

Revision #1 – Multiple Regression, Dummy Variable

Q1: Multiple Regression and Omitted Variable Bias (Chapter 3, Q9)

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*):

$$\log(\text{price}) = \beta_0 + \beta_1 \log(\text{nox}) + \beta_2 \text{rooms} + u.$$

- i. What are the probable signs of β_1 and β_2 ? What is the interpretation of β_1 ? Explain.
- ii. Why might *nox* [or more precisely, $\log(\text{nox})$] and *rooms* be negatively correlated? If this is the case, does the simple regression of $\log(\text{price})$ on $\log(\text{nox})$ produce an upward or a downward biased estimator of β_1 ?
- iii. Using the data in HPRICE2, the following equations were estimated:

$$\widehat{\log(\text{price})} = 11.71 - 1.043 \log(\text{nox}), n = 506, R^2 = .264.$$

$$\widehat{\log(\text{price})} = 9.23 - .718 \log(\text{nox}) + .306 \text{rooms}, n = 506, R^2 = .514.$$

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 ?

Q2: Multiple Regression (Inference) (Chapter 4, Q13)

The data in MEAPSINGLE were used to estimate the following equations relating school-level performance on a fourth-grade math test to socioeconomic characteristics of students attending school. The variable *free*, measured at the school level, is the percentage of students eligible for the federal free lunch program. The variable *medinc* is median income in the ZIP code, and *pctsgle* is percent of students not living with two parents (also measured at the ZIP code level). See also [Computer Exercise C11](#) in [Chapter 3](#).

$$\widehat{\text{math4}} = 96.77 - .833 \text{ pctsgle} \\ (1.60) \quad (.071) \\ n = 299, R^2 = .380$$

$$\widehat{\text{math4}} = 93.00 - .275 \text{ pctsgle} - .402 \text{ free} \\ (1.63) \quad (.117) \quad (.070) \\ n = 299, R^2 = .459$$

$$\widehat{\text{math4}} = 24.49 - .274 \text{ pctsgle} - .422 \text{ free} - .752 \text{ lmedinc} + 9.01 \text{ lexppp} \\ (59.24) \quad (.161) \quad (.071) \quad (5.358) \quad (4.04) \\ n = 299, R^2 = .472$$

$$\widehat{\text{math4}} = 17.52 - .259 \text{ pctsgle} - .420 \text{ free} + 8.80 \text{ lexppp} \\ (32.25) \quad (.117) \quad (.070) \quad (3.76) \\ n = 299, R^2 = .472.$$

- i. Interpret the coefficient on the variable *pctsgle* in the first equation. Comment on what happens when *free* is added as an explanatory variable.
- ii. Does expenditure per pupil, entered in logarithmic form, have a statistically significant effect on performance? How big is the estimated effect?

Q3: Multiple Regression Analysis with Qualitative Information (Chapter 7, Q8)

Suppose you collect data from a survey on wages, education, experience, and gender. In addition, you ask for information about marijuana usage. The original question is: "On how many separate occasions last month did you smoke marijuana?"

- i. Write an equation that would allow you to estimate the effects of marijuana usage on wage, while controlling for other factors. You should be able to make statements such as, "Smoking marijuana five more times per month is estimated to change wage by $x\%$."
- ii. Write a model that would allow you to test whether drug usage has different effects on wages for men and women. How would you test that there are no differences in the effects of drug usage for men and women?
- iii. Suppose you think it is better to measure marijuana usage by putting people into one of four categories: nonuser, light user (1 to 5 times per month), moderate user (6 to 10 times per month), and heavy user (more than 10 times per month). Now, write a model that allows you to estimate the effects of marijuana usage on wage.
- iv. Using the model in part (iii), explain in detail how to test the null hypothesis that marijuana usage has no effect on wage. Be very specific and include a careful listing of degrees of freedom.
- v. What are some potential problems with drawing causal inference using the survey data that you collected?