**Week 3**

**Pages 1-2**

I am Dr Chao-Yo Cheng. Welcome back.

In this video, we will discuss some fundamentals of logit regression, one of the most widely use regression techniques among quant social researchers when we analyze a dependent variable that can only take one of the two values, usually 0 or 1.

In Part 1, I will review some of the key fundamentals of OLS to explain the reasons for many quant researchers to believe that we should use a different regression model to study a response variable that can only take one of the two values. Quant researchers call this type of variables binary or binomial.

In Part 2, I will cover the key concepts and rationale behind logit or logistic regression. Our first task is to understand why logit regression allows us to model a binary or binomial dependent variable. We will also pay attention to some important practical issues when we apply the logit regression to infer the statistical relationship between our predictors and binary outcome variable. At the same time, I will also try to use logit regression to explain the rationale behind other generalized linear models. We will not cover all generalized linear models in advanced quants, but I hope today’s video will give you some basic ideas about these models.

**Page 3**

You should be relatively familiar with the key fundamentals of the least squares linear regression OLS by now. Depending on the number of predictors, you can either specify a bivariate linear regression model or a multiple linear regression model.

In practice, we usually use multiple linear regression model in research because it is very unlikely to use a single predictor or independent variable to produce any insightful understandings with respect to our outcome variable. What is more important is if we exclude some important variables in our model, we may get biased results (in terms of the estimated intercept and slopes).

Keep in mind – a relatively good model specification requires a solid understanding of the topic or the subject matter. It is also very important that we avoid some common pitfalls. For one thing, we should avoid including two predictors that are collinear with each other. We should also avoid depending on R-squared as a metric of model fit as the generic R-squared may be artificially inflated as we include more than one predictor. When we use multiple linear regression, we can instead use adjusted R-squared to evaluate model fit (although it is very important to know we should still take it with a grain of salt) and use ANOVA to compare the performance across different model specifications.

Finally, when we try to use different combinations of predictors to study our outcome variable and specify the linear regression models, it is also very common to call these models as the data generating process, or DGP, of the outcome variable.

**Page 4**

With certain assumptions, OLS is a very powerful regression technique because it can produce unbiased estimates for the intercept and slopes. The fitted regression line based on the estimated intercept and slopes can also produce the smallest sum of square errors, which means the least squares least squares can provide us with the unbiased results with smallest errors or residuals.

OLS can also model or describe more complicated data generating process such that the statistical relationship between the outcome and explanatory variables may not be a simple straight line. We can do this when we include some unique terms in our model specification, such as interaction or quadradic/polynomial terms.

**Page 5**

Sometimes you may have heard that we need to use generalized linear models for study non-linear statistical relationship, which can either be correlation or causality depending on the assumptions and data structure, but it is only part of the story. OLS, or the least squares linear regression, can model non-linear statistical relationship.

To demonstrate when OLS may not be our ideal modeling choice, we can start by considering a scenario in which we are trying to study a binary or binomial response or dependent variable.

A binary variable is a variable that can take only the value of 0 or 1 to represent the yes/no or the success/fail situation. Voter turnout is a good example because a voter can either vote in the elections or stay home and does not vote in the election. In this sense, a binary variable is a nominal or categorial variable.

A binomial variable, in contrast, is a variable calculating the number of yeses or successes in several identical, independent trials with a constant probability of having success or yes. We will talk more about probability in a moment, but usually a probability will be between 0 and 1. Say if based on our data in a constituency five out of ten eligible voters went to cast their vote, conventionally we can say 50% of voters turned out in the elections, but using the jargon of probability we will say the probability of a voter turning out is 0.5, using the data I just described. As a result, a binomial variable is a continuous variable bounded between 0 and 1.

