Chapter Review

Problems

1. The following equation was estimated using the data in CEOSAL1:

$$\widehat{\log(salary)} = 4.322 + .276\log(sales) + .0215\,roe - .00008\,roe^2$$
$$(.324)\ \ (.033)\qquad\qquad (.0129)\qquad (.00026)$$
$$n = 209,\ R^2 = .282.$$

   This equation allows *roe* to have a diminishing effect on $\log(salary)$. Is this generality necessary? Explain why or why not.

2. Let $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k$ be the OLS estimates from the regression of y_i on $x_{i1}, \ldots, x_{ik}$, $i = 1, 2, \ldots, n$. For nonzero constants $c_1, \ldots, c_k$, argue that the OLS intercept and slopes from the regression of $c_0 y_i$ on $c_1 x_{i1}, \ldots, c_k x_{ik}$, $i = 1, 2, \ldots, n$, are given by $\tilde{\beta}_0 = c_0\hat{\beta}_0, \tilde{\beta}_1 = (c_0/c_1)\hat{\beta}_1, \ldots, \hat{\beta}_k = (c_0/c_k)\hat{\beta}_k$. [*Hint*: Use the fact that the $\hat{\beta}_j$ solve the first order conditions in (3.13), and the $\tilde{\beta}_j$ must solve the first order conditions involving the rescaled dependent and independent variables.]

3. Using the data in RDCHEM, the following equation was obtained by OLS:

$$\widehat{rdintens} = 2.613 + .00030\,sales - .0000000070\,sales^2$$
$$(.429)\ \ (.00014)\qquad\quad (.0000000037)$$
$$n = 32,\ R^2 = .1484.$$

   i. At what point does the marginal effect of *sales* on *rdintens* become negative?

   ii. Would you keep the quadratic term in the model? Explain.

   iii. Define *salesbil* as sales measured in billions of dollars: $salesbil = sales/1{,}000$. Rewrite the estimated equation with *salesbil* and $salesbil^2$ as the independent variables. Be sure to report standard errors and the *R*-squared. [*Hint*: Note that $salesbil^2 = sales^2/(1{,}000)^2$.]

   iv. For the purpose of reporting the results, which equation do you prefer?

4. The following model allows the return to education to depend upon the total amount of both parents' education, called *pareduc*:

$$\log(wage) = \beta_0 + \beta_1 educ + \beta_2 educ \cdot pareduc + \beta_3 exper + \beta_4 tenure + u.$$

   i. Show that, in decimal form, the return to another year of education in this model is

$$\Delta \log(wage)/\Delta educ = \beta_1 + \beta_2 pareduc.$$

   What sign do you expect for β_2? Why?

   ii. Using the data in WAGE2, the estimated equation is

$$\widehat{\log(wage)} = \underset{(.13)}{5.65} + \underset{(.010)}{.047}\, educ + \underset{(.00021)}{.00078}\, educ \cdot pareduc + \underset{(.004)}{.019}\, exper + \underset{(.003)}{.010}\, tenure$$
$$n = 722,\ R^2 = .169.$$

   (Only 722 observations contain full information on parents' education.) Interpret the coefficient on the interaction term. It might help to choose two specific values for *pareduc*—for example, $pareduc = 32$ if both parents have a college education, or $pareduc = 24$ if both parents have a high school education—and to compare the estimated return to *educ*.

   iii. When *pareduc* is added as a separate variable to the equation, we get:

$$\widehat{\log(wage)} = \underset{(.38)}{4.94} + \underset{(.027)}{.097}\, educ + \underset{(.017)}{.033}\, pareduc - \underset{(.0012)}{.0016}\, educ \cdot pareduc + \underset{(.004)}{.020}\, exper + \underset{(.003)}{.010}\, tenure$$
$$n = 722,\ R^2 = .174.$$

   Does the estimated return to education now depend positively on parent education? Test the null hypothesis that the return to education does not depend on parent education.

