CHAPTER 6

SOLUTIONS TO PROBLEMS

6.1 This would make little sense. Performances on math and science exams are measures of outputs of the educational process, and we would like to know how various educational inputs and school characteristics affect math and science scores. For example, if the staff-to-pupil ratio has an effect on both exam scores, why would we want to hold performance on the science test fixed while studying the effects of *staff* on the math pass rate? This would be an example of controlling for too many factors in a regression equation. The variable *scill* could be a dependent variable in an identical regression equation.

6.2 (i) Because $\exp(-1.96\hat{\sigma}) < 1$ and $\exp(\hat{\sigma}^2/2) > 1$, the point prediction is always above the lower bound. The only issue is whether the point prediction is below the upper bound. This is the case when $\exp(\hat{\sigma}^2/2) \le \exp(1.96\hat{\sigma})$ or, taking logs, $\hat{\sigma}^2/2 \le 1.96\hat{\sigma}$, or $\hat{\sigma} \le 2(1.96) = 3.92$. Therefore, the point prediction is in the approximate 95% prediction interval for $\hat{\sigma} \le 3.92$. Because $\hat{\sigma}$ is the estimated standard deviation in the regression with log(*y*) as the dependent variable, 3.92 is a very large value for the estimated standard deviation of the error, which is on the order of 400 percent. Most of the time, the estimated SER is well below that.

(ii) In the CEO salary regression, $\hat{\sigma} = .505$, which is well below 3.92.

6.5 (i) The turnaround point is given by $\hat{\beta}_1/(2|\hat{\beta}_2|)$, or .0003/(.000000014) $\approx$ 21,428.57; remember, this is sales in millions of dollars.

(ii) Probably. Its *t* statistic is about –1.89, which is significant against the one-sided alternative H_0: $\beta_1 < 0$ at the 5% level (*cv* $\approx$ –1.70 with *df* = 29). In fact, the *p*-value is about .036.

(iii) Because *sales* gets divided by 1,000 to obtain *salesbil*, the corresponding coefficient gets multiplied by 1,000: (1,000)(.00030) = .30. The standard error gets multiplied by the same factor. As stated in the hint, $salesbil^2 = sales/1{,}000{,}000$, and so the coefficient on the quadratic gets multiplied by one million: (1,000,000)(.0000000070) = .0070; its standard error also gets multiplied by one million. Nothing happens to the intercept (because *rdintens* has not been rescaled) or to the R^2:

$$\widehat{rdintens} = \underset{(0.429)}{2.613} + \underset{(.14)}{.30\, salesbil} - \underset{(.0037)}{.0070\, salesbil^2}$$

$n = 32$, $R^2 = .1484$.

(iv) The equation in part (iii) is easier to read because it contains fewer zeros to the right of the decimal. Of course the interpretation of the two equations is identical once the different scales are accounted for.

6.6 The second equation is clearly preferred, as its adjusted *R*-squared is notably larger than that in the other two equations. The second equation contains the same number of estimated parameters as the first, and the one fewer than the third. The second equation is also easier to interpret than the third.

6.9 The generality is not necessary. The *t* statistic on roe^2 is only about –.30, which shows that roe^2 is very statistically insignificant. Plus, having the squared term has only a minor effect on the slope even for large values of *roe*. (The approximate slope is .0215 – .00016 *roe*, and even when *roe* = 25 – about one standard deviation above the average *roe* in the sample – the slope is .211, as compared with .215 at *roe* = 0.)

SOLUTIONS TO COMPUTER EXERCISES

C6.1 (i) The causal (or ceteris paribus) effect of *dist* on *price* means that $\beta_1 \geq 0$: all other relevant factors equal, it is better to have a home farther away from the incinerator. The estimated equation is

$$\widehat{\log(price)} = \underset{(0.65)}{8.05} + \underset{(.066)}{.365}\log(dist)$$

$$n = 142,\ R^2 = .180,\ \bar{R}^2 = .174,$$

which means a 1% increase in distance from the incinerator is associated with a predicted price that is about .37% higher.

(ii) When the variables log(*inst*), log(*area*), log(*land*), *rooms*, *baths*, and *age* are added to the regression, the coefficient on log(*dist*) becomes about .055 (*se* ≈ .058). The effect is much smaller now, and statistically insignificant. This is because we have explicitly controlled for several other factors that determine the quality of a home (such as its size and number of baths) and its location (distance to the interstate). This is consistent with the hypothesis that the incinerator was located near less desirable homes to begin with.

