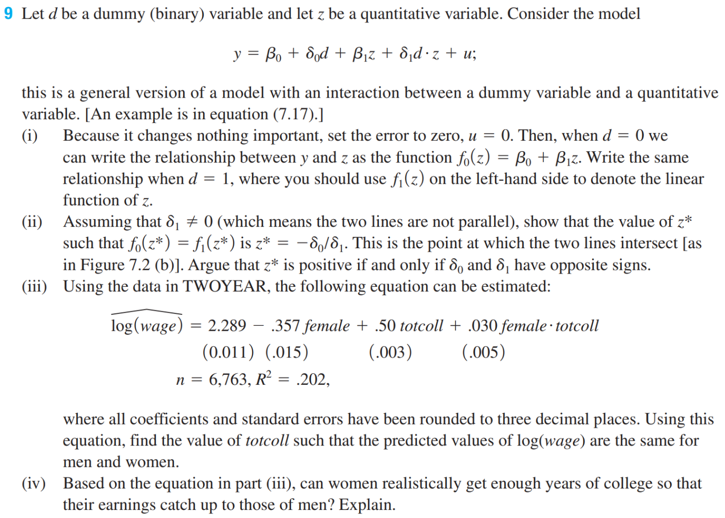
7.9 Problem Set 2

1. Chapter 7, Q 9



9 Let d be a dummy (binary) variable and let z be a quantitative variable. Consider the model

$$y = \beta_0 + \delta_0 d + \beta_1 z + \delta_1 d \cdot z + u;$$

this is a general version of a model with an interaction between a dummy variable and a quantitative variable. [An example is in equation (7.17).]

(i) Because it changes nothing important, set the error to zero, $u = 0$. Then, when $d = 0$ we can write the relationship between y and z as the function $f_0(z) = \beta_0 + \beta_1 z$. Write the same relationship when $d = 1$, where you should use $f_1(z)$ on the left-hand side to denote the linear function of z.

(ii) Assuming that $\delta_1 \neq 0$ (which means the two lines are not parallel), show that the value of z^* such that $f_0(z^*) = f_1(z^*)$ is $z^* = -\delta_0/\delta_1$. This is the point at which the two lines intersect [as in Figure 7.2 (b)]. Argue that z^* is positive if and only if δ_0 and δ_1 have opposite signs.

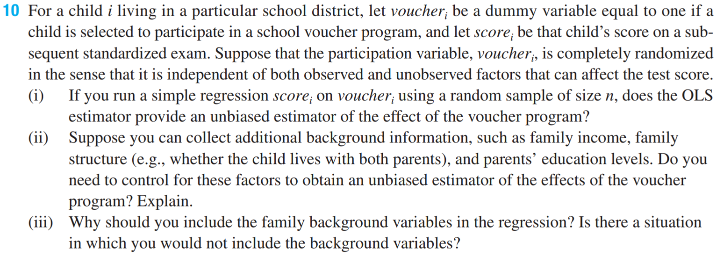
(iii) Using the data in TWOYEAR, the following equation can be estimated:

$$\begin{array}{l} \widehat{\log(wage)} = \underset{(0.011)}{2.289} - \underset{(.015)}{.357}\, female + \underset{(.003)}{.50}\, totcoll + \underset{(.005)}{.030}\, female \cdot totcoll \\ n = 6{,}763,\ R^2 = .202, \end{array}$$

where all coefficients and standard errors have been rounded to three decimal places. Using this equation, find the value of *totcoll* such that the predicted values of log(*wage*) are the same for men and women.

(iv) Based on the equation in part (iii), can women realistically get enough years of college so that their earnings catch up to those of men? Explain.

2. Chapter 7, Q 10



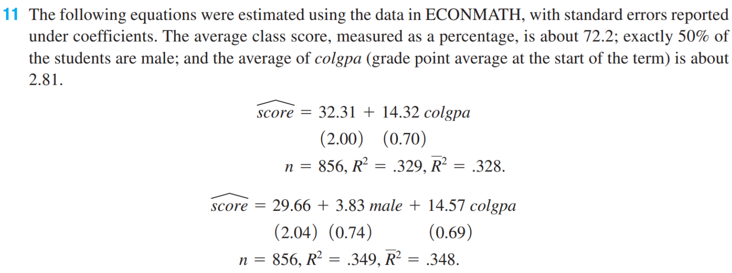
10 For a child i living in a particular school district, let $voucher_i$ be a dummy variable equal to one if a child is selected to participate in a school voucher program, and let $score_i$ be that child's score on a subsequent standardized exam. Suppose that the participation variable, $voucher_i$, is completely randomized in the sense that it is independent of both observed and unobserved factors that can affect the test score.

(i) If you run a simple regression $score_i$ on $voucher_i$ using a random sample of size n, does the OLS estimator provide an unbiased estimator of the effect of the voucher program?

(ii) Suppose you can collect additional background information, such as family income, family structure (e.g., whether the child lives with both parents), and parents' education levels. Do you need to control for these factors to obtain an unbiased estimator of the effects of the voucher program? Explain.

(iii) Why should you include the family background variables in the regression? Is there a situation in which you would not include the background variables?

