CHAPTER 4

SOLUTIONS TO PROBLEMS

4.1 (i) and (iii) generally cause the *t* statistics not to have a *t* distribution under H_0. Homoskedasticity is one of the CLM assumptions. An important omitted variable violates Assumption MLR.3. The CLM assumptions contain no mention of the sample correlations among independent variables, except to rule out the case where the correlation is one.

4.3 (i) Holding *profmarg* fixed, $\Delta\widehat{rdintens}$ = .321 Δlog(*sales*) = (.321/100)[$100 \cdot \Delta \log(sales)$] ≈ .00321(%Δ*sales*). Therefore, if %Δ*sales* = 10, $\Delta\widehat{rdintens}$ ≈ .032, or only about 3/100 of a percentage point. For such a large percentage increase in sales, this seems like a practically small effect.

(ii) H_0: $\beta_1 = 0$ versus H_1: $\beta_1 > 0$, where β_1 is the population slope on log(*sales*). The *t* statistic is .321/.216 ≈ 1.486. The 5% critical value for a one-tailed test, with *df* = 32 – 3 = 29, is obtained from Table G.2 as 1.699; so we cannot reject H_0 at the 5% level. But the 10% critical value is 1.311; since the *t* statistic is above this value, we reject H_0 in favor of H_1 at the 10% level.

(iii) Holding sales fixed, one percentage point increase in *profmarg* is estimated to increase R&D expenditures by 100(0.050) ≈ 5%. For a large percentage increase in sales, this is a small effect.

(iv) Not really. Its *t* statistic is only 1.087, which is well below even the 10% critical value for a one-tailed test.

4.5 (i) .412 ± 1.96(.094), or about .228 to .596.

(ii) No, because the value .4 is well inside the 95% CI.

(iii) Yes, because 1 is well outside the 95% CI.

4.7 (i) While the standard error on *hrsemp* has not changed, the magnitude of the coefficient has increased by half. The *t* statistic on *hrsemp* has gone from about –1.26 to –2.21, so now the coefficient is statistically less than zero at the 5% level. (From Table G.2 the 5% critical value with 40 *df* is –1.684. The 1% critical value is –2.423, so the *p*-value is between .01 and .05.)

(ii) If we add and subtract $\beta_2 \log(employ)$ from the right-hand-side and collect terms, we have

$$\begin{aligned}\log(scrap) &= \beta_0 + \beta_1 hrsemp + [\beta_2 \log(\text{sales}) - \beta_2 \log(employ)] \\ &\quad + [\beta_2 \log(employ) + \beta_3 \log(employ)] + u \\ &= \beta_0 + \beta_1 hrsemp + \beta_2 \log(sales/employ) \\ &\quad + (\beta_2 + \beta_3)\log(employ) + u,\end{aligned}$$

where the second equality follows from the fact that log(*sales*/*employ*) = log(*sales*) – log(*employ*). Defining $\theta_3 \equiv \beta_2 + \beta_3$ gives the result.

(iii) No. We are interested in the coefficient on log(*employ*), which has a *t* statistic of .2, which is very small. Therefore, we conclude that the size of the firm, as measured by employees, does not matter, once we control for training *and* sales per employee (in a logarithmic functional form).

(iv) The null hypothesis in the model from part (ii) is H0: $\beta_2 = -1$. The *t* statistic is [–.951 – (–1)]/.37 = (1 – .951)/.37 $\approx$ .132; this is very small, and we fail to reject whether we specify a one- or two-sided alternative.

4.9 (i) With *df* = 706 – 4 = 702, we use the standard normal critical value (*df* = ∞ in Table G.2), which is 1.96 for a two-tailed test at the 5% level. Now t_{educ} = –11.13/5.88 $\approx$ –1.89, so $|t_{educ}|$ = 1.89 < 1.96, and we fail to reject H0: β_{educ} = 0 at the 5% level. Also, $t_{age} \approx$ 1.52, so *age* is also statistically insignificant at the 5% level.

(ii) We need to compute the *R*-squared form of the *F* statistic for joint significance. But *F* = [(.113 – .103)/(1 – .113)](702/2) $\approx$ 3.96. The 5% critical value in the $F_{2,702}$ distribution can be obtained from Table G.3b with denominator *df* = ∞: *cv* = 3.00. Therefore, *educ* and *age* are jointly significant at the 5% level (3.96 > 3.00). In fact, the *p*-value is about .019, and so *educ* and *age* are jointly significant at the 2% level.

(iii) Not really. These variables are jointly significant, but including them only changes the coefficient on *totwrk* from –.151 to –.148.

(iv) The standard *t* and *F* statistics that we used assume homoskedasticity, in addition to the other CLM assumptions. If there is heteroskedasticity in the equation, the tests are no longer valid.

