Problem Set 5 - Solutions

Statistics and Econometrics

Question 1

Use the data in pntsprd.RData for this exercise.

1. The variable $favwin$ is a binary variable indicating whether the team favored by the Las Vegas point spread wins (Here is an explanation of spread betting: https://en.wikipedia.org/wiki/Spread_betting). A linear probability model to estimate the probability that the favored team wins is

$$P(favwin = 1|spread) = \beta_0 + \beta_1 spread.$$

   Explain why, if the spread incorporates all relevant information, we expect $\beta_0 = .5$.
2. Estimate the model from part 1 by OLS. Test $H_0 : \beta_0 = .5$ against a two-sided alternative. Use both the usual and heteroskedasticity-robust standard errors.
3. Now, estimate a probit model for $P(favwin = 1|spread)$. Interpret and test the null hypothesis that the intercept is zero. [*Hint*: Remember that $\Phi(0) = .5$.]
4. Use the probit model to estimate the probability that the favored team wins when $spread = 10$. Compare this with the LPM estimate from part 2.
5. Add the variables $favhome$, $fav25$, and $und25$ to the probit model and test joint significance of these variables using the likelihood ratio test. (How many df are in the chi-square distribution?) Interpret this result, focusing on the question of whether the spread incorporates all observable information prior to a game.

Solutions

1. If *spread* is zero, there is no favorite, and the probability that the team we (arbitrarily) label the favorite should have a 50% chance of winning.

2.

```r
load("pntsprd.RData")
spread.lpm <- lm(favwin ~ spread, data)
linearHypothesis(spread.lpm, "(Intercept) = 0.5")
```

```
## Linear hypothesis test
##
## Hypothesis:
## (Intercept) = 0.5
##
## Model 1: restricted model
## Model 2: favwin ~ spread
##
##   Res.Df    RSS Df Sum of Sq      F   Pr(>F)
## 1    552 90.102
## 2    551 88.904  1    1.1984 7.4276 0.006627 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
coeftest(spread.lpm, vcov = vcovHC(spread.lpm, "HC1"))
```

```
##
## t test of coefficients:
##
##              Estimate Std. Error t value  Pr(>|t|)
## (Intercept) 0.5769492  0.0316568  18.225 < 2.2e-16 ***
## spread      0.0193655  0.0019218  10.077 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
linearHypothesis(spread.lpm, "(Intercept) = 0.5", white.adjust = "hc1")
```

```
## Linear hypothesis test
##
## Hypothesis:
## (Intercept) = 0.5
##
## Model 1: restricted model
## Model 2: favwin ~ spread
##
## Note: Coefficient covariance matrix supplied.
##
##   Res.Df Df      F  Pr(>F)
## 1    552
## 2    551  1 5.9084 0.01539 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
predict(spread.lpm, newdata = data.frame(spread = 10))
```

```
##         1
## 0.7706044
```

The linear probability model estimated by OLS gives

$$\widehat{favwin} = \underset{\substack{(.028)\\ [.032]}}{.577} + \underset{\substack{(.0023)\\ [.0019]}}{.0194}\,spread,$$

$n = 553, R^2 = .111$. where the usual standard errors are in $(\cdot)$ and the heteroskedasticity-robust standard errors are in $[\cdot]$. Using the usual standard error, the t statistic for $H_0 : \beta_0 = .5$ is $(.577 - .5)/.028 = 2.75$, which leads to rejecting H_0 against a two-sided alternative at the 1% level (critical value ≈ 2.58). Using the robust standard error reduces the significance but nevertheless leads to strong rejection of H_0 at the 2% level against a two-sided alternative: $t = (.577 - .5)/.032 \approx 2.41$ (critical value ≈ 2.33).

As we expect, *spread* is very statistically significant using either standard error, with a t statistic greater than eight. If $spread = 10$ the estimated probability that the favored team wins is $.577 + .0194(10) = .771$.