Conceptually binary and binomial variables are not entirely the same, and to understand the binomial variable will require some basic understanding of the probability theory. However, from the data analytical point of view, binary and binomial variables are not that much different. Let’s use voter turnout as the example to illustrate this point. While conceptually we can define voter turnout as a binary or binomial variable, in practice of data analysis, we will have to create a binary variable in our dataset and use 1 and 0 to indicate voters who voted and did not vote respectively. And use the binary variable of voter turnout as our response variable in the regression. As such, I will use binary variable for the following discussion.

Before I continue, I must also highlight two caveats. First, treating voter turnout as a binomial requires us to assume several situations that may not be plausible in real life. First, we will have to assume each voter’s decision to turn out or not is independent of each other. In other words, my decision to vote in one election has nothing to do with your decision to vote. And this assumption is clearly not realistic, as our voting behavior, including our vote choices can be easily influenced by our relatives and friends. Next, even if each voter’s decision to turn out is independent of each other, we will need to assume everyone turns out with the same probability, which again is not very realistic. We will return to these assumptions at the end.

Second, I was using the example that 5 out of 10 voters turned out to infer the probability of a voter’s turning out is 0.5. This is actually a very specific way to understanding and describing probability. Some quant researchers may not agree with this. We will not cover much about this debate, but the only thing you should know is we may need to use another statistical paradigm if we formulate probability in another way.

**Page 6-8**

We can demonstrate why the linear regression will become problematic when we have a binary variable that can only take the value of 0 and 1. Here I will provide two most intuitive reasons. And in the following two slides I will provide visualized illustration.

The first reason why OLS can fail is because a linear function of a binary response variable can produce unrealistic predicted outcome. In other words, if we use OLS to study a binary outcome, the fitted line will include values below 0 and 1 along with the vertical axis.

The second reason is we can no longer satisfy the assumptions we need for the least squares linear regression when the outcome variable is either 0 or 1. That is, the variance of Y will no longer be the same across all levels/values of X.

Consider this scatterplot with two regression lines. The straight line is the fitted line produced by the least squares linear regression and the dash line is the fitted line provided by logistic or logit regression. A quick look can tell us already logit regression provides a better fit to the data when for each observation the dependent variable can only be 0 or 1.

But if we take a closer look at the straight line, you can see that OLS can generate predicted outcome or predicted Y above 1 and below 0, which in the case makes no sense at all.

Likewise, you can also see that the variance of Y is not the same. The variance of Y is quite small when X is close to -5 or 10, and will get bigger when we move close to the midpoint of X. In other words, now we can use the value of X to predict the residuals, which is a clear violation of the homoscedasticity assumption for the least squares linear regression.

**Page 10**

Now we know OLS may not be the best regression estimator we should use. In this part, I will use four steps to explain the basic setup of logit regression and how to interpret the estimated coefficients or estimated slopes.

In the first step, we will need to change the way we think about binary outcome variables. More specifically, we will need to consider the binary outcomes through the lens of probability. Next, I will explain how to transform probability into log odds before introducing logit regression. As a spoiler alert, logit regression allows us to study binary outcome variables by transforming it into log odds and consider log odds as a linear function of our predictors. As a result of this transformation process, the interpretation of the estimated slopes can be very challenging, so I will explain the process we take to read the estimated slopes very carefully and slowly in the final step.

**Page 11**

Let’s start by treating the binary variable, which can take the value of either 0 or 1, as a probability. We can use a lower p, more specifically, to present the probability of our outcomes takes the value of 1. As you can see at the bottom of this slide, we can use capital P with parenthesis and put Y equals 1 in the parenthesis to represent the probability of Y equals 1.

The main objective of using logit regression to study our binary outcome variable is to model the probability of Y equals one, or the probability of the event of our interest takes place because we use 1 to represent the occurrence of the given event.

According to the probability axioms, keep in mind any probability can only be between 0 and 1. And the sum of the probability of one event happening and the probability of this event not happening should be 1. These are the mathematical properties you should keep in mind, as we will need them in a minute.

**Page 12-14**

Now we will need to convert probability into log odds. Log odds, by definition, is the log of odds. If the probability of one event happening is lower p, then the odds of this event is defined as p divided by 1 minus p. The idea here is to get a sense the chance of one event taking place relative to the opposite scenario, which is this event does not take place.

