Lab 3

Barum Park

Department of Sociology
New York University

Feb. 8, 2018

Before We Start ...

- Any questions regarding last class?
- From this Lab onwards, I'll try to focus more on STATA code than on the models

Logistic Regression

- I'll try to demonstrate one last formulation of the binary logistic regression model
- This representation is called the **latent variable formulation** of the logistic regression model
- It appears in many textbooks, especially in the derivation of the probit model
- It will be helpful to understand ordered logistic regression in an intuitive way

- Suppose you have a "latent" (i.e., unobserved) outcome y^* which is continuous
- We assume that this latent variable is generated by the following equation

$$y^* = \alpha + \beta x + \epsilon^*, \quad \epsilon^* \sim \text{Logistic}(0,1)$$

- The "observed" outcome, y, is binary (either zero or one).

- Lastly, we assume that the latent variable is connected to the observed response in the following way:

$$y = \begin{cases} 1, & \text{if } y^* > 0 \\ 0, & \text{otherwise} \end{cases}$$

- You can think of the value 0 as a "threshold" (as $y^* > 0$ returns a 1 for y and $y^* \leq 0$ returns a 0 for y)
- This threshold is also arbitrary (we say, "unidentified") because

$$\begin{aligned} y^* > 0 &\implies \alpha + \beta x + \epsilon^* > 0 \\ &\implies \beta x + \epsilon^* > -\alpha \end{aligned}$$

Hence, we could let $-\alpha$ be the "threshold" and say that the "latent" regression has no constant

- It turns out that this model is the same model as the logistic regression we have learned so far!
- The derivation of this result is a little bit technical ...
- So let me convince you that these are the same models by simulation ..

Simulation Code

```
clear all
set seed
set obs 50000

gen x = rnormal()
gen u = runiform()
gen epsilonstar = ln(u/(1-u))
gen ystar = .5 + .8*x + epsilonstar

gen y = 0
replace y = 1 if ystar > 0
logit y x
```

Results

Logistic regression

Number of obs	=	10000
LR chi2(1)	=	1349.11
Prob > chi2	=	0.0000
Pseudo R2	=	0.1006

Log likelihood = -6030.4804

y	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
x	.8306022	.024909	33.35	0.000	.7817814	.8794231
_cons	.486941	.0221607	21.97	0.000	.4435069	.5303751

Ordered Logistic Regression

- When it comes to ordered logistic regression, we can use the same latent variable formulation
- But now, we have not only one threshold (0 in the previous example) but **many** thresholds
- For example, with 4 categories, we have

$$y^* = \alpha + \beta x + \epsilon^*$$

and

$$y = \begin{cases} 1, & \text{if } y^* < \tau_1^* \\ 2, & \text{if } \tau_1^* \leq y^* < \tau_2^* \\ 3, & \text{if } \tau_2^* \leq y^* < \tau_3^* \\ 4, & \text{if } \tau_3^* \leq y^* \end{cases}$$

- Note that we have 3 thresholds if there are 4 categories

Simulation Code

```
* generate cut-points
gen taustar1 = -3
gen taustar2 = .5
gen taustar3 = 5

* generate outcome (note that we are "replacing")
drop y
gen y = 1
replace y = 2 if ystar > taustar1
replace y = 3 if ystar > taustar2
replace y = 4 if ystar > taustar3

* run logistic regression
ologit y x
```

Results

Here are the results:

```
Ordered logistic regression                     Number of obs   =      50000
                                                LR chi2(1)      =    7086.14
                                                Prob > chi2     =     0.0000
Log likelihood = -41251.691                     Pseudo R2       =     0.0791
```

y	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
x	.7995737	.0101209	79.00	0.000	.779737	.8194103
/cut1	-3.494251	.0241167			-3.541519	-3.446983
/cut2	.005631	.009527			-.0130415	.0243035
/cut3	4.515017	.0380231			4.440493	4.589541

- Note that we have **no constant(!)** and all cutpoints are off by **approximately .5** from the specified τ_k^*s (which were $\{-3, .5, 5\}$. Why?

- Here is why. Consider the inequality

$$y^* < \tau_1^*$$

as $y^* = \alpha + \beta x + \epsilon^*$, we have

$$\alpha + \beta x + \epsilon^* < \tau_1^*$$

Subtracting α from both sides yields

$$\beta x + \epsilon^* < \tau_1^* - \alpha$$

- The left-hand side is y^* **without constant** and the right-hand side is the threshold minus the constant (**which is set to .5**)
- The same applies to all the other categories