(iii) When $[\log(inst)]^2$ is added to the regression in part (ii), we obtain (with the results only partially reported)

$$\widehat{\log(price)} = \underset{(2.65)}{-3.32} + \underset{(.062)}{.185}\log(dist) + \underset{(0.501)}{2.073}\log(inst) - \underset{(.0282)}{.1193}[\log(inst)]^2 + \ldots$$

$$n = 142,\ R^2 = .778,\ \bar{R}^2 = .764.$$

The coefficient on log(*dist*) is now very statistically significant, with a *t* statistic of about three. The coefficients on log(*inst*) and $[\log(inst)]^2$ are both very statistically significant, each with *t* statistics above four in absolute value. Just adding $[\log(inst)]^2$ has had a very big effect on the coefficient important for policy purposes. This means that distance from the incinerator and distance from the interstate are correlated in some nonlinear way that also affects housing price.

We can find the value of log(*inst*) where the effect on log(*price*) actually becomes negative: $2.073/[2(.1193)] \approx 8.69$. When we exponentiate this we obtain about 5,943 feet from the interstate. Therefore, it is best to have your home away from the interstate for distances less than just over a mile. After that, moving farther away from the interstate lowers predicted house price.

(iv) The coefficient on $[\log(dist)]^2$, when it is added to the model estimated in part (iii), is about -.0365, but its *t* statistic is only about -.33. Therefore, it is not necessary to add this complication.

C6.3 (i) Holding *exper* (and the elements in *u*) fixed, we have

$$\Delta \log(wage) = \beta_1 \Delta educ + \beta_3 (\Delta educ) exper = (\beta_1 + \beta_3 exper) \Delta educ,$$

or

$$\frac{\Delta \log(wage)}{\Delta educ} = (\beta_1 + \beta_3 exper).$$

This is the approximate proportionate change in *wage* given one more year of education.

(ii) H_0: $\beta_3 = 0$. If we think that education and experience interact positively – so that people with more experience are more productive when given another year of education – then $\beta_3 > 0$ is the appropriate alternative.

(iii) The estimated equation is

$$\widehat{\log(wage)} = \underset{(0.24)}{5.95} + \underset{(.0174)}{.0440}\, educ - \underset{(.0200)}{.0215}\, exper + \underset{(.00153)}{.00320}\, educ \cdot exper$$

$$n = 935, \quad R^2 = .135, \quad \overline{R}^2 = .132.$$

The *t* statistic on the interaction term is about 2.13,which gives a *p*-value below .02 against H_1: $\beta_3 > 0$. Therefore, we reject H_0: $\beta_3 = 0$ against H_1: $\beta_3 > 0$ at the 2% level.

(iv) We rewrite the equation as

$$\log(wage) = \beta_0 + \theta_1 educ + \beta_2 exper + \beta_3 educ(exper - 10) + u,$$

and run the regression log(*wage*) on *educ*, *exper*, and *educ*(*exper* – 10). We want the coefficient on *educ*. We obtain $\hat{\theta}_1 \approx .0761$ and $se(\hat{\theta}_1) \approx .0066$. The 95% CI for θ_1 is about .063 to .089.

C6.5 (i) The results of estimating the log-log model (but with *bdrms* in levels) are

$$\widehat{\log(price)} = \underset{(0.65)}{5.61} + \underset{(.038)}{.168}\log(lotsize) + \underset{(.093)}{.700}\log(sqrft) + \underset{(.028)}{.037}\,bdrms$$

$n = 88$, $R^2 = .634$, $\bar{R}^2 = .630$.

(ii) With *lotsize* = 20,000, *sqrft* = 2,500, and *bdrms* = 4, we have

$$\widehat{lprice} = 5.61 + .168 \cdot \log(20{,}000) + .700 \cdot \log(2{,}500) + .037(4) \approx 12.90$$

where we use *lprice* to denote log(*price*). To predict *price*, we use the equation $\hat{price} = \hat{\alpha}_0 \exp(\widehat{lprice})$, where $\hat{\alpha}_0$ is the slope on $\hat{m}_i \equiv \exp(\widehat{lprice})$ from the regression $price_i$ on $\hat{m}_i$, $i = 1,2, \ldots, 88$ (without an intercept). When we do this regression we get $\hat{\alpha}_0 \approx 1.023$. Therefore, for the values of the independent variables given above, $\widehat{price} \approx (1.023)\exp(12.90) \approx \$409{,}519$ (rounded to the nearest dollar). If we forget to multiply by $\hat{\alpha}_0$ the predicted price would be about $400,312.

