CHAPTER 3

SOLUTIONS TO PROBLEMS

3.1 (i) *hsperc* is defined so that the smaller it is, the lower the student's standing in high school. Everything else equal, the worse the student's standing in high school, the lower his/her expected college GPA.

(ii) Just plug these values into the equation

$$\widehat{colgpa} = 1.392 - .0135(20) + .00148(1050) = 2.676.$$

(iii) The difference between A and B is simply 140 times the coefficient on *sat*, because *hsperc* is the same for both students. So A is predicted to have a score .00148(140) $\approx$ .207 higher.

(iv) With *hsperc* fixed, $\Delta\widehat{colgpa} = .00148\Delta sat$. Now, we want to find Δsat such that $\Delta\widehat{colgpa} = .5$, so $.5 = .00148(\Delta sat)$ or $\Delta sat = .5/(.00148) \approx 338$. Perhaps not surprisingly, a large ceteris paribus difference in SAT score – almost two and one-half standard deviations – is needed to obtain a predicted difference in college GPA or a half a point.

3.3 (i) If adults trade off sleep for work, more work implies less sleep (other things equal), so $\beta_1 < 0$.

(ii) The signs of β_2 and β_3 are not obvious, at least to me. One could argue that more educated people like to get more out of life, and so, other things equal, they sleep less ($\beta_2 < 0$). The relationship between sleeping and age is more complicated than this model suggests, and economists are not in the best position to judge such things.

(iii) Since *totwrk* is in minutes, we must convert five hours into minutes: $\Delta totwrk$ = 5(60) = 300. Then *sleep* is predicted to fall by .148(300) = 44.4 minutes. For a week, 45 minutes less sleep is not an overwhelming change.

(iv) More education implies less predicted time sleeping, but the effect is quite small. If we assume the difference between college and high school is four years, the college graduate sleeps about 45 minutes less per week, other things equal.

(v) Not surprisingly, the three explanatory variables explain only about 11.3% of the variation in *sleep*. One important factor in the error term is general health. Another is marital status and whether the person has children. Health (however we measure that), marital status, and number and ages of children would generally be correlated with *totwrk*. (For example, less healthy people would tend to work less.)

3.5 (i) No. By definition, *study* + *sleep* + *work* + *leisure* = 168. Therefore, if we change *study*, we must change at least one of the other categories so that the sum is still 168.

(ii) From part (i), we can write, say, *study* as a perfect linear function of the other independent variables: *study* = 168 – *sleep* – *work* – *leisure*. This holds for every observation, so MLR.3 is violated.

(iii) Simply drop one of the independent variables, say *leisure*:

$$GPA = \beta_0 + \beta_1 study + \beta_2 sleep + \beta_3 work + u.$$

Now, for example, β_1 is interpreted as the change in *GPA* when *study* increases by one hour, where *sleep*, *work*, and *u* are all held fixed. If we are holding *sleep* and *work* fixed but increasing *study* by one hour, then we must be reducing *leisure* by one hour. The other slope parameters have a similar interpretation.

3.7 Only (ii), omitting an important variable, can cause bias, and this is true only when the omitted variable is correlated with the included explanatory variables. The homoskedasticity assumption, MLR.5, played no role in showing that the OLS estimators are unbiased. (Homoskedasticity was used to obtain the usual variance formulas for the $\hat{\beta}_j$.) Further, the degree of collinearity between the explanatory variables in the sample, even if it is reflected in a correlation as high as .95, does not affect the Gauss-Markov assumptions. Only if there is a *perfect* linear relationship among two or more explanatory variables is MLR.3 violated.

3.9 (i) $\beta_1 < 0$ because more pollution can be expected to lower housing values; note that β_1 is the elasticity of *price* with respect to *nox*. β_2 is probably positive because *rooms* roughly measures the size of a house. (However, it does not allow us to distinguish homes where each room is large from homes where each room is small.)