4.11 (i) In columns (2) and (3), the coefficient on *profmarg* is actually negative, although its *t* statistic is only about –1. It appears that, once firm sales and market value have been controlled for, profit margin has no effect on CEO salary.

(ii) We use column (3), which controls for most of the factors affecting salary. The *t* statistic on log(*mktval*) is about 2.05, which is just significant at the 5% level against a two-sided alternative. (We can use the standard normal critical value, 1.96.) So log(*mktval*) is statistically significant. Because the coefficient is an elasticity, a ceteris paribus 10% increase in market value is predicted to increase *salary* by 1%. This is not a huge effect, but it is not negligible, either.

(iii) These variables are individually significant at low significance levels, with $t_{ceoten} \approx 3.11$ and $t_{comten} \approx -2.79$.

(iv) Other factors fixed, another year as CEO with the company increases salary by about 1.71%. On the other hand, another year with the company, but not as CEO, lowers salary by about .92%. This second finding at first seems surprising, but could be related to the "superstar" effect: firms that hire CEOs from outside the company often go after a small pool of highly regarded candidates, and salaries of these people are bid up. More non-CEO years with a company makes it less likely the person was hired as an outside superstar.

4.13 (i) The equation predicts that if the percentage of students not living with two parents increases by 10%, then the percentage of students passing the math test falls by about 8.3 percentage points. When *free* is added as an explanatory variable, the coefficient on *pctsgle* falls to .275.

ii) The *t* statistic is 2.23, which is statistically significant at the 5% level. A one percentage increase in expenditure per pupil, holding the other variables fixed, increases the predicted performance by about 9 percent.

(iii) In all the four models, *pctsgle* has a significant effect on performance of students, and it is highly significant for the first model. The coefficient on *pctsgle* falls when other explanatory variables are added to the first model..

SOLUTIONS TO COMPUTER EXERCISES

C4.1 (i) Holding other factors fixed,

$$\begin{aligned}\Delta voteA &= \beta_1 \Delta \log(expendA) = (\beta_1/100)[100 \cdot \Delta \log(expendA)] \\ &\approx (\beta_1/100)(\%\Delta expendA),\end{aligned}$$

where we use the fact that $100 \cdot \Delta \log(expendA) \approx \%\Delta expendA$. So $\beta_1/100$ is the (ceteris paribus) percentage point change in *voteA* when *expendA* increases by one percent.

(ii) The null hypothesis is H$_0$: $\beta_2 = -\beta_1$, which means a z% increase in expenditure by A and a z% increase in expenditure by B leaves *voteA* unchanged. We can equivalently write H$_0$: $\beta_1 + \beta_2 = 0$.

(iii) The estimated equation (with standard errors in parentheses below estimates) is

$$\begin{array}{lllll}\widehat{voteA} = 45.08 & + \ 6.083 \log(expendA) & - \ 6.615 \log(expendB) & + .152\, prtystrA \\ \quad\quad\quad\quad (3.93) & (0.382) & (0.379) & (0.062)\end{array}$$

$$n = 173, \quad R^2 = .793.$$

The coefficient on log(*expendA*) is very significant (t statistic $\approx$ 15.92), as is the coefficient on log(*expendB*) (t statistic $\approx$ –17.45). The estimates imply that a 10% ceteris paribus increase in spending by candidate A increases the predicted share of the vote going to A by about .61 percentage points. [Recall that, holding other factors fixed, $\Delta\widehat{voteA} \approx (6.083/100)\%\Delta expendA$).] Similarly, a 10% ceteris paribus increase in spending by B reduces $\widehat{voteA}$ by about .66 percentage points. These effects certainly cannot be ignored.

While the coefficients on log(*expendA*) and log(*expendB*) are of similar magnitudes (and opposite in sign, as we expect), we do not have the standard error of $\hat{\beta}_1 + \hat{\beta}_2$, which is what we would need to test the hypothesis from part (ii).

(iv) Write $\theta_1 = \beta_1 + \beta_2$, or $\beta_1 = \theta_1 - \beta_2$. Plugging this into the original equation, and rearranging, gives

$$\widehat{voteA} = \beta_0 + \theta_1 \log(expendA) + \beta_2 [\log(expendB) - \log(expendA)] + \beta_3 prtystrA + u,$$

When we estimate this equation, we obtain $\hat{\theta}_1 \approx -.532$ and se($\hat{\theta}_1$) $\approx$ .533. The t statistic for the hypothesis in part (ii) is –.532/.533 $\approx$ –1. Therefore, we fail to reject H$_0$: $\beta_2 = -\beta_1$.