(iii) When we run the regression with all variables in levels, the *R*-squared is about .672. When we compute the correlation between $price_i$ and the $\hat{m}_i$ from part (ii), we obtain about .859. The square of this, or roughly .738, is the comparable goodness-of-fit measure for the model with log(*price*) as the dependent variable. Therefore, for predicting *price*, the log model is notably better.

C6.7 (i) If we hold all variables except *priGPA* fixed and use the usual approximation $\Delta(priGPA^2) \approx 2(priGPA)\cdot\Delta priGPA$, then we have

$$\begin{aligned}\Delta stndfnl &= \beta_2 \Delta priGPA + \beta_4 \Delta(priGPA^2) + \beta_6(\Delta priGPA)atndrte \\ &\approx (\beta_2 + 2\beta_4 priGPA + \beta_6 atndrte)\Delta priGPA;\end{aligned}$$

dividing by Δ*priGPA* gives the result. In equation (6.19) we have $\hat{\beta}_2 = -1.63$, $\hat{\beta}_4 = .296$, and $\hat{\beta}_6 = .0056$. When *priGPA* = 2.59 and *atndrte* = .82 we have

$$\frac{\widehat{\Delta stndfnl}}{\Delta priGPA} = -1.63 + 2(.296)(2.59) + .0056(.82) \approx -.092.$$

(ii) First, note that $(priGPA - 2.59)^2 = priGPA^2 - 2(2.59)priGPA + (2.59)^2$ and $priGPA(atndrte - .82) = priGPA \cdot atndrte - (.82)priGPA$. So we can write equation 6.18) as

$$\begin{aligned} stndfnl &= \beta_0 + \beta_1 atndrte + \beta_2 priGPA + \beta_3 ACT + \beta_4 (priGPA - 2.59)^2 \\ &\quad + \beta_4 [2(2.59)priGPA] - \beta_4 (2.59)^2 + \beta_5 ACT^2 \\ &\quad + \beta_6 priGPA(atndrte - .82) + \beta_6 (.82) priGPA + u \\ &= [\beta_0 - \beta_4 (2.59)^2] + \beta_1 atndrte \\ &\quad + [\beta_2 + 2\beta_4 (2.59) + \beta_6 (.82)] priGPA + \beta_3 ACT \\ &\quad + \beta_4 (priGPA - 2.59)^2 + \beta_5 ACT^2 + \beta_6 priGPA(atndrte - .82) + u \\ &\equiv \theta_0 + \beta_1 atndrte + \theta_2 priGPA + \beta_3 ACT + \beta_4 (priGPA - 2.59)^2 \\ &\quad + \beta_5 ACT^2 + \beta_6 priGPA(atndrte - .82) + u. \end{aligned}$$

When we run the regression associated with this last model, we obtain $\hat{\theta}_2 \approx$ -.091 [which differs from part (i) by rounding error] and $se(\hat{\theta}_2) \approx .363$. This implies a very small *t* statistic for $\hat{\theta}_2$.

C6.9 (i) The estimated equation is

$$\widehat{points} = \underset{(6.99)}{35.22} + \underset{(.405)}{2.364}\, exper - \underset{(.0235)}{.0770}\, exper^2 - \underset{(.295)}{1.074}\, age - \underset{(.451)}{1.286}\, coll$$

$n = 269, \quad R^2 = .141, \quad \bar{R}^2 = .128.$

(ii) The turnaround point is 2.364/[2(.0770)] ≈ 15.35. So, the increase from 15 to 16 years of experience would actually reduce salary. This is a very high level of experience, and we can essentially ignore this prediction: only two players in the sample of 269 have more than 15 years of experience.

(iii) Many of the most promising players leave college early, or, in some cases, forego college altogether, to play in the NBA. These top players command the highest salaries. It is not more college that hurts salary, but less college is indicative of super-star potential.

(iv) When age^2 is added to the regression from part (i), its coefficient is .0536 (se = .0492). Its *t* statistic is barely above one, so we are justified in dropping it. The coefficient on *age* in the same regression is –3.984 (se = 2.689). Together, these estimates imply a negative, increasing, return to *age*. The turning point is roughly at 74 years old. In any case, the linear function of *age* seems sufficient.

