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.

4.2 (i) H_0: $\beta_3 = 0$. H_1: $\beta_3 > 0$.

(ii) The proportionate effect on *salary* is .00024(50) = .012. To obtain the percentage effect, we multiply this by 100: 1.2%. Therefore, a 50 point ceteris paribus increase in *ros* is predicted to increase salary by only 1.2%. Practically speaking this is a very small effect for such a large change in *ros*.

(iii) The 10% critical value for a one-tailed test, using $df = \infty$, is obtained from Table G.2 as 1.282. The *t* statistic on *ros* is .00024/.00054 $\approx$ .44, which is well below the critical value. Therefore, we fail to reject H_0 at the 10% significance level.

(iv) Based on this sample, the estimated *ros* coefficient appears to be different from zero only because of sampling variation. On the other hand, including *ros* may not be causing any harm; it depends on how correlated it is with the other independent variables (although these are very significant even with *ros* in the equation).

4.3 (i) Holding *profmarg* fixed, $\Delta \widehat{rdintens}$ = .321 $\Delta\log(sales)$ = (.321/100)$[100 \cdot \Delta \log(sales)]$ $\approx$ .00321(%Δ*sales*). Therefore, if %Δ*sales* = 10, $\Delta \widehat{rdintens}$ $\approx$ .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 $\approx$ 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) Not really. Its *t* statistic is only 1.087, which is well below even the 10% critical value for a one-tailed test.

4.4 (i) H_0: $\beta_3 = 0$. H_1: $\beta_3 \neq 0$.

(ii) Other things equal, a larger population increases the demand for rental housing, which should increase rents. The demand for overall housing is higher when average income is higher, pushing up the cost of housing, including rental rates.

(iii) The coefficient on log(*pop*) is an elasticity. A correct statement is that "a 10% increase in population increases *rent* by .066(10) = .66%."

(iv) With $df = 64 - 4 = 60$, the 1% critical value for a two-tailed test is 2.660. The *t* statistic is about 3.29, which is well above the critical value. So β_3 is statistically different from zero at the 1% level.

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.

4.6 (i) With $df = n - 2 = 86$, we obtain the 5% critical value from Table G.2 with $df = 90$. Because each test is two-tailed, the critical value is 1.987. The *t* statistic for H_0: $\beta_0 = 0$ is about -.89, which is much less than 1.987 in absolute value. Therefore, we fail to reject $\beta_0 = 0$. The *t* statistic for H_0: $\beta_1 = 1$ is $(.976 - 1)/.049 \approx -.49$, which is even less significant. (Remember, we reject H_0 in favor of H_1 in this case only if $|t| > 1.987$.)

(ii) We use the SSR form of the *F* statistic. We are testing $q = 2$ restrictions and the *df* in the unrestricted model is 86. We are given $SSR_r = 209{,}448.99$ and $SSR_{ur} = 165{,}644.51$. Therefore,

$$F = \frac{(209{,}448.99 - 165{,}644.51)}{165{,}644.51} \cdot \left(\frac{86}{2}\right) \approx 11.37,$$

which is a strong rejection of H_0: from Table G.3c, the 1% critical value with 2 and 90 *df* is 4.85.

(iii) We use the *R*-squared form of the *F* statistic. We are testing $q = 3$ restrictions and there are $88 - 5 = 83$ *df* in the unrestricted model. The *F* statistic is $[(.829 - .820)/(1 - .829)](83/3) \approx 1.46$. The 10% critical value (again using 90 denominator *df* in Table G.3a) is 2.15, so we fail to reject H_0 at even the 10% level. In fact, the *p*-value is about .23.

(iv) If heteroskedasticity were present, Assumption MLR.5 would be violated, and the *F* statistic would not have an *F* distribution under the null hypothesis. Therefore, comparing the *F* statistic against the usual critical values, or obtaining the *p*-value from the *F* distribution, would not be especially meaningful.

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 –1.47 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 H_0: $\beta_2 = -1$. 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.

4.8 (i) We use Property VAR.3 from Appendix B: $\text{Var}(\hat{\beta}_1 - 3\hat{\beta}_2) = \text{Var}(\hat{\beta}_1) + 9\,\text{Var}(\hat{\beta}_2) - 6\,\text{Cov}(\hat{\beta}_1, \hat{\beta}_2)$.

