Problem Set 2

November 9, 2020

Rules

- Solutions to problem sets must be completed and submitted individually.
- The deadline for submission is **23.59 on 19 November 2020**. Submissions after the deadline will not be accepted.
- The datasets for the STATA exercises can be found at SUCourse.
- **Submissions must include two files: 1) one word file including your answers 2) one STATA log file.** Results in the log file and interpreted results must match. Otherwise, you will get zero points for those exercises. **Do not upload zipped files.**
- You are expected to answer all questions very clearly. Therefore, please be careful about use of language and writing. **For instance; if you are asked to interpret a coefficient, you need to interpret it in a way that someone who cannot see the data and does not know econometrics can understand what you mean.**
- Double check that you write your name/surname and student ID number.
- Failure in fulfilling any of these will result in a FAIL grade for that homework.

Question 1 (35 points)

Suppose that you are interested in estimating the causal relationship between y and x_1. For this purpose, you can collect data on a control variable, x_2. (For concreteness, you might think of y as final exam score, x_1 as class attendance and x_2 as SAT score.) Let $\tilde{\beta}_1$ be the simple regression estimate from y on x_1 and let $\hat{\beta}_1$ be the multiple regression estimate from y on x_1, controlling for x_2.

1) If x_1 is positively correlated with x_2, and x_2 has an effect on y, would you expect $\tilde{\beta}_1$ and $\hat{\beta}_1$ to be similar or different? Show the direction of potential bias in $\tilde{\beta}_1$ of the simple regression estimate when we omit x_2. (10 points)
2) If x_1 is highly correlated with x_2, would you expect $se(\tilde{\beta}_1)$ or $se(\hat{\beta}_1)$ to be smaller (or unclear)? Discuss your answer based on the formula of $se(\hat{\beta}_1)$. (5 points)
3) Discuss how the $se(\hat{\beta}_1)$ would be affected if you add more independent variables (in addition to x_1 and x_2) to the model. Suppose these new additional variables have have small correlation with x_1 and have high correlation with y. (5 points)
4) Discuss how the $se(\hat{\beta}_1)$ would be affected if the sample size was increased by four times (instead of: n=x, the sample size will be: n=4x). Calculate the approximate expected change in $se(\hat{\beta}_1)$. (5 points)
5) Discuss how the scenarios in question (2), (3) and (4) would effect the statistical significance of $\hat{\beta}_1$ and the size of confidence intervals constructed for $\hat{\beta}_1$. Discuss each scenario separately. (10 points)

Question 2 (40 points)

Use the data in DISCRIM.DTA to answer this question. These are zip code-level data on prices for various items at fast-food restaurants, along with characteristics of the zip code population, in New Jersey and Pennsylvania. The idea is to see whether fast-food restaurants charge higher prices in areas with a larger concentration of blacks.

1) Consider a model to explain the price of soda, *psoda*, in terms of the proportion of the population that is black and median income:

$$psoda = \beta_0 + \beta_1 prpblck + \beta_2 income + u,$$

Estimate this model by OLS and report the results, including the sample size and R-squared. Interpret the sign and magnitude of the coefficients on *prpblck* and *income*. (10 points)
2) Is the coefficient on *prpblck* statistically significant (different from zero) at 0.05 significance level? What

is the minimum significance level that we can say the coefficient is significantly different from zero? Is the coefficient significantly different from ”0.1” at 0.05 significance level? (10 points)

3) Compare the estimate from part (1) with the simple regression estimate from *psoda* on *prpblck*. Is the discrimination effect larger or smaller when you control for income? Explain why the effect is larger or smaller when you control for income (Hint: you are expected to use correlations in your explanation). (10 points)

4) A model with a constant price elasticity with respect to income may be more appropriate. Report estimates of the model

$$log(psoda) = \beta_0 + \beta_1 prpblck + \beta_2 log(income) + u,$$

If *prpblck* increases by 0.20 (20 percentage points), what is the estimated percentage change in *psoda*? (5 points)

5) Now add the variable *prppov* to the regression in part (1). Find the correlation between *income* and *prppov*. Is it roughly what you expected? Evaluate the following statement: ”Because *income* and *prppov* are so highly correlated, they have no business being in the same regression”. (5 points)

Question 3 (25 points)

The following model can be used to study whether campaign expenditures affect election outcomes:

$$voteA = \beta_0 + \beta_1 log(expendA) + \beta_2 log(expendB) + \beta_3 prtystrA + u,$$

where *voteA* is the percentage of the vote received by Candidate A, *expendA* and *expendB* are campaign expenditures by Candidates A and B, and *prtystrA* is a measure of party strength for Candidate A (the percentage of the most recent presidential vote that went to A’s party)

1) Estimate the given model using the data in VOTE1.DTA and report the results in usual form. Interpret the sign and magnitude of the estimated effect of A’s expenditures on the outcome (β_1). Do you think the estimated effect is the causal effect? What is the main assumption we make to say it is a causal effect? Discuss the validity of that assumption by providing an example. (10 points)

2) Interpret the explanatory power (R^2) of the estimated model. Is it an evidence for causal effect of expendA on voteA? (5 points)

3) Based on the estimation results, discuss the statistical significance of the estimated coefficient for A’s expenditure (at 0.05 significance level). Explain your answer by using different approaches: 1) use reported t-value, 2) use reported p-value, 3) use reported confidence interval (10 points)