Applied Quantitative Analysis II
Lab 2

Eric G. Zhou

Spring 2020

NYU Wagner School of Public Service
NYU Grossman School of Medicine

AQA II: Lab 2

Reminder for the office hour

12:15 - 1:15 pm, Fridays, Room 5154 C, Puck Building

Questions about the lecture, coding, your **project**

If you want to discuss anything in detail, better send me an email about the question ahead, so our meeting can be constructive

AQA II: Lab 2

A few changes from last time

- We will take a 10-minute break during 7:25 - 7:35
- I will ask more questions, and we should do more interaction
- Now, you guys could nominate topics that I cover more in detail(theory, programming or both). I will choose what to cover if no feedback from you.

Today's ontents

- Quiz
- More on the logit/probit regression
- Interpretation of logit/probit models
- Programming in Stata with GSS data

A gentle reminder that, when writing papers or assignment, please cite any stats or math equations from your textbooks, my slides may have mistakes.

Quiz

1.What does randomization achieve in studying the treatment effects between the treatment and control group?

2.What happens to the analysis above if your participants randomly drop out of the study before the end of treatment? What if the attrition is not random?

3.What happens when you include two highly correlated regressors in one OLS regression model?

Quiz answers

1.What does randomization of treatment achieve in studying the treatment effects between the treatment and control group?

The randomization of treatment will eliminate the systematic differences between the two groups. Or, the treatment variable thus becomes exogenous to the outcome, which allows us to estimate the treatment effect unbiasedly.

2.What happens to the analysis above if your participants randomly drop out of the study before the end of treatment? What if the attrition is not random?

If it's random, then the treatment effect is still unbiased, but due to the reduction in the number of observation we lose some statistical power, which is actually very important for the design purposes of RCT. One implication of this consequence is your estimates will be biased towards the null if large number of participants attrit. If it's not random, then your estimates are biased. Think about what we could do?

Quiz answers

3.What happens when you include two highly correlated regressors in one OLS regression model?

This leads to another question, what's the data generating process? If there are no issues of endogeniety, estimators are unbiased but standard errors are exploding. The problem is we never know the DGP, otherwise we don't need regressions. So, it's safe to say never include two highly correlated regressors in your regression model. At least it's safe in the econometric/statistic sense.

People do include such variables due to theoretical reasons. Know your goal and audience is always very important.

Odds ratio, log odds, probability and some odd names

This part is basic and important

Odds is defined as $\frac{p}{1-p}$

Log odds(or logged odds) is $ln(\frac{p}{1-p})$

Now consider two probabilities p_1 & p_2, Odds ratio is defined as $\frac{\frac{p_1}{1-p_1}}{\frac{p_2}{1-p_2}}$ (Note there are some disagreement on this definition)

What the Logistic model(in stata) does is to fit the data with $\frac{p}{1-p} = \exp(x_j\beta + \beta_0)$ and obtain β using Maximum Likelihood Estimations. This is something we have to find out before jumping to interpretations, i.e., the statistical packages vary, unless you hand-code the estimator.

Brief recap on Logit/Probit theory

The identification condition

$$P = p(x) = E[y|x_1, x_2, ..., x_k]$$

therefore, we fit the data with a logistic regression model in the form of

$$\ln(\frac{p}{1-p}) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k$$

importantly, remember $ln(x) = y \iff x = e^y$

And conversely, we could have

$$p = \frac{e^{x'\beta}}{1 + e^{x'\beta}}$$

Brief recap on Logit/Probit theory

We should ask ourselves two questions:

- How did people come up with this estimation method?
- What happens to the error term ϵ_i?

Brief recap on Logit/Probit theory

In social science, the intuition was from several famous economists, McFadden, Train, Pakes. Specifically, it starts with the concept of individual choice and utility function.

Luce (1959) first derived famous Independence from Irrelevant Alternative(IIA) assumption and choice probability. Marschak (1960) implied that the axioms are consistent with utility maximization. They strive to model individual choices, behavior and uncertainties.

Consider the utility function for individual i with J alternatives U_{ij}, with the observable part V_{ij} and the unobservable error term e_{ij}

$$U_{ij} = \underbrace{V_{ij}}_{\text{observable}} + e_{ij}, \forall j \tag{1}$$

Brief recap on Logit/Probit theory

Now, let's consider the utility of the same individual i but for another alternative m, and similarly,

$$U_{im} = \underbrace{V_{im}}_{\text{observable}} + e_{im}, \forall m \tag{2}$$

Researchers want to know what is the probability of individual i choose alternative j over m, given a series of observables featured in V_i?

