Applied Quantitative Analysis II
Lab 1

Eric G. Zhou

Spring 2020

NYU Wagner School of Public Service
NYU Grossman School of Medicine

Let's find us an office hour

What about 12:15 - 1:15 pm Fridays? Plus I will stay around for another 30 minutes after our lab

Email: geng.zhou@nyu.edu

Christina Nelson, Tuesdays, 10am-12pm, Puck room 4186, also by appointment

Email: cn1097@nyu.edu

Climate Survey

- Can we please go around room introduce your name, your area of interests(in general)? And thesis topic(or pending topics)?
- How many of you are in the AQR program?
- How many of you intend to use Stata, R, Python as a programming language?
- What do you plan to do after graduation? If academic, what programs? If industry, what positions?

These questions will help me to determine what topics to cover more in detail, and what resources may be more interesting. But send me an email about anything anyways.

AQA II: Lab 1

Expectations

- The main goal for the lab section is to practice what we learn in the lecture with practical STATA programming exercises, therefore, to be the master of these techniques, and apply them appropriately in your master thesis as your progress through the semester
- The participation grade of this class also expands to lab, so we encourage you to ask questions, share feedback or comments, in class or send me an email
- We will do weekly lab assignments, exercises, and quizzes(not graded but essential to learning)
- Assignments can be written in Stata or R, but we will talk about solutions only in Stata; your code should run through smoothly and be properly documented

AQA II: Lab 1

Contents

- Review of the OLS regression
- Review of the Logit/Probit regression
- How to download and use the GSS data?
- Programming in Stata with GSS data

A gentle reminder that please cite any stats or math equations from your textbook, my slides may have mistakes.

The ordinary Least Square regression

Assumptions of the model $y_i = x_{i1}\beta_1 + x_{i2}\beta_2 + \cdots + x_{ik}\beta_k + \epsilon_i$

- Linearity: the model specifies a linear relationship between y and Xs.
- Full rank: There is no exact linear relationship among any of the independent variables in the model.
- Exogeneity of the independent variables: $E[\epsilon_i | x_{j1}, x_{j2}, \cdots, x_{jk}] = 0$.
- Homoscedasticity and nonautocorrelation: Each disturbance, ϵ_i has the same finite variance, σ, and is uncorrelated with every other ϵ_i.
- Data generation: the data in X_{jk} may be any mixture of constants and random variables.
- Normal distribution: ϵ_i is normally distributed.

How to derive the OLS estimator?

Assume that all of the six assumptions are met.

$$Y = X'\beta + \epsilon \tag{1}$$

The least squares coefficient vector minimizes the sum of squared residuals, so we choose beta such that the coefficient vector minimizes the SSR.

$$\min_{\beta} S(\beta) = e'e = (y - X\beta)'(y - X\beta) \tag{2}$$

Now switching terms and it's easy to see

$$S(\beta) = y'y - 2y'X\beta + \beta'X'X\beta \tag{3}$$

How to derive the OLS estimator?

Does this ring a bell now? It's quadratic, so the necessary condition for a minimum is

$$\frac{\partial S(\beta)}{\partial \beta} = -2X'y + 2X'X\beta = 0 \tag{4}$$

And this is the famous equation all of you have known

$$\beta_{ols} = (X'X)^{-1}X'y \tag{5}$$

But can we tell if it is unbiased? Yes, but in reality it depends on the assumptions.

$$\beta_{ols} = (X'X)^{-1}X'(X\beta + \epsilon) = \underbrace{\beta}_{\text{true value}} + \underbrace{(X'X)^{-1}X'\epsilon}_{\text{bias} = 0 \text{ if exogeneity holds}} \tag{6}$$

What does a random assignment do in theory?

The word random

Watch out for this word anywhere! One of the most important word in methodology.

In a Randomized Control Trial(RCT), the treatment variable is randomly assigned to the treatment and control group.

What does this achieve?

In a survey that has the random sampling procedure within a certain population, the sample (assuming large enough) would allow us to make statistical inferences about the underlying population.

What does this mean?

Can you think of another important "random" item?