Let’s use one example to reinforce our understanding.

Say today the probability of raining is 0.8, which is perhaps not unusual for London. Then we know the probability of not raining is 0.2, using the axioms we just discussed in the previous slide.

The odds of raining today will then be 0.8 divided by 0.2, which equals 4. This number tells us it is four times more likely to be rainy today than not rainy.

We can obtain the log odds, of the log of odds, by taking the natural log of this odds, which is 4. Natural log is a log transformation with the base of e. Lower e here refers to a mathematical constant commonly used for log operations in algebra.

You may be wondering why we must take so many steps to transform the dependent variable from a binary variable to log odds. You will see in a minute, but if we can plot the relationship between a probability between 0 and 1 and the corresponding log odds, you can see that transforming probability into log odds basically helps us get around one of the issues that make OLS fail for binary outcomes: while a binary outcome variable or a probability can only be 0 or 1, which makes any predictions below 0 and above 1 unacceptable, now the log odds is no longer constrained between 0 and 1 (if you take a look at the vertical axis).

**Page 15**

This is the magic of logit regression. Intuitively speaking, logit regression allows us to use a linear function to model log odds. In other words, log odds can be represented as a linear function of our predictors.

The process or the function we use to transform our original binary variable is called the logit link function, or the logit function. Logit regression obtains its name because we will have to use the logit function to transform the original binary dependent variable so it can be modeled or represented with a linear function.

In addition to logit regression, there are many generalized linear models out there, but each of these generalized linear models shares one thing: They use different link functions to transform the outcome variable so the outcome variable can be modeled as a linear function of our predictors. Using the same logic, a generalized linear model where the corresponding link function keeps the original outcome variable is just OLS. This tells you that OLS is actually one of the generalized linear models.

**Page 16**

Although many researchers have been using the type of dependent variable to inform their modeling choice, which leads to a very popular statement that “one must use logit regression when the dependent variable is 0 or 1,” it is very crucial we are transparent about the assumptions we need for logit regression.

The textbook has used different examples to discuss the assumptions of logit regression, and here I will just highlight three most important ones. The reason why I want to highlight these ones is because these assumptions are no less unrealistic or controversial compared with the assumptions that we need to hold for the least squares linear regression.

In particular, two assumptions here can be contested. For one thing, as we have discussed using voter turnout as the example, it is very hard to imagine the occurrence of a particular event for all observations in the dataset is independent of each other. Also, it may be a bit strong to assume the log odds must be a linear function of the predictors.

**Page 17**

Although logit regression provides us with a convenient modeling technique to study binary dependent variables, it can be very tricky to interpret the estimated coefficients, especially when it comes to the estimated slopes. I will do it step by step and use one example to illustrate these steps in a minute. We are also going use the problem set to reinforce the procedures we need to take to interpret the estimated slopes produced by logit regression.

Consider this simple bivariate logit regression again. The lower p represents, as a reminder, the probability that Y or the dependent variable equals 1. The logit link function allows us to transform this probability into a log of odds of Y or the dependent variable being 1.

Using the same logic we use to interpret the estimated slope for a least squares linear regression, which tells us the slope corresponds to the changes on the left hand side when we move the predictor by 1, say from 0 to 1 or from 1 to 2. Our initial starting point with respect to X does not matter here as long as we move X by one unit.

Therefore, we can see that our beta here corresponds to beta changes in log odds when we move X by one. Here we use X equals 0 and 1 for the purpose of illustration. In other words, beta is the difference between two log odds, one log odds is for X equals 1 and the other log odds is for X equals 0. The difference between two different log odds is the same as the log of one odds divided by the other. This operation is possible based on one of basic rules for logarithm.