(ii) $t = (\hat{\beta}_1 - 3\hat{\beta}_2 - 1)/\text{se}(\hat{\beta}_1 - 3\hat{\beta}_2)$, so we need the standard error of $\hat{\beta}_1 - 3\hat{\beta}_2$.

(iii) Because $\theta_1 = \beta_1 - 3\beta_2$, we can write $\beta_1 = \theta_1 + 3\beta_2$. Plugging this into the population model gives

$$\begin{aligned} y &= \beta_0 + (\theta_1 + 3\beta_2)x_1 + \beta_2 x_2 + \beta_3 x_3 + u \\ &= \beta_0 + \theta_1 x_1 + \beta_2(3x_1 + x_2) + \beta_3 x_3 + u. \end{aligned}$$

This last equation is what we would estimate by regressing *y* on x_1, $3x_1 + x_2$, and x_3. The coefficient and standard error on x_1 are what we want.

4.9 (i) With $df = 706 - 4 = 702$, we use the standard normal critical value ($df = \infty$ 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 H_0: $\beta_{educ} = 0$ 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) ≈ 3.96. The 5% critical value in the $F_{2,702}$ distribution can be obtained from Table G.3b with denominator $df = \infty$: 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.

4.10 (i) We need to compute the F statistic for the overall significance of the regression with n = 142 and k = 4: F = [.0395/(1 – .0395)](137/4) ≈ 1.41. The 5% critical value with 4 numerator df and using 120 for the numerator df, is 2.45, which is well above the value of F. Therefore, we fail to reject H_0: $\beta_1 = \beta_2 = \beta_3 = \beta_4 = 0$ at the 10% level. No explanatory variable is individually significant at the 5% level. The largest absolute t statistic is on *dkr*, $t_{dkr} \approx$ 1.60, which is not significant at the 5% level against a two-sided alternative.

(ii) The F statistic (with the same df) is now [.0330/(1 – .0330)](137/4) ≈ 1.17, which is even lower than in part (i). None of the t statistics is significant at a reasonable level.

(iii) It seems very weak. There are no significant t statistics at the 5% level (against a two-sided alternative), and the F statistics are insignificant in both cases. Plus, less than 4% of the variation in *return* is explained by the independent variables.

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 the most 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 $t_{comten} \approx -2.79$. 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.

SOLUTIONS TO COMPUTER EXERCISES

4.12 (i) Holding other factors fixed,

$$\Delta voteA = \beta_1 \Delta \log(expendA) = (\beta_1/100)[100 \cdot \Delta \log(expendA)]$$
$$\approx (\beta_1/100)(\%\Delta expendA),$$

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

$$\widehat{voteA} = \underset{(3.93)}{45.08} + \underset{(0.382)}{6.083} \log(expendA) - \underset{(0.379)}{6.615} \log(expendB) + \underset{(.062)}{.152}\, prtystrA$$

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

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: $\beta_2 = -\beta_1$.

4.13 (i) In the model

$$\log(salary) = \beta_0 + \beta_1 LSAT + \beta_2 GPA + \beta_3 \log(libvol) + \beta_4 \log(cost) + \beta_5 rank + u,$$

the hypothesis that *rank* has no effect on log(*salary*) is H_0: $\beta_5 = 0$. The estimated equation (now with standard errors) is

$$\widehat{\log(salary)} = \underset{(0.53)}{8.34} + \underset{(.0040)}{.0047}\,LSAT + \underset{(.090)}{.248}\,GPA + \underset{(.033)}{.095}\,\log(libvol)$$

$$+ \underset{(.032)}{.038}\,\log(cost) \quad - \underset{(.0003)}{.0033}\,rank$$

$n = 136, \quad R^2 = .842.$

The *t* statistic on *rank* is –11, which is very significant. If *rank* decreases by 10 (which is a move up for a law school), median starting salary is predicted to increase by about 3.3%.

(ii) *LSAT* is not statistically significant (*t* statistic $\approx$ 1.18) but *GPA* is very significance (*t* statistic $\approx$ 2.76). The test for joint significance is moot given that *GPA* is so significant, but for completeness the *F* statistic is about 9.95 (with 2 and 130 *df*) and *p*-value $\approx$ .0001.

(iii) When we add *clsize* and *faculty* to the regression we lose five observations. The test of their joint significant (with 2 and 131 – 8 = 123 *df*) gives $F \approx .95$ and *p*-value $\approx$ .39. So these two variables are not jointly significant unless we use a very large significance level.