Review of the Logit/Probit regression

Let's begin with a binary dependent outcome y that takes either 1 or 0, and p denotes the probability event $y = 1$ occurs, and $1 - p$ for $y = 0$.

Since we want to model the probabilities, we parametrize the probabilities p to depend on a regressor vector x and a parameter vector β, so the single-index form model we would use with conditional probability can be written as

$$p_i \equiv \mathsf{Pr}[y_i = 1|x] = F(x'\beta) \tag{7}$$

Consequently, the Logit model specifies $p = \Lambda(x'\beta) = \frac{e^{x'\beta}}{1+e^{x'\beta}}$

Whereas, the Probit model specifies $p = \Phi(x'\beta) = \int_{-\infty}^{x'\beta} \phi(z)dz$, where $\Phi(\cdot)$ is the standard normal cdf.

Review of the Logit/Probit regression

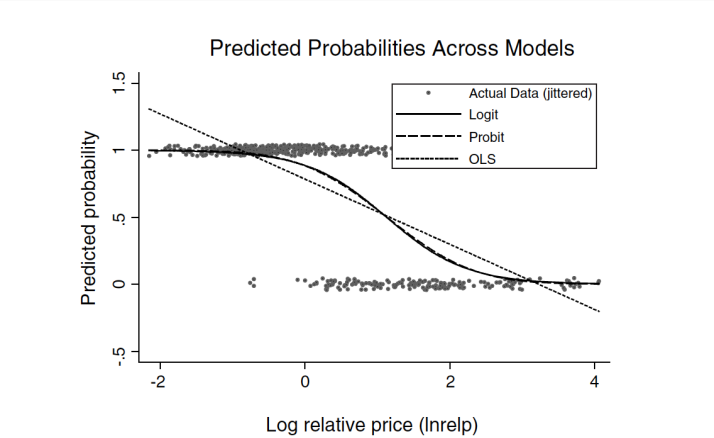
The figure below from Cameron & Trivedi shows the comparison of several popular linear probability models

Table 14.3. *Binary Outcome Data: Commonly Used Models*

Model	**Probability** ($p = \Pr[y = 1|\mathbf{x}]$)	**Marginal Effect** ($\partial \mathrm{p}/\partial \mathrm{x_j}$)
Logit	$\Lambda(\mathbf{x}'\boldsymbol{\beta}) = \dfrac{e^{\mathbf{x}'\boldsymbol{\beta}}}{1+e^{\mathbf{x}'\boldsymbol{\beta}}}$	$\Lambda(\mathbf{x}'\boldsymbol{\beta})[1-\Lambda(\mathbf{x}'\boldsymbol{\beta})]\beta_j$
Probit	$\Phi(\mathbf{x}'\boldsymbol{\beta}) = \int_{-\infty}^{\mathbf{x}'\boldsymbol{\beta}} \phi(z)dz$	$\phi(\mathbf{x}'\boldsymbol{\beta})\beta_j$
Complementary log-log	$C(\mathbf{x}'\boldsymbol{\beta}) = 1 - \exp(-\exp(\mathbf{x}'\boldsymbol{\beta}))$	$\exp(-\exp(\mathbf{x}'\boldsymbol{\beta}))\exp(\mathbf{x}'\boldsymbol{\beta})\beta_j$
Linear probability	$\mathbf{x}'\boldsymbol{\beta}$	β_j

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.



Odds ratio, log odds, probability and some odd names

This part is even more important because this is directly related to whether we could correctly interpret the model. More on this next week.

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.

AQA II: Lab 1

Download and use GSS data

The General Social Survey is a very powerful dataset manage by the NORC at the University of Chicago.

Let's go to their website at `https://gss.norc.org/`

Stata programming

Some general tips for coding/programming:

Write a codebook that stores solutions to difficult problems, useful functions, graphs, or even for self-growth

Be sure to understand the math/stats first then perform the analysis, the more you know the better estimations and interpretations you could have

Online resources, github, stackoverflow, stata journal, mannuals, R bloggers, etc.

Collaborate with peers, debug and appreciate each others' codes, work on project together