Statistics and Econometrics: Problem Set 5

Siow Meng Low

10 November 2016

Question 1

Question 1a: Linear Probability Model

The equation of the linear probability model is given as:

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

The *spread* variable is the Las Vegas point spread (used for sports betting purpose). When $spread = 0$, neither team is favoured by the betting system. If the *spread* variable incorporates all relevant information (e.g. historical performance, favoured team is at home), then neither team has an advantage over the other since there is no favourite. In this case, we would expect the probability of winning (for the "favoured" team) to be 0.5, since both teams are equally likely to win the match.

When $spread = 0$, we could rewrite the equations as:

$$P(favwin = 1|spread = 0) = \beta_0$$

As discussed earlier, we expect the probability to be 0.5 when $spread = 0$. Since this probability is equal to β_0, we expect $\beta_0 = 0.5$ as well.

Question 1b: OLS Estimation of the Linear Probability Model

Table 1 displays the model information of the linear probability model. The left column contains the information of usual standard errors (in parentheses) whereas the right column shows the robust standard errors (in parentheses).

Table 1: Question 1b - LPM with Standard Errors and Robust Standard Errors

	Probability of Favoured Team's Win	
	Usual Standard Errors	Robust Standard Errors
	(1)	(2)
spread	0.0194***	0.0194***
	(0.0023)	(0.0019)
Constant	0.5769***	0.5769***
	(0.0282)	(0.0317)
N	553	553
R^2	0.1107	0.1107
Adjusted R^2	0.1091	0.1091
Residual Std. Error (df = 551)	0.4017	0.4017
F Statistic (df = 1; 551)	68.5691***	68.5691***

Notes: ***Significant at the 1 percent level.
**Significant at the 5 percent level.
*Significant at the 10 percent level.

The null hypothesis and alternative hypothesis are:

- H_0: $\beta_0 = 0.5$
- H_1: $\beta_0 \neq 0.5$

We would need to do a two-tail test to test the null hypothesis. The critical value at 5% significance level (for two-tail test) is thus 1.96 (assume normal distribution since we have large degree of freedom, > 120).

Using the usual standard error, the t-statistic is $\dfrac{\hat{\beta}_0 - 0.5}{se(\hat{\beta}_0)} = 2.7254$ and the corresponding p-Value is 0.0064. The t-statistic is a lot larger than the critical value (and also p-Value is a lot smaller than 0.05), hence we reject $H_0 : \beta_0 = 0.5$

Using robust standard error, the t-statistic is $\dfrac{\hat{\beta}_0 - 0.5}{robust\, se(\hat{\beta}_0)} = 2.4307$ and the corresponding p-Value is 0.0151. The t-statistic is a larger than the critical value (and also p-Value is smaller than 0.05), hence we also reject $H_0 : \beta_0 = 0.5$

In conclusion, we reject the null hypothesis $H_0 : \beta_0 = 0.5$ after doing a two-tail test using both usual and robust standard errors.

Question 1c: Probit Model

Table 2 displays the model information of the probit model for $P(favwin = 1|spread)$.

Table 2: Probit Model for Question 1c

	Probability of Favoured Team's Win
spread	0.0925***
	(0.0121)
Constant	−0.0106
	(0.1035)
N	553
Log Likelihood	−263.5622
Akaike Inf. Crit.	531.1244
Notes:	***Significant at the 1 percent level.
	**Significant at the 5 percent level.
	*Significant at the 10 percent level.

$Pseudo\, R^2 = 0.1294$, Percentage Correctly Predicted = 76.31%

The Probit model uses the standard normal cumulative distribution function $G(z) = \Phi(z)$. In our case, the equation of the probit model is:

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

Let us consider the case when $spread = 0$, then the equation becomes $P(favwin = 1|spread = 0) = \Phi(\beta_0)$. If the intercept is zero, the probability becomes $P(favwin = 1|spread = 0) = \Phi(0) = 0.5$.

Similar to reasoning in Section 1a, we would exepect the probability of winning (by "favoured" team) to be 0.5 when $spread = 0$ and if *spread* incorporates all relevant information. By testing the null hypothesis $H_0 : \beta_0 = 0$ for the probit model, we can examine whether the probability is indeed 0.5 when $spread = 0$

The null hypothesis and alternative hypothesis are:

- H_0: $\beta_0 = 0$
- H_1: $\beta_0 \neq 0$

Since the MLE estimator of Probit is asymptotically normal, we could use the two-tail Z-test to test the null hypothesis. The critical value at 5% significance level (for two-tail test) is thus 1.96.

The test statistic is $\frac{\hat{\beta}_0}{se(\hat{\beta}_0)}$ = -0.1023 and the corresponding p-Value is 0.9185. The absolute value of the test statistic is 0.1023 and it is way smaller than the critical value (and also p-Value is extremely large), hence we fail to reject the null hypothesis that the intercept is zero, $H_0 : \beta_0 = 0$

Note that the result of this hypothesis testing indicates that the probability of "favoured" team's win is likely to be near to 0.5 when $spread = 0$, when we use Probit model. This conclusion is different from Question 1b when we tested the Linear Probability Model.

Question 1d: Probability Prediction when *spread* = 10

When $spread = 10$, the probit model estimates the probability: $\hat{P}(favwin = 1|spread = 10) = \Phi(\hat{\beta}_0 + \hat{\beta}_1 spread) = \Phi(-0.0106 + 0.0925(10)) = 0.8197$

The LPM model estimates: $\hat{P}(favwin = 1|spread = 10) = \hat{\beta}_0 + \hat{\beta}_1 spread = 0.5769 + 0.0194(10) = 0.7706$

The probability estimated by the probit model is a bit higher than the LPM model.

