Problem Set 5

Statistics and Econometrics

Deadline: 11am, 2 December 2020

General Guideline

What we are looking for in the assignments is a demonstration that you can understand the econometrics and statistics questions and can solve them with R or conceptually. That means effective programming to get correct results is needed, but at the same time, clear explanations of economics/business concepts in well presented reports are equally important when assessing your work. In particular, you will be marked for successful (correct) programming (not the style of coding), good understanding of related concepts, and clear interpretations and explanations of results.

Please submit a pdf or html file converted from R markdown/notebook after you program in R.

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.

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?

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.