Solutions to Tutorial Questions - Week 5

Statistics and Econometrics

Question 1

Use the data in loanapp.RData for this exercise. The binary variable to be explained is *approve*, which is equal to one if a mortgage loan to an individual was approved. The key explanatory variable is *mortno*, a dummy variable equal to one if the applicant had no mortgage history.

1. Estimate a probit model of *approve* on *mortno*. Find the estimated probability of loan approval for those with no mortgage history and those who had mortgage before.
2. Now, add the variables *hrat*, *obrat*, *loanprc*, *unem*, *male*, *married*, *dep*, *sch*, *cosign*, *chist*, *pubrec*, and *vr* to the probit model. Is *mortno* still statistically significant?
3. Estimate the model from part 2 by logit.
4. Estimate the partial effects of *mortno* for probit and logit.

Solutions

1.

```
load("loanapp.RData")
loan.probit <- glm(approve ~ mortno, family = "binomial"(link = "probit"), data)
stargazer(loan.probit, header = FALSE, type = 'latex', title = "Question 1.1")
```

Table 1: Question 1.1

	Dependent variable:
	approve
mortno	0.394***
	(0.084)
Constant	1.051***
	(0.042)
Observations	1,989
Log Likelihood	−728.734
Akaike Inf. Crit.	1,461.467
Note:	*p<0.1; **p<0.05; ***p<0.01

```
predict(loan.probit, newdata = data.frame(mortno = 0), type = "response")
```

```
##         1
## 0.8532731
```

```
predict(loan.probit, newdata = data.frame(mortno = 1), type = "response")
```

```
##         1
## 0.9257576
```

As there is only one explanatory variable that takes on just two values, there are only two different predicted values: the estimated probabilities of loan approval for those with or without mortgage history. Rounded to three decimal places these are .853 for those who had mortgage before and .926 for those without mortgage history. These will be identical to the fitted values from the linear probability model. This must always be the case when the independent variables in a binary response model are mutually exclusive and exhaustive binary variables. Then, the predicted probabilities, whether we use the LPM, probit, or logit models, are simply the cell frequencies. (In other words, .853 is the proportion of loans approved for those who had mortgage before and .926 is the proportion approved for those with no mortgage history.)

2.

```
loan.probit2 <- glm(approve ~ mortno + hrat + obrat + loanprc + unem + male + married
                    + dep + sch + cosign + chist + pubrec + vr,
                    family = "binomial"(link = "probit"), data)
```

With the set of controls added, the probit estimate on *mortno* becomes about .249 (se $\approx$ .096). Therefore, *mortno* is still very significant.

3.

```
loan.logit <- glm(approve ~ mortno + hrat + obrat + loanprc + unem + male + married
                  + dep + sch + cosign + chist + pubrec + mortlat1 + mortlat2 + vr,
                  family = "binomial"(link = "logit"), data)
stargazer(loan.probit2, loan.logit, header = FALSE, type = 'latex', title = "Questions 1.2 and 1.3")
```

When we use logit instead of probit, the coefficient (standard error) on *mortno* becomes .475 (.189).

4.

```
mean(1/sqrt(2*pi) * exp(-loan.probit2$linear.predictors^2/2)) * loan.probit2$coefficients[2]
```

```
##     mortno
## 0.04217366
```

```
mean(loan.logit$fitted.values * (1 - loan.logit$fitted.values)) * loan.logit$coefficients[2]
```

```
##     mortno
## 0.04280316
```

We calculate average partial effects from both probit and logit models for comparison. The average effect from probit is given by

$$n^{-1}\sum_{i=1}^{n}\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{(\hat{\beta}_0+\hat{\beta}_1 x_{i1}+\cdots+\hat{\beta}_k x_{ik})^2}{2}\right)\hat{\beta}_{mortno}.$$

The average effect from logit is given by

$$n^{-1}\sum_{i=1}^{n}G(\hat{\beta}_0+\hat{\beta}_1 x_{i1}+\cdots+\hat{\beta}_k x_{ik})\cdot[1-G(\hat{\beta}_0+\hat{\beta}_1 x_{i1}+\cdots+\hat{\beta}_k x_{ik})]\hat{\beta}_{mortno},$$

where $G(z)=\frac{\exp(z)}{1+\exp(z)}$. Using the sample estimates, the average effect of *mortno* from probit is .0422, and the effect from logit is .0428. So results from the two model are consistent.