3. Chapter 7, Q 11



11 The following equations were estimated using the data in ECONMATH, with standard errors reported under coefficients. The average class score, measured as a percentage, is about 72.2; exactly 50% of the students are male; and the average of *colgpa* (grade point average at the start of the term) is about 2.81.

$$\begin{array}{rcl} \widehat{score} & = & \underset{(2.00)}{32.31} + \underset{(0.70)}{14.32}\, colgpa \\ n = 856, & R^2 = .329, & \bar{R}^2 = .328. \end{array}$$

$$\begin{array}{rcl} \widehat{score} & = & \underset{(2.04)}{29.66} + \underset{(0.74)}{3.83}\, male + \underset{(0.69)}{14.57}\, colgpa \\ n = 856, & R^2 = .349, & \bar{R}^2 = .348. \end{array}$$

$$\widehat{score} = \underset{(2.86)}{30.36} + \underset{(3.96)}{2.47}\, male + \underset{(0.98)}{14.33}\, colgpa + \underset{(1.383)}{0.479}\, male \cdot colgpa$$

$$n = 856,\ R^2 = .349,\ \overline{R}^2 = .347.$$

$$\widehat{score} = \underset{(2.86)}{30.36} + \underset{(0.74)}{3.82}\, male + \underset{(0.98)}{14.33}\, colgpa + \underset{(1.383)}{0.479}\, male \cdot (colgpa - 2.81)$$

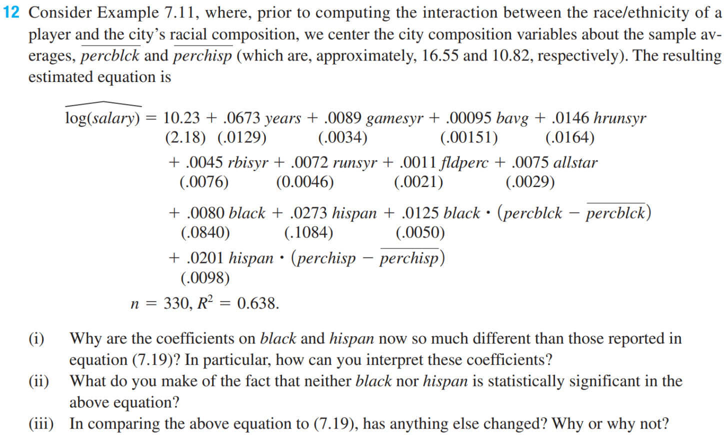
$$n = 856,\ R^2 = .349,\ \overline{R}^2 = .347.$$

(i) Interpret the coefficient on *male* in the second equation and construct a 95% confidence interval for β_{male}. Does the confidence interval exclude zero?

(ii) In the second equation, why is the estimate on *male* so imprecise? Should we now conclude that there are no gender differences in *score* after controlling for *colgpa*? [*Hint*: You might want to compute an *F* statistic for the null hypothesis that there is no gender difference in the model with the interaction.]

(iii) Compared with the third equation, why is the coefficient on *male* in the last equation so much closer to that in the second equation and just as precisely estimated?

4. Chapter 7, Q 12



12 Consider Example 7.11, where, prior to computing the interaction between the race/ethnicity of a player and the city's racial composition, we center the city composition variables about the sample averages, $\overline{percblck}$ and $\overline{perchisp}$ (which are, approximately, 16.55 and 10.82, respectively). The resulting estimated equation is

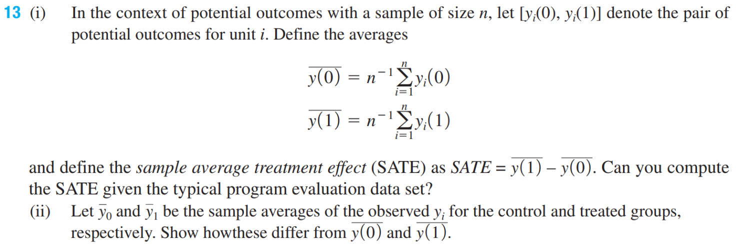
$$\begin{aligned}
\widehat{\log(salary)} = {} & \underset{(2.18)}{10.23} + \underset{(.0129)}{.0673}\, years + \underset{(.0034)}{.0089}\, gamesyr + \underset{(.00151)}{.00095}\, bavg + \underset{(.0164)}{.0146}\, hrunsyr \\
& + \underset{(.0076)}{.0045}\, rbisyr + \underset{(0.0046)}{.0072}\, runsyr + \underset{(.0021)}{.0011}\, fldperc + \underset{(.0029)}{.0075}\, allstar \\
& + \underset{(.0840)}{.0080}\, black + \underset{(.1084)}{.0273}\, hispan + \underset{(.0050)}{.0125}\, black \cdot (percblck - \overline{percblck}) \\
& + \underset{(.0098)}{.0201}\, hispan \cdot (perchisp - \overline{perchisp}) \\
& n = 330,\ R^2 = 0.638.
\end{aligned}$$

(i) Why are the coefficients on *black* and *hispan* now so much different than those reported in equation (7.19)? In particular, how can you interpret these coefficients?

(ii) What do you make of the fact that neither *black* nor *hispan* is statistically significant in the above equation?

(iii) In comparing the above equation to (7.19), has anything else changed? Why or why not?

5. Chapter 7, Q 13



13 (i) In the context of potential outcomes with a sample of size n, let $[y_i(0), y_i(1)]$ denote the pair of potential outcomes for unit i. Define the averages

$$\overline{y(0)} = n^{-1}\sum_{i=1}^{n} y_i(0)$$

$$\overline{y(1)} = n^{-1}\sum_{i=1}^{n} y_i(1)$$

and define the *sample average treatment effect* (SATE) as $SATE = \overline{y(1)} - \overline{y(0)}$. Can you compute the SATE given the typical program evaluation data set?

(ii) Let $\bar{y}_0$ and $\bar{y}_1$ be the sample averages of the observed y_i for the control and treated groups, respectively. Show howthese differ from $\overline{y(0)}$ and $\overline{y(1)}$.

After the 3rd problem set, I will assign presentations for these problem sets and discuss solutions together.