Finally, using the basic rules of logarithm again, we can take the exponent of beta to derive the fraction in which one odds was divided by the other. The right hand side of the last equation is odds ratio. Our interpretation will be based on the odds ratio, or more simply, the exponent of the estimated slope. We can complete the calculation very easily in R.

**Page 18-19**

Let’s use an example to reinforce our understanding.

Say we have collected data from a political party in a country for the past 100 elections, and we want to understand the correlation between the number of hours a party spends on televised campaign and the election results, which can either be the party winning or losing the election. In other words, we want to know if the number of hours for televised commercials is correlated with the probability that a party wins the election.

The predictor is the number of hours on TV campaigns, and the outcome variable is a binary variable that takes the value of 1 if the party wins. Suppose we derive the estimated slope as 0.33. And based on the discussion in the previous slide, we know the odds-ratio is the exponent of 0.33, which is about 1.39.

What does this result tell us? Should we recommend the party spend more money on increasing the hours of televised campaigns?

Well, keep in mind the odds ratio is the ratio of two different odds. The odds in the numerator can be considered as the odds of winning when the party spends one hour on TV campaigns. And the odds in the denominator can be considered as the odds of winning when the party spends zero hour on TV campaigns. Therefore, if the odds ratio is larger than one, it means the numerator is larger than the denominator, which means the odds of winning goes up when the party spends on additional hour of televised campaigns.

In conclusion, the party definitely should spend more hours promoting their candidates on TV.

The table here lists all three possible scenarios when we try to read odds ratio. If the odds ratio is 1, we will know increasing the hour of campaigning by one does not change the odds of winning. If the odds ratio is smaller than one, then we will know increasing the hour of campaigning by one will actually reduce the odds of winning the elections.

**Page 20**

We have covered the basic setup of logit regression and some key principles of statistical inference for logit regression. Before I conclude this lecture, let me wrap up some loose ends.

First, we will work on the problem set to demonstrate how to conduct significance tests on odds ratios, but the same statistical tools we use for linear regression, such as the p-value or confidence intervals, are applicable. We can also use similar tools to carry out model comparison. It might be tricky to evaluate model fit for logit and other non-linear generalized linear models, and we will discuss this in detail in class.

Next, odds ratios are for sure a very obscure and challenging statistical concept. We will discuss how to use predicted probabilities as the alternative to report your results from logit regression.

Finally, I would like to challenge the conventional wisdom that we can only use logit or probit to study binary outcome variables. Quant scholars in some subject areas are very strict about using logit or probit regression when the outcome variable can only be 0 or 1, but it is important to recognize now every quant researcher shares this view. For one thing, it has been very clear that logit regression and other non-linear generalized models can also have unrealistic assumptions. Also, it is possible to use linear regression to study binary outcomes as we have come up with some tools to address the pitfalls we highlighted, especially the violation of the homoscedasticity assumption. When we use OLS to model binary outcomes, we will call it a linear probability model.

A recent review article also highlighted some other caveats when we use logit and other non-linear generalized models. The review article is quite technical, but one takeaway lesson is GLMs will create more biased results compared with OLS when the model excludes some important predictors, which we can never know for sure. And the setup of non-linear models may make the estimated coefficients across different models using the same data not comparable.

**Pages 22-24**

Here is the list some of the very common GLMs in practical or applied quant social research. As I have mentioned, these GLMs use different link functions to transform different types of dependent variables so they can be models by a linear function.

If you would like to see how you can use R to use these models, the examples provided by UCLA Statistical Methods and Data Analytics are very thorough and easy to follow.

If you would like to learn more about the underlying mathematical reasoning, please consult the book written by Professor Andrew Gelman and his coauthors.

In practice, which model we should use depends on the type of dependent variables, but as I have mentioned before we should also be careful with the relevant assumptions as no model is assumption free.

Finally, please allow me to use George Box to conclude our discussion. These two sections are retrieved from his 1976 article. These two sections are very relevant to the points I repeated make throughout this lecture – all models are wrong, but some are useful. The key is to know the pros and cons and become a well-informed quantitative social researchers.