Daniel Carpenter

ID: 113009743

PS2

Warm-Up 3.7

1. **Which of the following can cause OLS estimators to be biased?**
    a. Omitting an important variable.

Exercises 3.9

The following equation describes the median housing price in a community in terms of amount of pollution (*nox* for nitrous oxide) and the average number of rooms in houses in the community (rooms):

1. **What are the probable signs of B_1 and B_2? What is the interpretation of B_1? Explain.**
    a. B_1: Potentially positive. It is unclear to me that there would be more or less pollution for a household with higher incomes. A person with a high wage may invest in more carbon efficient mechanisms, thus reducing pollution. However, they may own more cars or machines that cause greater pollution.
    b. B_2: Positive. I expect the price of the house to rise with the number of rooms in the household. Therefore, the owners need a higher wage to buy the expanded house.

2. **Why might *nox* [or more precisely, log(*nox*] and rooms be negatively correlated? If this is the case, does the simple regression of log(*price)* on log(*nox)* produce an upward or a downward biased estimator of B_1?**
    a. A negative correlation may exist due to the investment in more carbon neutral household items. They also may have better insulation, which would probably have lower emissions in the long run.
    b. If it is negatively correlated, it would produce a downward sloping graph.

3. **Is the relationship between the simple and multiple regression estimates of the elasticity of price with respect to nox what you would have predicted, given your answer in part (ii)? Does this mean that −.718 is definitely closer to the true elasticity than −1.043?**
    a. The simple regression model was unexpected. However, the multiple regression model seems more accurate. The elasticity changed, meaning that when adding number of rooms, the change lessened for the change in *nox*.
    b. It does not mean that it is *definitely* closer to the true elasticity, but it most likely does mean that it is closer to the true elasticity. When adding more variables, the interpreter can expect a more true model.

Exercises 7.4

The data used are in CEOSAL1, where finance, consprod, and utility are binary variables indicating the financial, consumer products, and utilities industries. The omitted industry is transportation.

1. **Compute the approximate percentage difference in estimated salary between the utility and transportation industries, holding sales and roe fixed.**
   a. -28.3% difference in utility

2. **Use equation (7.10) to obtain the exact percentage difference in estimated salary between the utility and transportation industries and compare this with the answer obtained in part (i).**
   a. -24.7%

3. **What is the approximate percentage difference in estimated salary between the consumer products and finance industries? Write an equation that would allow you to test whether the difference is statistically significant.**
   a. 2.3% Difference

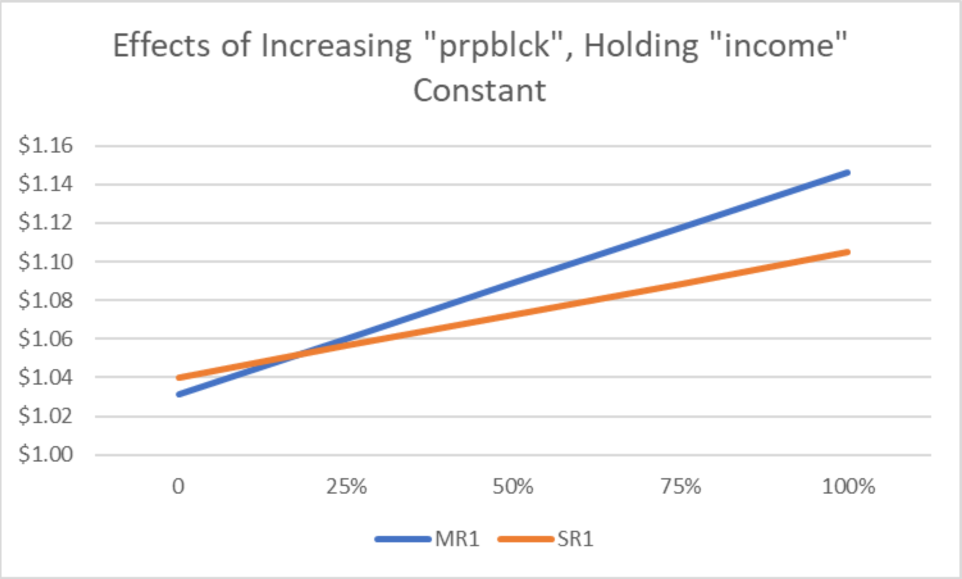
3.C8

Use the data in DISCRIM 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. **Find the average values of prpblck and income in the sample, along with their standard deviations. What are the units of measurement of prpblck and income?**
    a. Proportion of Blacks in Zip Code
        i. Avg.: 11% of zip code black
        ii. St.Dev: 18% standard deviation
        iii. Units of Measurement: Percentage representation of zip code
    b. Median Income per Zip Code
        i. Avg.: $47,053.78 per household
        ii. St.Dev: $13,179.29 per household
        iii. Units of Measurement: Dollars per household

