of our outcome or dependent variable.

Similarly, some researchers believe we can use linear regression to explain the outcome of our interest. Using the same example from the previous slide, we can say it may be the case that a politician’s competence may not reliably predict how many votes they will receive, but it does help researchers explain a politician’s performance in elections. In this case, we do not care much about perception of competence as a predictor but just treat it as one of the factors that we can use to understand voting behaviors in democratic elections. In this case, X is better to be understood as an explanatory variable.

Before I move on to the discussion on the statistical inference of linear regression, I also would like to make two additional points. The slope we estimate using different regression techniques, including the least squares, is only one way for us to measure the correlation between X and Y. Also, with certain assumptions, the slope can tell us the causal effect of X. We will talk more about this later this term, but in the field of causal inference, X is usually understood as the causal or the treatment variable.

**Pages 8-10**

Now we are all clear about the basic setup and data analytical objectives of linear regression, it is time to address a number of questions regarding statistical inference. That is, how do we estimate the intercept and slope using the data we have? How can we know the intercept and slopes we estimated based on the data we have are the best ones? What can we really learn from our regression coefficients for the intercept and slopes?

The (ordinary) least squares, or OLS, is one of the most common techniques we use to estimate the intercept and slope. The intercept and slope estimated using OLS have some very nice properties. One of the most important properties is OLS can produce “unbiased” point estimates, which means OLS can give us the estimated intercept and slope close to the “true” one. Remember our data is always a sample of some well-defined or unknown population, but with some assumptions or conditions we can be more or less certain that, theoretically speaking, the estimated intercept and slope produced by OLS are very similar to, if not exactly identical with, the true intercept and slope that we will never know for sure.

The first condition is straightforward. To use linear regression is to say we assume there is a linear relationship between the predictor and the outcome.

Next, we need the error of each observation to be independent from each other. For instance, if we are trying to use the number of hours you study for a school test to predict the final mark you receive, the estimated error, or the residual, of one student should not depend on the residual of another student.

Meanwhile, we need the value of actual Y to be normally distributed across different levels or values of X – that is, the value of actual Y recorded in the dataset given a specific value/level of X should be included in the corresponding normal distribution of Y at the same value/level of X.

Finally, we need the variance or standard error of our Y to be equal across different values or levels of the predictor – in the textbook they call the condition “equal variance,” and a more technical is homoscedasticity.

Keep in mind: These are the conditions for the estimated regression coefficients for the intercept and slopes to be “unbiased,” which means they are close to or nearly the same as the true ones in the population. The conditions we have gone through above do not necessarily allow OLS to produce efficient regression coefficients or point estimates for the intercept and slopes. To say a point estimate or a regression coefficient is efficient is to say the coefficients or estimates have a relatively small variance or standard error, and the variance or standard error of our coefficients or estimates usually has more to do with the statistical power or the sample size. Very often we can reduce the variance or standard error of our regression estimate by increasing the sample size.

Page 9

This scatterplot summarizes the conditions listed in the previous slide. Again, this is a bivariate linear regression. Each small black dot in the plot is a data point or observation. The thick blue straight line says the values of Y can be approximated by a linear function of X, which is the linearity assumption.

Next, although it is a bit hard to see here, the distance between each black dot and the blue regression line does not have a clear pattern and certainly does not depend on each other. That is, the residual of one observation has nothing to do with the residual of another observation. This condition is the independence assumption.

Furthermore, you can see that across different levels of X, the values of Y follow the same normal distribution. The red curves in the scatterplot show the normality assumption.

Finally, given that they follow the same normal distribution at different mean values as we increase the value of X, it means the errors have the same variance or standard error across different levels/values of X. This is the homoscedasticity distribution. The reason for us to say so is because a normal distribution is defined by its mean and variance, to say two normal distributions at different mean value are identical is also saying they have the same variance.

Often scholars will combine the independence, normality, and homoscedasticity assumptions to say the outcome variable must be independently and identically (normally) distributed across all levels/values of the predictor. This is known as the i.i.d. condition.

Page 10

These assumptions and conditions are surely important as we hope to obtain unbiased estimates for the intercept and slope of the linear regression model.

But I want to make this clear: the violation of these assumptions is not the end of the world. As you can see here: Econometricians, statisticians or quant researchers in other fields have developed many statistical techniques we can use to “correct” or “adjust” our estimates when one or more assumptions are not satisfied.

For instance, generalized linear models can be quite useful if our data do not allow for the linearity assumption. After all, the relationship between X and Y is not always linear. There are also other statistical techniques that we can use to relax these assumptions when we conduct regression analysis. We will not cover these techniques in this module, but you are welcome to request tutorials or additional references if you are interested in these techniques.