C4.3 (i) The estimated model is

$$\begin{array}{llll}\widehat{\log(price)} = & 4.77 & + .000379\, sqrft & + .0289\, bdrms \\ & (0.10) & (.000043) & (.0296)\end{array}$$

$$n = 88, \ R^2 = .588.$$

Therefore, $\hat{\theta}_1$ = 150(.000379) + .0289 = .0858, which means that an additional 150 square foot bedroom increases the predicted price by about 8.6%.

(ii) $\beta_2 = \theta_1 - 150\beta_1$, and so

$$\begin{aligned}\log(price) &= \beta_0 + \beta_1 sqrft + (\theta_1 - 150\beta_1) bdrms + u \\ &= \beta_0 + \beta_1 (sqrft - 150\, bdrms) + \theta_1 bdrms + u.\end{aligned}$$

(iii) From part (ii), we run the regression

$$\log(price) \text{ on } (sqrft - 150\, bdrms),\ bdrms,$$

and obtain the standard error on *bdrms*. We already know that $\hat{\theta}_1$ = .0858; now we also get se($\hat{\theta}_1$) = .0268. The 95% confidence interval reported by my software package is .0326 to .1390 (or about 3.3% to 13.9%).

C4.5 (i) If we drop *rbisyr*, the estimated equation becomes

$$\widehat{\log(salary)} = \underset{(0.27)}{11.02} + \underset{(.0121)}{.0677}\, years + \underset{(.0016)}{.0158}\, gamesyr + \underset{(.0011)}{.0014}\, bavg + \underset{(.0072)}{.0359}\, hrunsyr$$

$$n = 353, \quad R^2 = .625.$$

Now *hrunsyr* is very statistically significant (*t* statistic ≈ 4.99), and its coefficient has increased by about two and one-half times.

(ii) The equation with *runsyr*, *fldperc*, and *sbasesyr* added is

$$\widehat{\log(salary)} = \underset{(2.00)}{10.41} + \underset{(.0120)}{.0700}\, years + \underset{(.0027)}{.0079}\, gamesyr + \underset{(.00110)}{.00053}\, bavg + \underset{(.0086)}{.0232}\, hrunsyr + \underset{(.0051)}{.0174}\, runsyr + \underset{(.0020)}{.0010}\, fldperc - \underset{(.0052)}{.0064}\, sbasesyr$$

$$n = 353, \quad R^2 = .639.$$

Of the three additional independent variables, only *runsyr* is statistically significant (*t* statistic = .0174/.0051 ≈ 3.41). The estimate implies that one more run per year, other factors fixed, increases predicted salary by about 1.74%, a substantial increase. The stolen bases variable even has the "wrong" sign with a *t* statistic of about –1.23, while *fldperc* has a *t* statistic of only .5. Most major league baseball players are pretty good fielders; in fact, the smallest *fldperc* is 800 (which means .800). With relatively little variation in *fldperc*, it is perhaps not surprising that its effect is hard to estimate.

(iii) From their *t* statistics, *bavg*, *fldperc*, and *sbasesyr* are individually insignificant. The *F* statistic for their joint significance (with 3 and 345 *df*) is about .69 with *p*-value ≈ .56. Therefore, these variables are jointly very insignificant.

C4.7 (i) The minimum value is 0, the maximum is 99, and the average is about 56.16.

(ii) When *phsrank* is added to (4.26), we get the following:

$$\widehat{\log(wage)} = 1.459 - .0093\, jc + .0755\, totcoll + .0049\, exper + .00030\, phsrank$$

$$\qquad (0.024) \quad (.0070) \quad (.0026) \qquad (.0002) \qquad (.00024)$$

n = 6,763, R^2 = .223.

So *phsrank* has a *t* statistic equal to only 1.25; it is not statistically significant. If we increase *phsrank* by 10, log(*wage*) is predicted to increase by (.0003)10 = .003. This implies a .3% increase in *wage*, which seems a modest increase given a 10 percentage point increase in *phsrank*. (However, the sample standard deviation of *phsrank* is about 24.)

(iii) Adding *phsrank* makes the *t* statistic on *jc* even smaller in absolute value, about 1.33, but the coefficient magnitude is similar to (4.26). Therefore, the basic point remains unchanged; the return to a junior college is estimated to be somewhat smaller, but the difference is not significant at standard significant levels.

(iv) The variable *id* is just a worker identification number, which should be randomly assigned (at least roughly). Therefore, *id* should not be correlated with any variable in the regression equation. It should be insignificant when added to (4.17) or (4.26). In fact, its *t* statistic is very low, about .54, and the two-sided *p*-value is 0.587.