Question 1e: Joint Significance

Table 3 displays the model information of the probit model for $P(favwin = 1|spread, favhome, fav25, und25)$, after adding in independent variables *favhome*, *fav25*, *und25*.

Table 3: Question 1e - Probit Model with Additional Variables

	Probability of Favoured Team's Win
spread	0.0879***
	(0.0128)
favhome	0.1486
	(0.1369)
fav25	0.0031
	(0.1588)
und25	−0.2198
	(0.2513)
Constant	−0.0552
	(0.1292)
N	553
Log Likelihood	−262.6418
Akaike Inf. Crit.	535.2835
Notes:	***Significant at the 1 percent level.
	**Significant at the 5 percent level.
	*Significant at the 10 percent level.

$Pseudo\, R^2 = 0.1325$, Percentage Correctly Predicted = 75.95%

Next, we would like to test the joint significance of the three new independent variables. The unrestricted model would be this model and the restricted model is the probit model in Section 1c. Since we have added in 3 new independent variables, the likelihood ratio statistics will follow a chi-square distribution of 3 degrees of freedom: $LR = 2(L_{ur} - L_r) \sim \chi^2_3$

For a chi-square (df = 3) distribution, the critical value at 5% significance level is 7.8147

The value of the Likelihood Ratio Statistic is: $2(L_{ur} - L_r) = 2(-262.642 - -263.562) = 1.841$ and the corresponding p-Value is 0.6061.

The test statistic is way smaller than the critical value (and the p-Value is a lot larger than 0.05), hence we fail to reject the null hypothesis $H_0 : \beta_{favhome} = 0, \beta_{fav25} = 0, \beta_{und25} = 0$. In other words, these three variables are jointly insignificant.

From the testing results, the variable *spread* seems to incorporate the observable information contained in the three independent variables *favhome*, *fav25*, and *und25*. Consequently, when *spread* has been controlled for, these three variables becomes jointly very insignificant (and can be dropped from the model).

Note that since we only did the joint significance test for three independent variables (*favhome*, *fav25*, and *und25*), we can't say that *spread* incorporates "all" other observable information (e.g. *neutral*, *fregion*, *uregion*) as well. Further join significance test can be performed if we would like to find out.

Question 2

Question 2a: First Differencing

The equation of the basic model is given as:

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

The first-differenced equation is thus:

$$\Delta hrsemp_i = \delta_1 \Delta d88_t + \delta_2 \Delta d89_t + \beta_1 \Delta grant_i + \beta_2 \Delta grant_{i,-1} + \beta_3 \Delta log(employ_i) + \Delta u_i$$

From the above equation, it is clear that we would use a first-differenced model without intercept for estimation. Table 4 displays the model information of the first-differenced estimation.

Table 4: First-Differenced Model for Question 2a

	Difference in Hours of Job Training per Employee
d88	−0.7583
	(1.9185)
d89	4.0345
	(3.1821)
grant	32.3537***
	(2.8800)
grant_1	1.1741
	(5.1711)
lemploy	0.3491
	(4.7005)
N	255
R^2	0.4751
Adjusted R^2	0.4658
F Statistic	45.2588*** (df = 5; 250)
Notes:	***Significant at the 1 percent level.
	**Significant at the 5 percent level.
	*Significant at the 10 percent level.

There are a total of 157 firms in the dataset. However, some of observations have missing values (i.e. NA) for *hrsemp* and *lemploy*. For first-differenced estimation, a firm must have valid observations (i.e. no missing

values for the dependent and independent variables) for at least 2 time periods in order to be used in the estimation. Therefore, the total number of firms that are used in the estimation is 131 and the number of observations used is 255

If each firm had data on all variables for all the three periods, the number of observations used would be 157(3 - 1) = 314 observations.

Question 2b: Coefficient of $\Delta grant_i$

From Table 4, the coefficient of $\Delta grant_i$, is estimated to be $\hat{\beta}_1 = 32.35$. This means that a firm which has newly received a grant for the current year (and did not receive a grant for the previous year), is estimated to increase the job training hours per employee by 32.35 hours.

The p-Value for $H_0 : \beta_1 = 0$ is extremely small: 5.642e-24, this means that $\Delta grant_i$ is statistically very significant.

Question 2c: Significance $\Delta grant_{i,-1}$

From Table 4, we can see that $\Delta grant_{i,-1}$ is statistically insignificant. The p-Value for $H_0 : \beta_2 = 0$ is a high value: 0.8206, indicating it is statistically very insignificant.

This should not be surprising because a grant is intended to encourage the firm to train their employees in the same year. It is likely that the grant awarded in the previous year would only be valid (and used up) within that same year, hence it would not have a effect on the training hours of current year.

Question 2d: Training at Larger Firms

From Table 4, we can see that the coefficient of $\Delta log(employ_i)$, is estimated to be $\hat{\beta}_3 = 0.3491$. This means on average, every 1% increase in the number of employees only result in the increase of 0.0035 training hours per employee. This difference is practically very small.

In addition, the p-Value for $H_0 : \beta_3 = 0$ is an extremely high value: 0.941. In other words, we cannot reject the null hypothesis that $\Delta log(employ_i)$ is statistically insignificant. Due to the insignificance of $\Delta log(employ_i)$ in our estimates, we conclude that there is no difference between the training hours (per employee) in large firms and small firms.