Question 2

Using the data in rental.RData for this exercise. The data on rental prices and other variables for college towns are for the years 1980 and 1990. The idea is to see whether a stronger presence of students affects rental rates. The fixed effects model is

$$\log(rent_{it})=\beta_0+\delta_0 y90+\beta_1\log(pop_{it})+\beta_2\log(avginc_{it})+\beta_3 pctstu_{it}+a_i+u_{it},$$

Table 2: Questions 1.2 and 1.3

	Dependent variable:	
	approve	
	probit	*logistic*
	(1)	(2)
mortno	0.249***	0.475**
	(0.096)	(0.189)
hrat	0.013*	0.022*
	(0.007)	(0.013)
obrat	−0.033***	−0.061***
	(0.006)	(0.011)
loanprc	−1.058***	−2.050***
	(0.240)	(0.462)
unem	−0.034*	−0.054*
	(0.018)	(0.033)
male	0.007	0.034
	(0.108)	(0.202)
married	0.232**	0.443**
	(0.094)	(0.177)
dep	−0.075*	−0.137*
	(0.038)	(0.072)
sch	0.041	0.093
	(0.094)	(0.176)
cosign	0.082	0.103
	(0.238)	(0.440)
chist	0.611***	1.107***
	(0.094)	(0.169)
pubrec	−0.843***	−1.436***
	(0.125)	(0.214)
mortlat1		−0.212
		(0.463)
mortlat2		−0.762
		(0.564)
vr	−0.236***	−0.423***
	(0.080)	(0.152)
Constant	2.438***	4.478***
	(0.304)	(0.579)
Observations	1,971	1,971
Log Likelihood	−612.266	−611.075
Akaike Inf. Crit.	1,252.532	1,254.149

where *pop* is city population, *avginc* is average income, and *pctstu* is student population as a percentage of city population (during the school year).

1. Estimate the equation as if we have a cross sectional data set (i.e., without a_i) and report the results. What do you make of the estimate on the 1990 dummy variable? What do you get for $\hat{\beta}_{pctstu}$?
2. Now estimate the equation using first-difference estimation. Compare your estimate of β_{pctstu} with that from part 1. Does the relative size of the student population appear to affect rental prices?

Solutions

1.

```
load("rental.RData")
rent.ols <- lm(log(rent) ~ y90 + log(pop) + log(avginc) + pctstu, data)
```

Estimating the equation without a_i, we obtain

$$\widehat{\log(rent)} = \underset{(.535)}{-.569} + \underset{(.035)}{.262}\, y90 + \underset{(.023)}{.041}\, \log(pop) + \underset{(.053)}{.571}\, \log(avginc) + \underset{(.001)}{.005}\, pctstu,$$

$n = 128, R^2 = 0.861, \bar{R}^2 = 0.857$. The positive and very significant coefficient on $y90$ simply means that, other things in the equation fixed, nominal rents grew by over 26% over the 10 year period. The coefficient on *pctstu* means that a one percentage point increase in *pctstu* increases rent by half a percent (.5%). The t statistic of five shows that, at least based on the usual analysis, *pctstu* is very statistically significant.

2.

```
rent.fd <- plm(log(rent) ~ y90 + log(pop) + log(avginc) + pctstu, data,
               index = c("city", "year"), effect = "individual", model = "fd")
```

The equation estimated by first difference is

$$\widehat{\Delta \log(rent)} = \underset{(.037)}{.386} + \underset{(.088)}{.072}\, \Delta \log(pop) + \underset{(.066)}{.310}\, \Delta \log(avginc) + \underset{(.0041)}{.0112}\, \Delta pctstu,$$

$n = 64, R^2 = .322, \bar{R}^2 = .302$. Interestingly, the effect of *pctstu* is over twice as large as we estimated in part 1. Now, a one percentage point increase in *pctstu* is estimated to increase rental rates by about 1.1%. While we have differenced away a_i, there may be other unobservables that change over time and are correlated with $\Delta pctstu$.