(iv) If we want to just determine the effect of numerical ranking on starting law school salaries, we should control for other factors that affect salaries and rankings. The idea is that there is some randomness in rankings, or the rankings might depend partly on frivolous factors that do not affect quality of the students. LSAT scores and GPA are perhaps good controls for student quality. However, if there are differences in gender and racial composition across schools, and systematic gender and race differences in salaries, we could also control for these. However, it is unclear why these would be correlated with *rank*. Faculty quality, as perhaps measured by publication records, could be included. Such things do enter rankings of law schools.

4.14 (i) The estimated model is

$$\widehat{\log(price)} = \underset{(0.10)}{11.67} + \underset{(.000043)}{.000379}\,sqrft + \underset{(.0296)}{.0289}\,bdrms$$

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

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

$$\log(price) = \beta_0 + \beta_1 sqrft + (\theta_1 - 150\,\beta_1) bdrms + u$$

$$= \beta_0 + \beta_1 (sqrft - 150\ bdrms) + \theta_1 bdrms + u.$$

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

log(*price*) on (*sqrft* – 150 *bdrms*) and *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%).

4.15 The *R*-squared from the regression *bwght* on *cigs*, *parity*, and *faminc*, using all 1,388 observations, is about .0348. This means that, if we mistakenly use this in place of .0364, which is the *R*-squared using the same 1,191 observations available in the unrestricted regression, we would obtain F = [(.0387 – .0348)/(1 – .0387)](1,185/2) ≈ 2.40, which yields *p*-value ≈ .091 in an *F* distribution with 2 and 1,1185 *df*. This is significant at the 10% level, but it is incorrect. The correct *F* statistic was computed as 1.42 in Example 4.9, with *p*-value ≈ .242.

4.16 (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 $\approx$ 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 $\approx$ .56. Therefore, these variables are jointly very insignificant.

4.17 (i) In the model

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

the null hypothesis of interest is H_0: $\beta_2 = \beta_3$.

(ii) Let $\theta_2 = \beta_2 - \beta_3$. Then we can estimate the equation

$$\log(wage) = \beta_0 + \beta_1 educ + \theta_2 exper + \beta_3 (exper + tenure) + u$$

to obtain the 95% CI for θ_2. This turns out to be about .0020 ± 1.96(.0047), or about -.0072 to .0112. Because zero is in this CI, θ_2 is not statistically different from zero at the 5% level, and we fail to reject H_0: $\beta_2 = \beta_3$ at the 5% level.

4.18 (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:

$$\widehat{\log(wage)} = \underset{(0.024)}{1.459} - \underset{(.0070)}{.0093}\, jc + \underset{(.0026)}{.0755}\, totcoll + \underset{(.0002)}{.0049}\, exper + \underset{(.00024)}{.00030}\, phsrank$$

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 base point remains unchanged: the return to a junior college is estimated to be somewhat smaller, but the difference is not significant and 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 about .54.

4.19 (i) There are 2,017 single people in the sample of 9,275.

(ii) The estimated equation is

$$\widehat{nettfa} = -43.04 + .799\, inc + .843\, age$$
$$(4.08) \quad (.060) \quad (.092)$$

$$n = 2{,}017, \; R^2 = .119.$$

The coefficient on *inc* indicates that one more dollar in income (holding *age* fixed) is reflected in about 80 more cents in predicted *nettfa*; no surprise there. The coefficient on *age* means that, holding income fixed, if a person gets another year older, his/her *nettfa* is predicted to increase by about \$843. (Remember, *nettfa* is in thousands of dollars.) Again, this is not surprising.

(iii) The intercept is not very interesting, as it gives the predicted *nettfa* for $inc = 0$ and $age = 0$. Clearly, there is no one with even close to these values in the relevant population.

(iv) The *t* statistic is $(.843 - 1)/.092 \approx -1.71$. Against the one-sided alternative H_1: $\beta_2 < 1$, the p-value is about .044. Therefore, we can reject H_0: $\beta_2 = 1$ at the 5% significance level (against the one-sided alternative).

(v) The slope coefficient on *inc* in the simple regression is about .821, which is not very different from the .799 obtained in part (ii). As it turns out, the correlation between *inc* and *age* in the sample of single people is only about .039, which helps explain why the simple and multiple regression estimates are not very different; refer back to page 79 of the text.