5. In Example 4.2, where the percentage of students receiving a passing score on a tenth-grade math exam (*math*10) is the dependent variable, does it make sense to include *sci11*—the percentage of eleventh graders passing a science exam—as an additional explanatory variable?

6. When $atndrte^2$ and *ACT*·*atndrte* are added to the equation estimated in (6.19), the *R*-squared becomes .232. Are these additional terms jointly

significant at the 10% level? Would you include them in the model?

7. The following three equations were estimated using the 1,534 observations in 401K:

$$\widehat{prate} = \underset{(.78)}{80.29} + \underset{(.52)}{5.44}\, mrate + \underset{(.045)}{.269}\, age - \underset{(.00004)}{.00013}\, totemp$$

$$R^2 = .100, \overline{R}^2 = .098.$$

$$\widehat{prate} = \underset{(1.95)}{97.32} + \underset{(0.51)}{5.02}\, mrate + \underset{(.044)}{.314}\, age - \underset{(.28)}{2.66} \log(totemp)$$

$$R^2 = .144, \overline{R}^2 = .142.$$

$$\widehat{prate} = \underset{(.78)}{80.62} + \underset{(.52)}{5.34}\, mrate + \underset{(.045)}{.290}\, age - \underset{(.00009)}{.00043}\, totemp + \underset{(.00000000010)}{.0000000039}\, totemp^2$$

$$R^2 = .108, \overline{R}^2 = .106.$$

Which of these three models do you prefer? Why?

8. Suppose we want to estimate the effects of alcohol consumption (*alcohol*) on college grade point average (*colGPA*). In addition to collecting information on grade point averages and alcohol usage, we also obtain attendance information (say, percentage of lectures attended, called *attend*). A standardized test score (say, *SAT*) and high school GPA (*hsGPA*) are also available.

   i. Should we include *attend* along with *alcohol* as explanatory variables in a multiple regression model? (Think about how you would interpret β_{alcohol}.)

   ii. Should *SAT* and *hsGPA* be included as explanatory variables? Explain.

9. If we start with (6.38) under the CLM assumptions, assume large *n*, and ignore the estimation error in the $\widehat{\beta}_j$, a 95% prediction interval for y^0 is $[\exp(-1.96\widehat{\sigma})\exp(\widehat{logy^0}), \exp(1.96\widehat{\sigma})\exp(\widehat{\log y^0})]$. The point prediction for y^0 is $\hat{y}^0 = \exp(\widehat{\sigma}_2)\exp(\widehat{logy^0})$.

   i. For what values of $\widehat{\sigma}$ will the point prediction be in the 95% prediction interval? Does this condition seem likely to hold in most applications?

   ii. Verify that the condition from part (i) is satisfied in the CEO salary example.

10. The following two equations were estimated using the data in MEAPSINGLE. The key explanatory variable is *lexppp*, the log of expenditures per student at the school level.

$$\widehat{math4} = 24.49 + 9.01\,lexppp - .422\,free - .752\,lmedinc - .274\,pctsgle$$
$$(59.24)\ (4.04)\qquad (.071)\qquad (5.358)\qquad (.161)$$
$$n = 229, R^2 = .472, \overline{R}^2 = .462.$$

$$\widehat{math4} = 149.38 + 1.93\,lexppp - .060\,free - 10.78\,lmedinc - .397\,pctsgle + .667\,read4$$
$$(41.70)\ (2.82)\qquad (.054)\qquad (3.76)\qquad (.111)\qquad (.042)$$
$$n = 229, R^2 = .749, \overline{R}^2 = .743.$$

i. If you are a policy maker trying to estimate the causal effect of per-student spending on math test performance, explain why the first equation is more relevant than the second. What is the estimated effect of a 10% increase in expenditures per student?

ii. Does adding *read4* to the regression have strange effects on coefficients and statistical significance other than β_{lexppp}?

iii. How would you explain to someone with only basic knowledge of regression why, in this case, you prefer the equation with the smaller adjusted *R*-squared?