Applied Quantitative Analysis II
Lab 4

Eric G. Zhou

Spring 2020

NYU Wagner School of Public Service
NYU Grossman School of Medicine

AQA II: Lab 4

General reminder

Goal of this course: write a good thesis paper for the AQR program

By March 1st, you should all **secure** your datasets for analysis. Let Mike or I know if you have any issues

March - late April, we should be focusing on conducting the analysis, sensitivity analysis, diagnostics, visualizations – good work takes time

AQA II: Lab 4

Contents for today

Quick talk on assignment 2 Quiz Latent variable regression

AQA II: Lab 4

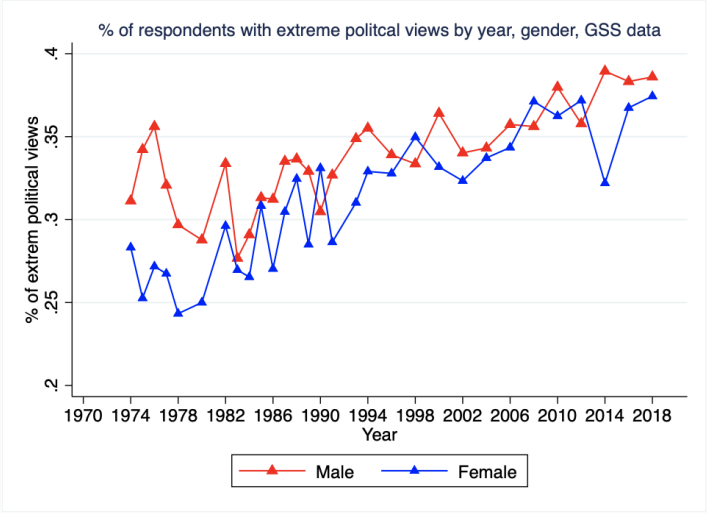
Assignment 2

Glad you all have turned in! It was long but should not be hard.

I thought I corrected the wording of Question 2.(1), but by the virtue of this mistake, the results still reveal something about the society...

The graph and story that I intended to have you guys make was like this

Assignment 2



Assignment 2

Source	SS	df	MS	Number of obs	=	55,068
				F(33, 55034)	=	15.29
Model	110.627295	33	3.35234227	Prob > F	=	0.0000
Residual	12063.0574	55,034	.219192815	R-squared	=	0.0091
				Adj R-squared	=	0.0085
Total	12173.6847	55,067	.221070417	Root MSE	=	.46818

extrem	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
sex						
Female	-.024569	.0040214	-6.11	0.000	-.032451	-.016687
age	-.0005875	.0006461	-0.91	0.363	-.0018539	.0006788
c.age#c.age	.0000182	6.41e-06	2.84	0.004	5.65e-06	.0000308
educ	.0083063	.0006797	12.22	0.000	.006974	.0096385

Assignment 2

Logistic regression — Number of obs = 55,068
LR chi2(33) = 501.89
Prob > chi2 = 0.0000
Log likelihood = -34668.151 — Pseudo R2 = 0.0072

extrem	Odds Ratio	Std. Err.	z	P>|z|	[95% Conf.	Interval]
sex						
Female	.8943986	.0163871	-6.09	0.000	.8628503	.9271003
age	.9976685	.0029386	-0.79	0.428	.9919255	1.003445
c.age#c.age	1.000079	.0000291	2.70	0.007	1.000022	1.000136
educ	1.038334	.0032291	12.10	0.000	1.032024	1.044682

Quiz

Quiz 1. Why random sampling of an underlying population ensures representativeness if conducted correctly? Give an example where random sampling is not feasible? What is your alternative strategy for this situation?

Quiz 2. Explain the advantages and disadvantages of using OLS or logit/probit to estimate probability

Quiz 3. What happens to your regression models when your key independent variable is measured with significant errors?

Quiz

Quiz 1. Why random sampling of an underlying population ensures representativeness if conducted correctly? Give an example where random sampling is not feasible? What is your alternative strategy for this situation?

Quiz

Quiz 1. Answer

In short, the random sampling within a certain population ensures that each unit has the equal chance to be included in the sample, and therefore your sample statistic can be inferred to describe the population statistics. We never could measure the population statistics directly. Why? It's costly to do so, if not entirely impossible. As in regression, $\hat{\beta}$ is merely an estimate we derived for the more important and interesting β of the population parameter. Think about your generalizability and external validity all the time – the discussion of this part should weigh half of your research (as in importance not length).