(v) The OLS results are

$$\widehat{\log(wage)} = 6.78 + .078\, points + .218\, exper - .0071\, exper^2 - .048\, age - .040\, coll$$

(.85) (.007) (.050) (.0028) (.035) (.053)

$n = 269$, $R^2 = .488$, $\bar{R}^2 = .478$

(vi) The joint F statistic produced by Stata is about 1.19. With 2 and 263 *df*, this gives a p-value of roughly .31. Therefore, once scoring and years played are controlled for, there is no evidence for wage differentials depending on age or years played in college.

C6.11 (i) The results of the OLS regression are

$$\widehat{ecolbs} = \underset{(0.38)}{1.97} - \underset{(0.59)}{2.93}\, ecoprc + \underset{(0.71)}{3.03}\, regprc$$

$n = 660$, $R^2 = .036$, $\bar{R}^2 = .034$

As predicted by economic theory, the own price effect is negative and the cross price effect is positive. In particular, an increase in *ecoprc* of .10, or 10 cents per pound, reduces the estimated demand for eco-labeled apples by about .29 lbs. A ceteris paribus increase of 10 cents per lb. for regular applies increases the estimated demand for eco-labeled apples by about .30 lbs. These effects, which are essentially the same magnitude but of opposite sign, are fairly large.

(ii) Each price variable is individually statistically significant with t statistics greater than four (in absolute value) in both cases. The p-values are zero to at least three decimal places.

(iii) The fitted values range from a low of about .86 to a high of about 2.09. This is much less variation than *ecoblbs* itself, which ranges from 0 to 42 (although 42 is a bit of an outlier). There are 248 out of 660 observations with *ecolbs* = 0 and these observations are clearly not explained well by the model.

(iv) The R-squared is only about 3.6% (and it does not really matter whether we use the usual or adjusted R-squared). This is a very small explained variation in *ecolbs*. So the two price variables do not do a good job of explaining why $ecolbs_i$ varies across families.

(v) When *faminc*, *hhsize*, *educ*, and *age* are added to the regression, the R-squared only increases to about .040 (and the adjusted R-squared falls from .034 to .031). The p-value for the joint F test (with 4 and 653 *df*) is about .63, which provides no evidence that these additional variables belong in the regression. Evidently, in addition to the two price variables, the factors that explain variation in *ecolbs* (which is, remember, a counterfactual quantity), are not captured by the demographic and economic variables collected in the survey. Almost 97 percent of the variation is due to unobserved "taste" factors.

C6.13 (i) The estimated equation is

$$\widehat{math4} = \underset{(19.96)}{91.93} + \underset{(2.10)}{3.52}\, lexppp - \underset{(0.94)}{5.40}\, lenroll - \underset{(.015)}{.449}\, lunch$$

$$n = 1{,}692,\ R^2 = .3729,\ \bar{R}^2 = .3718$$

The *lenroll* and *lunch* variables are individually significant at the 5% level, regardless of whether we use a one-sided or two-sided test; in fact, their *p*-values are very small. But *lexppp*, with $t = 1.68$, is not significant against a two-sided alternative. Its one-sided *p*-value is about .047, so it is statistically significant at the 5% level against the positive one-sided alternative.

(ii) The range of fitted values is from about 42.41 to 92.67, which is much narrower than the rage of actual math pass rates in the sample, which is from zero to 100.

(iii) The largest residual is about 51.42, and it belongs to building code 1141. This residual is the difference between the actual pass rate and our best prediction of the pass rate, given the values of spending, enrollment, and the free lunch variable. If we think that per pupil spending, enrollment, and the poverty rate are sufficient controls, the residual can be interpreted as a "value added" for the school. That is, for school 1141, its pass rate is over 51 points higher than we would expect, based on its spending, size, and student poverty.

(iv) The joint *F* statistic, with 3 and 1,685 *df*, is about .52, which gives *p*-value ≈ .67. Therefore, the quadratics are jointly very insignificant, and we would drop them from the model.

(v) The beta coefficients for *lexppp*, *lenroll*, and *lunch* are roughly .035, –.115, and –.613, respectively. Therefore, in standard deviation units, *lunch* has by far the largest effect. The spending variable has the smallest effect.