What would you do?

Brief recap on Logit/Probit theory

Researchers want to know what is the probability of individual i choose alternative j over m, given a series of observables featured in V_i?

While it's intuitive to think given the individual is rational, i.e. utility maximizing, he would only choose j if $U_{ij} \geq U_{im}$, or vice versa. This is not enough. So, McFadden (1974) developed the following framework.

$$P_{ij} = Prob(U_i j \geq U_i m) = Prob(V_{ij} + \epsilon_{ij} \geq V_{im} + \epsilon_{im}), \forall j \neq m \tag{3}$$

$$P_{ij} = Prob(\epsilon_{ij} - \epsilon_{im} \geq V_{im} - V_{ij}), \forall j \neq m$$

Brief recap on Logit/Probit theory

Then, we assume the error term ϵ_i follows the type-I extreme distribution, and independent from each other. And the density function for this term is

$$f(\epsilon_i) = e^{-\epsilon_i} e^{-e^{-\epsilon_i}} \tag{4}$$

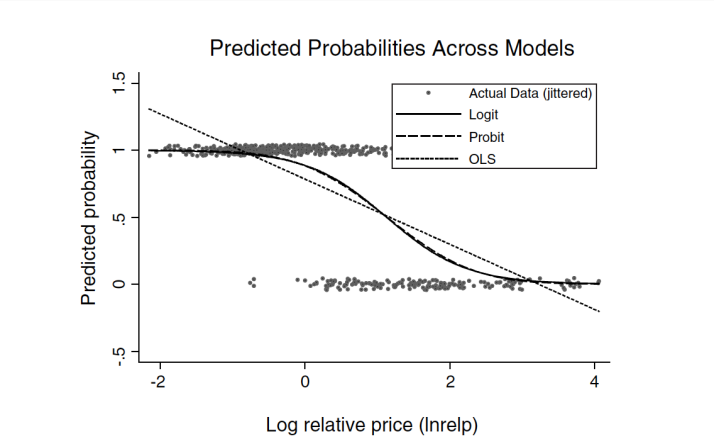
Why? Because it has nice properties, such as

$$F(\epsilon^*_{ijm} = \epsilon_{ij} - \epsilon_{im}) = \frac{e^{\epsilon^*_{ijm}}}{1 + e^{\epsilon^*_{ijm}}} \tag{5}$$

Evidently, equation (5) is identical to the logit distribution. We are making implicit assumptions about the function forms of the error term in the utility function. Identification is usually hard, and always require instrumental variables. Read BLP(1995) if you are interested.

Review of the Logit/Probit regression

Below is a graph of predicted probabilities over a regressor with the same data and specifications, showing the differences among Logit, Probit and OLS.



Interpretation of the Logit/Probit regression

Remember we fit the logit regression as

$$\ln(\frac{p}{1-p}) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k$$

To better illustrate, let's consider a simpler model, and x_1 is binary.

$$\ln(\frac{p}{1-p}) = \beta_0 + \beta_1 x_1$$

So, $Odds(p(x_1 = 1)) = \frac{p(x_1=1)}{1-p(x_1=1)} = e^{\beta_0+\beta_1}$ and $Odds(p(x = 0)) = e^{\beta_0}$

By now, you should know how to interpret β_1. Anyone wants to give a shot?

Interpretation of the Logit/Probit regression

By now, you should know how to interpret β_1. Anyone wants to give a shot?

$$OR(x_1) = \frac{Odds(p(x_1 = 1))}{Odds(p(x = 0))} = e^{\beta_1}$$

$$\beta_1 = ln(OR(x_1))$$

In this case, $beta_1$ is called a logit coefficient. If you use the **logistic** command in Stata, it gives you Odds Ratio automatically.

In written formality, we should interpret the logit/logistic regression coefficient as such:

The model predicts that every unit increase in x is associated with a(n) increase/decrease in the odds that y = 1 by a factor of [amount of the coef.]

Interpretation of the Logit/Probit regression

We do not have a R-square as OLS for a logit model. But people have come up with ways of quantifying the model fit. See more at

`https://stats.idre.ucla.edu/other/mult-pkg/faq/general/faq-what-are-pseudo-r-squareds/`

Remember don't interpret the Pseudo-R-Square of Logit the same way as that of the OLS model.

Interpretation of the Logit/Probit regression

Probit is the same as Logit, except that due to different assumptions, the objective function does not have a closed form, so the coef. you obtain is numerically optimized. I encourage you to look at it yourself with resources in textbooks and online. Let me know if you have any questions!

Any other questions before we move to stata?