Random sampling is not feasible if you do not know the entire list of the population. What can we do? Multi-stage stratified/clustered random sampling(do not forget to adjust your standard error).

Quiz

Quiz 2. Explain the advantages and disadvantages of using OLS or logit/probit to estimate probability

Quiz

Quiz 2. Explain the advantages and disadvantages of using OLS or logit/probit to estimate probability

OLS is the best linear unbiased efficient estimator given the Markov-Gaussian Theorem, and it provides constant marginal effect. However, OLS could result in unrealistic estimates of probability at the two extremes.

Logit/probit models the probability within (0,1), however the marginal effect is not constant, which is fine if you understand the margin command in Stata. MLE estimators are highly biased if you did not specify the model correctly.

Logit/probit has worse properties when applied with panel data.

Quiz

Quiz 3. What happens to your regression models when your key independent variable is measured with significant errors?

Quiz

Quiz 3. What happens to your regression models when your key independent variable is measured with significant errors?

Two cases. If the error is random, your regression estimate is biased toward the null, assume the error is mean-zero. If not, then your estimate is biased toward unknown direction. Though many research do not address or comment on this issue in their paper, but you should, especially with survey data that involves "recall xxxx" – recall bias.

Solution? Go and collect better, and more data.

Latent variable regression

Discrete dependent-variable models are often cast in the form of **index function models**.

Consider the decision of making a large purchase, as the price increase when would you buy? What about others given the same cost?

Latent variable regression

Discrete dependent-variable models are often cast in the form of **index function models**.

Consider the decision of making a large purchase, as the price increase when would you buy? What about others given the same cost?

Let's consider an unobserved variable y^* that measures the difference between benefit and cost such that $y* = x'\beta + \epsilon$

And individuals only decide to purchase if net benefit is positive, which arrives in data as

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

$$y = 0 \text{ if } y* \leq 0 \tag{1}$$

Latent variable regression

In this set up, $x'\beta$ is the index function. ϵ is an innocent normalization. And naturally, the **latent regression** part is still $* = x'\beta + \epsilon$, as we do not directly observe except the eventual choice.

In a general form, let us view this model with an arbitrary finite threshold value a, then we model the probability with the condition

$$\begin{aligned} \text{Prob}(y = 1|x) &= \text{Prob}(y* > a|x) = \text{Prob}(x'\beta + \epsilon > a|x) \\ &= \text{Prob}(\epsilon < x'\beta - a|x) \\ &= F(x'\beta) \end{aligned}$$

This should look familiar to you – it's the logit model if you specify ϵ as the Logit cdf, and Probit if standard normal cdf.

Latent variable regression

Everything goes around comes around. Do not fear strange names, after AQA-II you should know most of them.

This is the simplest model set up and it turns out we have many other very important derivatives of this model, such as multinomial logit, nested logit, etc.

Multinomial logistic regression

Remember the random utility model set up

$$U_{ij} = z'_{ij} + \epsilon_{ij} \tag{2}$$

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

$$U_{ij} > U_{ik} \ \forall k \neq j \tag{3}$$

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

$$F(\epsilon_{ij}) = \exp(-\exp(-\epsilon_{ij})) \tag{4}$$

Multinomial logistic regression

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

$$\text{Prob}(Y_i = j) = \frac{\exp(z'_{ij}\theta)}{\sum_{j=1}^{J} \exp(z'_{ij}\theta)} \tag{5}$$

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

$$\text{Prob}(Y_i = j) = \frac{\exp(x'_{ij}\beta + w'_i\alpha)}{\sum_{j=1}^{J} \exp(x'_{ij}\beta + w'_i\alpha)} \tag{6}$$

Does this remind you of the fixed effects model?

Multinomial logistic regression

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

$$\ln L = \sum_{i=1}^{n} \sum_{j=1}^{J} \ln \text{Prob}(Y_i = j) \tag{7}$$

Then we derive the Hessian and Jacobian matrix as

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

$$\frac{\partial^2 \ln L}{\partial \beta \partial \beta'} = \sum_{i=1}^{n} \sum_{j=1}^{J} P_{ij} (x_{ij} - \bar{x}_i)(x_{ij} - \bar{x}_i)' \tag{9}$$

Multinomial logistic regression

We need numerical methods to find the optimum β, so this model was not estimable until the late 1990s.

Newton's methods

AQA II: Lab 4

Multinomial logistic regression

Questions?