**Pages 11-12**

Page 11

Now we are clear about the assumptions, it is time for us to discuss the more important issues of regression analysis: How do we obtain useful information from our analysis? Sure, the regression analysis, especially the least squares, may produce unbiased regression coefficients and point estimates of the intercept and slope, but it is very crucial for us to review some of the key principles when it comes of the task of statistical inferences.

In this lecture, I will focus on two issues related to statistical inference of least squares linear regression, or OLS. One concerns the question of statistical significance and the other is about the question of regression model fit.

First, after we carry out any linear regression analysis, we are keen to find out if the point estimate is significant or not. Given that the intercept usually does not deliver much substantive implications, at this moment we will focus on the regression coefficient for the estimated slope, or beta.

To determine whether the estimated slope is statistically significant or not, we will have to conduct hypothesis testing. You should have learned the rationale and procedures of hypothesis testing in other modules by this point.

As a reminder, the hypothesis testing for OLS is a two-tailed test. As we want to use OLS to estimate the slope, or the correlation between X and Y, the null hypothesis is always beta equals zero. And the alternative hypothesis will be beta is not zero. By saying our estimated slope is statistically significant is basically to say the estimated beta is statistically different from zero for us to say we can reject the null hypothesis.

To determine whether our estimated betas are statistically significance in this sense, we can either use p-value or confidence intervals. If you use p-value, which is defined as a conditional probability that we observe the pattern in the data given that the null hypothesis is true, we will need this probability to be super low for us to say the null hypothesis is so unlikely that we must reject it. And as you may know the conventional threshold is .05.

If you think the idea of p-value is very convoluted, which is very much of the case, an alternative is to find out the confidence intervals of a regression estimate to infer their level of statistical significance. We will then have to assume the sampling distribution of our regression estimates to be normal and use the standard error of this estimate to construct the intervals. The key idea is to make sure the interval, which shows the range of our point estimates or regression coefficients, does not include zero so we can reject the null hypothesis that beta equals zero.

Before I continue, I want to make one thing very, very clear: Statistical significance is neither the only one nor the most crucial aspect of regression analysis. In addition to statistical significance, we should also pay attention to the sign and size of our regression coefficient. A statistically significant coefficient for the slope may be so small that in the end our predictor only has a trivial correlation or causal effect on the outcome variable.

Page 12

We should also pay attention to the question of the model fit. There are many techniques to evaluate whether the proposed model is a good fit, and which one we use will demand a clear conceptualization and operationalization of goodness-of-fit.

A very common indicator for goodness-of-fit is R-squared, which can be casually defined as a fraction of the variance of predicted outcome over the variance of the actual outcome in the data. If our model is a good fit of the data or the outcome variable, the idea here is to make sure this fraction is very close to 1 so we can say the model has done a great job explaining or accounting for the outcome variable of interest.

Alternatively, we can also calculate the sum of squared residuals (SSR) to see if the model performs well by reliably predicting the actual values of our outcome variable, which requires the residual of each observation or data point to be as small as possible. OLS, or the least squares linear regression, is known to produce estimated coefficients for the intercept and slope that result in smallest SSR. This is where the name of “least squares” came from in the first place.

**Page 13-16**

Page 13

So far, we have used bivariate linear regression to go over some of the key concepts and principles of linear regression analysis because many things we have covered so far can be extended to multiple linear regression, which is the linear regression model including more than one predictor or independent variable. There is a number of additional notable distinctions between bivariate and multiple linear regression, and here I want to discuss them more carefully.

Again, multiple linear regression using the least squares has become the most powerful regression technique as it provides us with the means to present and study nuanced, complicated interactive dynamics between different predictors with respect to the outcome or dependent variables of our interest.

We can formally or mathematically specify a multiple linear model in many ways, one of which requires some knowledge of calculus and linear algebra. To simplify our discussion, here is a simple algebraic representation without using any complicated mathematical notations and concepts. The only notation that may look unfamiliar to you is the subscript attached to each predictor – here the number means different predictors. For instance, X one means the first predictor and beta one is the slope or regression coefficient associated with the first predictor we have in the model.

Page 14

For any regression coefficient, as I have explained earlier, it can be understood in different ways, such as forecasting, prediction and causation. Although the underlying idea of multiple linear regression is not so different from bivariate linear regression, but we will have to consider the meaning of slopes with a bit of more caution.

I am not intended to go over the formal foundation in detail, but here is the intuition: In the case of multiple linear regression, given that we have more than one slope, each estimated coefficient for a corresponding slope should be understood as follows: We are using the estimated coefficient to study the correlation between a given predictor and the outcome or the causal impact of a given predictor on the outcome, while at the same time taking into account the correlation and impact of other competing or rival factors in relation to our outcome or dependent variable.