C4.9 (i) The results from the OLS regression, with standard errors in parentheses, are

$$\begin{aligned}\widehat{\log(psoda)} = \; & -1.46 \;+\; .073\,prpblck \;+\; .137\log(income) \;+\; .380\,prppov \\ & \;(0.29) \quad\; (.031) \qquad\qquad (.027) \qquad\qquad\quad (.133)\end{aligned}$$

n = 401, R^2 = .087.

The *p*-value for testing H0: $\beta_1 = 0$ against the two-sided alternative is about .018, so that we reject H0 at the 5% level but not at the 1% level.

(ii) The correlation is about −.84, indicating a strong degree of multicollinearity. Yet each coefficient is very statistically significant: the *t* statistic for $\hat{\beta}_{\log(income)}$ is about 5.1 and that for $\hat{\beta}_{prppov}$ is about 2.86 (two-sided *p*-value = .004).

(iii) The OLS regression results when log(*hseval*) is added are

$$\begin{aligned}\widehat{\log(psoda)} = \; & -.84 \;+\; .098\,prpblck \;-\; .053\log(income) \\ & \;(.29) \quad\; (.029) \qquad\qquad (.038) \\ & \;+\; .052\,prppov \;+\; .121\log(hseval) \\ & \qquad (.134) \qquad\qquad (.018)\end{aligned}$$

n = 401, R^2 = .184.

The coefficient on log(*hseval*) is an elasticity: a one percent increase in housing value, holding the other variables fixed, increases the predicted price by about .12 percent. The two-sided *p*-value is zero to three decimal places.

(iv) Adding log(*hseval*) makes log(*income*) and *prppov* individually insignificant (at even the 15% significance level against a two-sided alternative for log(*income*), and *prppov* is does not have a *t* statistic even close to one in absolute value). Nevertheless, they are jointly significant at the 5% level because the outcome of the $F_{2,396}$ statistic is about 3.52 with *p*-value = .030. All of the control variables – log(*income*), *prppov*, and log(*hseval*) – are highly correlated, so it is not surprising that some are individually insignificant.

(v) Because the regression in (iii) contains the most controls, log(*hseval*) is individually significant, and log(*income*) and *prppov* are jointly significant, (iii) seems the most reliable. It holds fixed three measure of income and affluence. Therefore, a reasonable estimate is that if the proportion of blacks increases by .10, *psoda* is estimated to increase by 1%, other factors held fixed.

C4.11 (i) The estimated equation, with standard errors in parentheses below coefficient estimates, is

$$\begin{array}{lllllllll}\widehat{educ} = & 8.24 & + & .190\ motheduc & + & .109\ fatheduc & + & .401\ abil & + .0506\ abil^2 \\ & (0.29) & & (.028) & & (.020) & & (.030) & (.0083)\end{array}$$

n = 1,230, R^2 = .444.

The null hypothesis of a linear relationship between *educ* and *abil* is $\mathrm{H}_0 : \beta_4 = 0$ and the alternative is that H_0 does not hold. The *t* statistic is about $.0506/.0083 \approx 6.1$, which is a very large value for a *t* statistic. The *p*-value against the two-sided alternative is zero to more than four decimal places.

(ii) We could rewrite the model by defining, say, $\theta_1 = \beta_1 - \beta_2$ and then substituting in $\beta_1 = \theta_1 + \beta_2$, just as we did with the example in Section 4.4. These days, it is easier to use a special command in statistical software. The estimated difference in the coefficients is about .081. The instructor could use the lincom command in Stata to get a *t* statistic of about 1.94 and an associated two-sided *p*-value of about .053. So there is some evidence against the null hypothesis.

(iii) The instructor could use the test command in Stata to test the joint significance of the tuition variables. With 2 and 1,223 degrees of freedom the instructor get an *F* statistic of about .84 with association *p*-value of about .43. Thus, the tuition variables are jointly insignificant at any reasonable significance level.

(iv) Not surprisingly, the correlation between *tuit*17 and *tuit*18 is very high, about .981: there is very little change in tuition over a year that cannot be explained by a common inflation factor. The instructor could generate the variable *avgtuit* = (*tuit*17 + *tuit*18)/2, and then added it to the regression from part (i). The coefficient on *avgtuit* is about .016 with $t = 1.29$. This certainly helps with statistical significance but the two-sided *p*-value is still only about .20.

(v) The positive coefficient on *avgtuit* does not make a lot of sense if we think that, all other things fixed, higher tuition makes it less likely that people go to college. But we are only controlling for parents' levels of education and a measure of ability. It could be that higher tuition indicates higher quality of the state colleges. Or, it could be that tuition is higher in states with higher average incomes, and higher family incomes lead to higher education. In any case, the statistical link is not very strong.