3.

```r
spread.probit <- glm(favwin ~ spread, family = "binomial"(link = "probit"), data)
stargazer(spread.probit, header = FALSE, type = 'latex', title = "Question 1.3 - Probit Model")
```

```r
linearHypothesis(spread.probit, "(Intercept) = 0")
```

```
## Linear hypothesis test
##
## Hypothesis:
```

Table 1: Question 1.3 - Probit Model

	Dependent variable:
	favwin
spread	0.092^{***}
	(0.012)
Constant	-0.011
	(0.103)
Observations	553
Log Likelihood	-263.562
Akaike Inf. Crit.	531.124
Note:	*p<0.1; **p<0.05; ***p<0.01

```
## (Intercept) = 0
##
## Model 1: restricted model
## Model 2: favwin ~ spread
##
##   Res.Df Df  Chisq Pr(>Chisq)
## 1    552
## 2    551  1 0.0105     0.9185
```

In the Probit model

$$P(favwin = 1|spread) = \Phi(\beta_0 + \beta_1 spread),$$

where $\Phi(\cdot)$ denotes the standard normal cdf, if $\beta_0 = 0$ then

$$P(favwin = 1|spread) = \Phi(\beta_1 spread)$$

and, in particular, $P(favwin = 1|spread = 0) = \Phi(0) = .5$. This is the analog of testing whether the intercept is .5 in the LPM. From the table, the t statistic for testing $H_0 : \beta_0 = 0$ is only about $-.102$, so we do not reject H_0.

4.

```
predict(spread.probit, newdata = data.frame(spread = 10), type = "response")
```

```
##         1
## 0.8196512
```

When $spread = 10$, the predicted response probability from the estimated Probit model is $\Phi[-.0106 + .0925(10)] = \Phi(.9144) \approx .820$. This is above the estimate for the LPM.

5.

```
spread.probit2 <- glm(favwin ~ spread + favhome + fav25 + und25,
                      family = "binomial"(link = "probit"), data)
logLik(spread.probit2)
```

```
## 'log Lik.' -262.6418 (df=5)
```

```
1 - pchisq(spread.probit$deviance - spread.probit2$deviance, 3)
```

```
## [1] 0.6060875
```

```r
linearHypothesis(spread.probit2, c("favhome = 0", "fav25 = 0", "und25 = 0"))
```

```
## Linear hypothesis test
##
## Hypothesis:
## favhome = 0
## fav25 = 0
## und25 = 0
##
## Model 1: restricted model
## Model 2: favwin ~ spread + favhome + fav25 + und25
##
##   Res.Df Df Chisq Pr(>Chisq)
## 1    551
## 2    548  3 1.851     0.6039
```

When $favhome$, $fav25$, and $und25$ are added to the Probit model, the value of the log-likelihood becomes -262.64. Therefore, the likelihood ratio statistic is $2[-262.64-(-263.56)] = 2(263.56-262.64) = 1.84$. The p-value from the χ^2_3 distribution is about .61, so $favhome$, $fav25$, and $und25$ are jointly very insignificant. Once *spread* is controlled for, these other factors have no additional power for predicting the outcome.

Question 2

For this exercise, we use jtrain.RData to determine the effect of the job training grant on hours of job training per employee. The basic model for the three years is

$$\begin{aligned} hrsemp_{it} &= \beta_0 + \delta_1 d88_t + \delta_2 d89_t + \beta_1 grant_{it} \\ &\quad +\beta_2 grant_{i,t-1} + \beta_3 \log(employ_{it}) + a_i + u_{it}, \end{aligned}$$

where $hrsemp_{it}$ indicates the average number of hours training per employee for firm i in time period t; $grant_{it}$ is a dummy variable, which is equal to 1 if firm i received a job training grant in time period t, and 0 otherwise; $employ_{it}$ indicates the number of employees at firm i in time period t.