Consider the example where we investigate the correlation between a politician’s perceived competence and their performance in elections again. Say now we want to include one more predictor or independent variable in the model, and this variable is the economic situation of the politician’s constituency. Once we include both predictors, then the estimated slope of the competence variable should be understood as the correlation between competence and electoral performance while at the same time we also consider for the influence of the economic situation in this politician’s constituency on the election outcomes in our regression analysis. It is possible that once we account for a constituency’s economic situation, a politician’s perceived competence will have little to do with their performance in the election.

A similar concept you may have heard of is intersectionality, as it suggests that an individual’s race and gender can have a non-trivial impact on various outcomes in their life.

Page 15

When we know it is possible to include more than one predictor in the linear model, we will need to consider two questions more carefully: Which variables we should include and how these variables relate to each other in the model. Many quantitative researchers dedicate their entire career to studying the statistical techniques and tools of model and variable selection. Even though model comparison is indeed a very important topic to pursue, but this will not be our focus here, as many model comparison techniques often require more assumptions and subject a lot of debates.

When it comes to for variable selection, the key thing to keep in mind is to avoid omitting or forgetting to include any important predictors regarding the dependent variable of interest in the model, because the result of doing so is that the estimated coefficient of the slope of a predictor can become very small in terms of the size and will no longer be statistically significant once we account for other predictors. Quantitative researchers who focus on statistical learning and causal inference have also proposed some criteria for us to select variables in the regression model.

As we include more than one variable in the linear regression model, it is also very important we do not include two predictors highly correlated with each other because doing so will make your model contain redundant predictors and make our estimated coefficients misleading and unreliable. This is what we call collinearity, and we can use the variance inflation factor as a possible benchmark to detect the presence of collinearity in the model, or we can just check the pairwise correlation of all predictors. If two predictors are highly collinear or correlated with each other, the usual practice is to drop one of them.

It is also important to highlight that sometimes having more predictors may artificially increase the value of the R-squared so that R-squared no longer serves as an informative metric of model fit. Adjusted R-squared is useful to address this situation as it includes one additional term in the formula to penalize the inclusion of too many variables in your model. In other words, including more predictors may in the end shrink the value of the R-squared or the marginal utility of adding one predictor will decline as you include more predictors.

The inclusion of more than one predictor in the model can also allow us to model more complicated statistical relationships, some of which can be non-linear. More specifically, we can try to model non-linear data generation process of our outcome variables by including the interaction term and the polynomial term in our model.

A linear function with an interaction term will allow some dependency between different predictors so the estimated slope of a predictor can change under different situations. For instance, maybe the correlation between political competence and election results will change when we consider the economic structure of a constituency, which can be either agriculture or industrial. There is also a good example in the textbook for you to consider.

A linear function with polynomial terms basically allows the predictor to have a non-linear relationship with the dependent variable, and you will see an example in the next slide.

Page 16

Here is a quick example of a linear function with quadradic terms. In this case, we set the intercept at one and the slopes of X and X-squared at 2 and 0.5, respectively.

Here is the interesting question to ask yourself: If multiple linear regression allows us to include more complicated interactive dynamics between different predictors such that the statistical relationship between predictors and the dependent variable is no longer linear, then why do we need generalized linear model? We can talk about this briefly when we see each other in class, but we will address this issue in more depth again in the following week. The hint is: Generalized linear model allows us to relax more assumptions of OLS in addition to the linearity one.

Page 17

Now this concludes our “practical” review of multiple linear regression.

Multiple linear regression is a convenient tool, but by now I hope it should be clear to you there is not a guarantee that your statistical analysis or your quant research will be perfect or 100% scientific. Regression analysis is not completely a black and white matter.

When you try to carry out your own analysis, there will be many decisions you have to responsibly make and the decisions you make may not have a clear statistical reason underneath. Regardless of the model we use, we are bound to take many assumptions, and some of the assumptions can be controversial and unrealistic. And eventually, doing quant social research will demand a solid knowledge of the subject matter, which often can only be obtained with high-quality qualitative research. You will need to come up with a good substantive reason for your model setup, including the predictors you wish to include and how these predictors will be combined to inform a reasonable data generation process of the dependent variable.

Here are some principles or steps you can consider, and I will not read them out one by one. The key lesson is: There is no perfect model. Nonetheless, knowing that by no means invalidates the analytical value or need of using statistical models to understand the social world. Meanwhile, be prepared that any solid quant research will require you to use multiple statistical techniques based on a good understanding of both quantitative and qualitative social research. A table with one regression model including all predictors will not get you very far.

I would like to end this review by quoting George Box again. He is one of the most famous statisticians of all time. All models are wrong, but some of them are useful. Statistical models are approximations of the real life.

This is Dr Chao-Yo Cheng. See you in class.