Assignment #3

ECON 2023 Introductory Econometric

March 18, 2023

1 Multiple Regression

1. Using the data in GPA2 on 4,137 college students, the following equation was estimated by OLS:

$$\widehat{colgpa} = 1.392 - 0.315hsperc + 0.0148sat$$
$$n = 4,137, R^2 = 0.273$$

where *colgpa* is measured on a four-point scale, *hsperc* is the percentile in the high school graduating class (defined so that, for example, $hsperc = 5$ means the top 5% of the class), and *sat* is the combined math and verbal scores on the student achievement test.

   (a) Why does it make sense for the coefficient on *hsperc* to be negative?

   (b) What is the predicted college GPA when $hsperc = 20$ and $sat = 1,050$?

   (c) Suppose that two high school graduates, A and B, graduated in the same percentile from high school, but Student A's SAT score was 140 points higher (about one standard deviation in the sample). What is the predicted difference in college GPA for these two students? Is the difference large?

   (d) Holding *hsperc* fixed, what difference in SAT scores leads to a predicted *colga* difference of .50, or one-half of a grade point? Comment on your answer.

2. The following model is a simplified version of the multiple regression model used by Biddle and Hamermesh (1990) to study the trade-off between time spent sleeping and working and to look at other factors affecting sleep:

$$sleep = \beta_0 + \beta_1 totwrk + \beta_2 educ + \beta_3 age + u,$$

where *sleep* and *totwrk* (total work) are measured in minutes per week and *educ* and *age* are measured in years.

   (a) If adults trade off sleep for work, what is the sign of β_1?

   (b) What signs do you think β_2 and β_3 will have?

   (c) Using the data in SLEEP75, the estimated equation is

$$\widehat{sleep} = 3,638.25 - 0.148totwrk - 11.13educ + 2.20age$$
$$n = 706, R^2 = .113.$$

   (d) Discuss the sign and magnitude of the estimated coefficient on educ.

   (e) Would you say *totwrk*, *educ*, and *age* explain much of the variation in sleep? What other factors might affect the time spent sleeping? Are these likely to be correlated with *totwrk*?

3. Suppose that the population model determining y is

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + u_i$$

and this model satisfies Assumptions MLR.1 through MLR.4. However, we estimate the model that omits x_3. Let $\tilde{\beta}_0, \tilde{\beta}_1$, and $\tilde{\beta}_2$ be the OLS estimators from the regression of y on x_1 and x_2. Show that the expected value of $\tilde{\beta}_1$ (given the values of the independent variables in the sample) is

$$E(\tilde{\beta}_1) = \beta_1 + \beta_3 \frac{\sum_{i=1}^{n} \hat{r_{i1}} x_{i3}}{\sum_{i=1}^{n} \hat{r_{i1}}^2}$$

where the $\hat{r_{i1}}$ are the OLS residuals from the regression of x_1 on x_2. [Hints: $\tilde{\beta}_1 = \frac{\sum_{i=1}^{n} \hat{r_{i1}} y_i}{\sum_{i=1}^{n} \hat{r_{i1}}^2}$]

2 Software Problem Set

1. Use the data in **HPRICE1** to estimate the model

$$price = \beta_0 + \beta_1 sqrft + \beta_2 bdrms + u,$$

where *price* is the house price measured in thousands of dollars.

   (a) Write out the results in equation form.

   (b) What is the estimated increase in price for a house with one more bedroom, holding square footage constant?

   (c) What is the estimated increase in price for a house with an additional bedroom that is 140 square feet in size? Compare this to your answer in part (b).

   (d) What percentage of the variation in price is explained by square footage and number of bedrooms?

   (e) The first house in the sample has $srft = 2,438$ and $bdrms = 4$. Find the predicted selling price for this house from the OLS regression line.

   (f) The actual selling price of the first house in the sample was \$300,000 (so $price = 300$). Find the residual for this house. Does it suggest that the buyer underpaid or overpaid for the house?

2. Use the data in **MEAPSINGLE** to study the effects of single-parent households on student math performance. These data are for a subset of schools in southeast Michigan for the year 2000. The socioeconomic variables are obtained at the ZIP code level (where ZIP code is assigned to schools based on their mailing addresses).

   (a) Run the simple regression of *math4* on *petsgle* and report the results in the usual format. Interpret the slope coefficient. Does the effect of single parenthood seem large or small?

   (b) Add the variables *Imedinc* and *free* to the equation. What happens to the coefficient on *petsgle*? Explain what is happening.

   (c) Find the sample correlation between *Imedinc* and *free*. Does it have the sign you expect?

   (d) Does the substantial correlation between *Imedinc* and *free* mean that you should drop one from the regression to better estimate the causal effect of single parenthood on student performance? Explain.