1. Estimate the equation using fixed effects estimation (i.e., model = “within”). How many firms are used in the estimation? How many total observations would be used if each firm had data on all variables (in particular, $hrsemp$) for all three time periods?
2. Interpret the coefficient on $grant_{it}$ and comment on its significance.
3. Is it surprising that $grant_{i,t-1}$ is insignificant? Explain.
4. Do larger firms train their employees more or less, on average? How big are the differences in training?

Solutions

1.

```r
load("jtrain.RData")
hrsemp.fe <- plm(hrsemp ~ d88 + d89 + grant + grant_1 + log(employ),
                 data, index = c("fcode", "year"), model = "within")

stargazer(hrsemp.fe, header = FALSE, type = 'latex', title = "Question 2.1")
```

```r
data.p <- pdata.frame(data, index = c("fcode", "year"))
pdim(data.p)
```

```
## Balanced Panel: n = 157, T = 3, N = 471
```

Table 2: Question 2.1

	Dependent variable:
	hrsemp
d88	−1.099
	(1.983)
d89	4.090
	(2.481)
grant	34.228***
	(2.858)
grant_1	0.504
	(4.127)
log(employ)	−0.176
	(4.288)
Observations	390
R^2	0.491
Adjusted R^2	0.208
F Statistic	48.206*** (df = 5; 250)
Note:	*p<0.1; **p<0.05; ***p<0.01

```
data.new <- data %>% select(hrsemp, fcode, year, d88, d89, grant, grant_1, employ) %>% na.omit
na.stats <- tapply(data.new$hrsemp, data.new$fcode, length)
table(na.stats)
```

```
## na.stats
##   1   2   3
##   4   7 124
```

There are 124 firms with all three years of data, 7 firms with two years of data and 4 firms with one year of data. 22 firms in the sample have missing information from all three years and are not used at all. Due to missing data, we use only $390(124 \times 3 + 7 \times 2 + 4)$ observations in the fixed-effects estimation. If we had information for all 157 firms, we would have 471 total observations in estimating the equation.

2. The coefficient on *grant* means that if a firm received a grant for the current year, it trained each worker an average of 34.228 hours more than it would have otherwise. This is a practically large effect, and the t statistic is very large.

3. Since a grant last year was used to pay for training last year, it is perhaps not surprising that the grant does not carry over into more training this year. It would if inertia played a role in training workers.

4. The coefficient on the employees variable is very close to zero: a 10% increase in *employ* decreases predicted hours per employee by about 0.0176. [Recall: $\Delta\widehat{hrsemp} \approx -(.176/100)(\%\Delta employ)$.] This is pratically small, and the t statistic is also rather small.

Question 3

As discussed in the lecture, we can potentially evaluate performance of Logit/Probit models based on measures calculated from the confusion matrix, such as precision and recall. However, in many cases, these two measures may send conflicting messages. Let's say we have two models (M1 and M2), where Precision of M1 = 0.8 and Recall of M1 = 0.4, whereas Precision of M2 = 0.4 and Recall of M2 = 0.8. How would you choose between the two models? Explain.

Solution

The choice between the two models depends on the costs associated with the Type I error (as measured by precision) and Type II error (as measured by recall). Using the example on Slide 14, if we mistakenly predict a non-pregnant customer to be pregnant (Type I error), we may send her some unnecessary flyers/brochures/emails of baby products, where the cost is relatively small. However, if we mistakenly predict a pregnant customer to be non-pregnant (Type II error), she may do her shopping of pushchair, baby products and maternity wear from Target's competitors, and thus the company loses all the profit from selling those products. So in the context of this example, I would say minimizing Type II error is more important, and thus M2 is preferred.

Of course, the preceding discussion is based on the assumption that we have to choose one between M1 and M2 with certain thresholds being used to convert probabilities into classifications. In practice, the errors are sensitive to the threshold being used, meaning that a larger threshold typically leads to a smaller number of Type I errors, but a greater number of Type II errors. One can have a more comprehensive understanding of the performance of a model by checking out ROC (Receiver Operator Characteristic Curve), which you will learn in the advanced machine learning class.