Applied Quantitative Analysis II
Lab 5

Eric G. Zhou

Spring 2020

NYU Wagner School of Public Service
NYU Grossman School of Medicine

AQA II: Lab 5

General reminder

The project should be taking shape

Assignment 3 will be available tonight on NYU Class and due in two weeks

Prepare for the potential outbreak of COV-19 Virus in US (CDC has warned the public)

AQA II: Lab 5

Contents today

Multinomial logit

Ordered logit

Interpretations m/ologit

Stata programming

Multinomial logistic regression

Remember the random utility model set up

$$U_{ij} = z'_{ij} + \epsilon_{ij} \tag{1}$$

We assume if individual i choose product j over k, then we must have

$$U_{ij} > U_{ik} \ \forall k \neq j \tag{2}$$

By imposing the structure on the error term, we could make the estimations work nicely

$$F(\epsilon_{ij}) = \exp(-\exp(-\epsilon_{ij})) \tag{3}$$

Multinomial logistic regression

The probability of individual i choosing product j out of the total J options is:

$$\text{Prob}(Y_i = j) = \frac{\exp(z'_{ij}\theta)}{\sum_{j=1}^{J} \exp(z'_{ij}\theta)} \tag{4}$$

This model is also known as the conditional logit model. $z_{ij} = [x_{ij}, w_i]$, where x_{ij} varies across the choices and individuals, but w_i contains the individual characteristics that do not vary across i. So sometimes, this model can be modified by partition

$$\text{Prob}(Y_i = j) = \frac{\exp(x'_{ij}\beta + w'_i\alpha)}{\sum_{j=1}^{J} \exp(x'_{ij}\beta + w'_i\alpha)} \tag{5}$$

Does this remind you of the fixed effects model?

Multinomial logistic regression

How do we find our coefficient? Define $\bar{x}_i = \sum_{j=1}^{J} P_{ij} x_{ij}$

$$\ln L = \sum_{i=1}^{n} \sum_{j=1}^{J} \ln \text{Prob}(Y_i = j) \tag{6}$$

Then we derive the Hessian and Jacobian matrix as

$$\frac{\partial \ln L}{\partial \beta} = \sum_{i=1}^{n} \sum_{j=1}^{J} (x_{ij} - \bar{x}_i) \tag{7}$$

$$\frac{\partial^2 \ln L}{\partial \beta \partial \beta'} = \sum_{i=1}^{n} \sum_{j=1}^{J} P_{ij} (x_{ij} - \bar{x}_i)(x_{ij} - \bar{x}_i)' \tag{8}$$

Multinomial logistic regression

How do we find our coefficient? β can be found via the Newton Method

Asymptotically, $\hat{\beta} \sim N(\beta, -H(\beta))$, where the variance term is the diagonal of the famous **Fisher information matrix**, i.e., the Hessian of β.

To ensure internal validity, we need two relatively strong assumptions:

1. The usual exogeneity assumption and the functional form for the error term being Gumbel for Logit and S-Normal for Probit.

2. The Independence from Irrelevant Alternatives (IIA). The odds ratios is independent from any irrelevant alternatives – convenient for estimation but not realistic in practice. E.g., the red bus, blue bus problem.

Multinomial logistic regression
Strength:

Powerful tool to study categorical outcomes, but rarely used now except for predictions or multi-stage models.

Limitations:

Require high computation power given large dataset

The degree of freedom is limited if you have many regressors or categories for your outcome

Rely on strong assumptions that are rarely plausible in reality (Simulation methods are helpful but very complex)

Ordered logistic regression

Many synonyms, ordered, ordinal logit, someone even calls this "ordered multinomial logit", same stuff.

It is essentially the latent variable regression but with multiple cutoff points (Num. of cutoffs = Num. of Ordered Class - 1)

Note that there is no constant value included in the regressors.

Recall the set up of the latent variable regressions:

$$y = 1 \text{ if } y* > 0$$

$$y = 0 \text{ if } y* \leq 0 \tag{9}$$

Ordered logistic regression

Imagine that CNN is conducting a political polling and ask how would you assess the current president?

The answers are:

1. Very good
2. Good
3. Neither good nor bad
4. Poor
5. Very poor

The answers are ordered. It's natural to apply ordered logit/probit model in this case.

Ordered logistic regression

You will observe people selecting any of the answers above and a bunch of observable regressors Xs. A description for this decision process is a running utility given the observables cutoff by the 5 categories. Change of the observables is associated with the change of outcome.

1. Very good, if $U_i > k_1$
2. Good, if $k_1 > U_i > k_2$
3. Neither good nor bad, if $k_2 > U_i > k_3$
4. Poor, if $k_3 > U_i > k_4$
5. Very poor, if $k_4 > U_i$

Where, $k = \beta' X + \epsilon$, and $k_1 > k_2 > k_3 > k_4$.

Ordered logistic regression

A visual example is as such

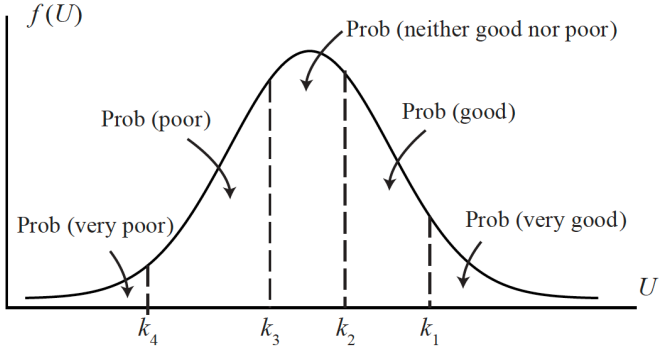

Figure 7.1: Distribution of opinion about president's job.

Ordered logistic regression

Like before, if we assume the error term has the CDF as $F(\epsilon) = exp(\epsilon)/(1 + exp(\epsilon))$.

The probability of the answer "Very good" based on our data is:

$$\text{Prob ("very poor")} = \text{Prob}(\epsilon < k_4 - \beta' x) = \frac{e^{k_4 - \beta' x}}{1 + e^{k_4 - \beta' x}}$$

Whereas,

$$\text{Prob ("poor")} = \text{Prob}(k_4 < U < k_3) = \frac{e^{k_3 - \beta' x}}{1 + e^{k_3 - \beta' x}} - \frac{e^{k_4 - \beta' x}}{1 + e^{k_4 - \beta' x}}$$

The other pairs of categories follow similarly.

Ordered logistic regression

Assumptions:

1. Usual assumptions on exogeniety and functional form of the error term
2. Parallel Line assumptions. Below is a paragraph I quote from UCLA stata website.

"One of the assumptions underlying ordered logistic (and ordered probit) regression is that the relationship between each pair of outcome groups is the same. In other words, ordered logistic regression assumes that the coefficients that describe the relationship between, say, the lowest versus all higher categories of the response variable are the same as those that describe the relationship between the next lowest category and all higher categories, etc. This is called the proportional odds assumption or the parallel regression assumption."