2. **Consider a model to explain the price of soda, psoda, in terms of the proportion of the population that is black and median income: Estimate this model by OLS and report the results in equation form, including the sample size and R-squared. (Do not use scientific notation when reporting the estimates.) Interpret the coefficient on prpblck. Do you think it is economically large?**
    a. psoda = .956 + .115(prpblck) + 0.0000016(income)
    b. The coefficient on “prpblck” means that with one percentage increase in the proportion of black individuals in the zip code, the result will be a $0.00115 in the price of soda.
    c. This is very insignificant. When plugging in various levels of proportions of blacks or incomes, the price of soda increases by extremely small amounts.

3. **Compare the estimate from part (ii) with the simple regression estimate from psoda on prpblck. Is the discrimination effect larger or smaller when you control for income?**
    a. Psoda = 1.04 + 0.0649(prpblck)
    b. When holding income constant, we are able to see as the proportion of black population increases, the multivariate regression produces a higher price level of soda.



c.

4. **A model with a constant price elasticity with respect to income may be more appropriate. Report estimates of the model**
    a. psoda = 1 + 0.00000007(income)

5. **If prpblck increases by .20 (20 percentage points), what is the estimated percentage change in psoda? (Hint: The answer is 2.xx, where you fill in the "xx.")**
    a. 2.44%

6. **Now add the variable prppov to the regression in part (iv). What happens to ?**
    a. The coefficient of prpblck decreases from 0.122 to 0.0728

7. **Find the correlation between log(income) and prppov. Is it roughly what you expected?**
    a. Negative correlation. Yes.

8. Evaluate the following statement: "Because and prppov are so highly correlated, they have no business being in the same regression."
    a. False. They may experience a high correlation but adding in other variables will allow to have a greater picture of the causal effects of the discrimination. They may be correlated, but this may not be the cause.

7.C8

Use the data in LOANAPP 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 white, a dummy variable equal to one if the applicant was white. The other applicants in the data set are black and Hispanic.

1. **To test for discrimination in the mortgage loan market, a linear probability model can be used. If there is discrimination against minorities, and the appropriate factors have been controlled for, what is the sign of b1?**
   a. There would be a positive correlation between number of white individuals applying on approval rates.

2. **Regress approve on white and report the results in the usual form. Interpret the coefficient on white. Is it statistically significant? Is it practically large?**
   a. Approval = 0.708 + 0.201(white)
   b. If you are a white individual, there is a 20.1% greater probability of you getting accepted.

3. **As controls, add the variables hrat, obrat, loanprc, unem, male, married, dep, sch, cosign, chist, pubrec, mortlat1, mortlat2, and vr. What happens to the coefficient on white? Is there still evidence of discrimination against nonwhites?**
   a. The coefficient of white decreases from 0.201 to 0.129. Although the coefficient decreased by nearly 35%, it does not push the variable to 0. Thus, there is still evidence of discrimination of approval rates.

4. **Now, allow the effect of race to interact with the variable measuring other obligations as a percentage of income (obrat). Is the interaction term significant?**
   a. White = 0.993 – 0.00458(obrat)
   b. The effect that obrat has on white seems small.

5. **Using the model from part (iv), what is the effect of being white on the probability of approval when *obrat* = 32, which is roughly the mean value in the sample? Obtain a 95% confidence interval for this effect.**
   a. Confused on how to calculate, will be reaching out soon.
   b. Confidence interval results: -31.9, -31.17.

Exercise 7.5

In Example 7.2, let noPC be a dummy variable equal to one if the student does not own a PC, and zero otherwise.

1. **If noPC is used in place of PC in equation (7.6), what happens to the intercept in the estimated equation? What will be the coefficient on noPC? (Hint: Write and plug this into the equation icon.)**
   a. colGPA = B_0 + S_0 – S_0noPC + B_1hsGPA + B_2ACT
      1.26 + 0.157 = 1.417
      noPC Coefficient: –0.157

2. **What will happen to the R-squared if noPC is used in place of PC?**
   a. Nothing will happen to the R squared.

3. **Should PC and noPC both be included as independent variables in the model? Explain.**
   a. No. We are unable to use two dummy variables in one equation because it will cause perfect collinearity, thus breaking a Gauss Markov assumption.