(ii) If we assume that *rooms* increases with quality of the home, then log(*nox*) and *rooms* are negatively correlated when poorer neighborhoods have more pollution, something that is often true. We can use Table 3.2 to determine the direction of the bias. If $\beta_2 > 0$ and Corr(x_1,x_2) < 0, the simple regression estimator $\tilde{\beta}_1$ has a downward bias. But because $\beta_1 < 0$, this means that the simple regression, on average, overstates the importance of pollution. [E($\tilde{\beta}_1$) is more negative than β_1.]

(iii) This is what we expect from the typical sample based on our analysis in part (ii). The simple regression estimate, –1.043, is more negative (larger in magnitude) than the multiple regression estimate, –.718. As those estimates are only for one sample, we can never know which is closer to β_1. But if this is a "typical" sample, β_1 is closer to –.718.

3.11 From equation (3.22), we have

$$\tilde{\beta}_1 = \frac{\sum_{i=1}^{n} \hat{r}_{i1} y_i}{\sum_{i=1}^{n} \hat{r}_{i1}^2},$$

where the $\hat{r}_{i1}$ are defined in the problem. As usual, we must plug in the true model for y_i:

$$\tilde{\beta}_1 = \frac{\sum_{i=1}^{n} \hat{r}_{i1}(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + u_i}{\sum_{i=1}^{n} \hat{r}_{i1}^2}.$$

The numerator of this expression simplifies because $\sum_{i=1}^{n} \hat{r}_{i1}$ = 0, $\sum_{i=1}^{n} \hat{r}_{i1} x_{i2}$ = 0, and $\sum_{i=1}^{n} \hat{r}_{i1} x_{i1}$ = $\sum_{i=1}^{n} \hat{r}_{i1}^2$. These all follow from the fact that the $\hat{r}_{i1}$ are the residuals from the regression of x_{i1} on x_{i2}: the $\hat{r}_{i1}$ have zero sample average and are uncorrelated in sample with x_{i2}. So the numerator of $\tilde{\beta}_1$ can be expressed as

$$\beta_1 \sum_{i=1}^{n} \hat{r}_{i1}^2 + \beta_3 \sum_{i=1}^{n} \hat{r}_{i1} x_{i3} + \sum_{i=1}^{n} \hat{r}_{i1} u_i.$$

Putting these back over the denominator gives

$$\tilde{\beta}_1 = \beta_1 + \beta_3 \frac{\sum_{i=1}^{n} \hat{r}_{i1} x_{i3}}{\sum_{i=1}^{n} \hat{r}_{i1}^2} + \frac{\sum_{i=1}^{n} \hat{r}_1 u_i}{\sum_{i=1}^{n} \hat{r}_{i1}^2}.$$

Conditional on all sample values on x_1, x_2, and x_3, only the last term is random due to its dependence on u_i. But E(u_i) = 0, and so

$$\mathrm{E}(\tilde{\beta}_1) = \beta_1 + \beta_3 \frac{\sum_{i=1}^{n} \hat{r}_{i1} x_{i3}}{\sum_{i=1}^{n} \hat{r}_{i1}^2},$$

which is what we wanted to show. Notice that the term multiplying β_3 is the regression coefficient from the simple regression of x_{i3} on $\hat{r}_{i1}$.

3.13 (i) For notational simplicity, define $s_{zx} = \sum_{i=1}^{n}(z_i - \overline{z})x_i$; this is not quite the sample covariance between z and x because we do not divide by n – 1, but we are only using it to simplify notation. Then we can write $\tilde{\beta}_1$ as

$$\tilde{\beta}_1 = \frac{\sum_{i=1}^{n}(z_i - \overline{z})y_i}{s_{zx}}.$$

This is clearly a linear function of the y_i: take the weights to be $w_i = (z_i - \overline{z})/s_{zx}$. To show unbiasedness, as usual we plug $y_i = \beta_0 + \beta_1 x_i + u_i$ into this equation, and simplify:

$$\begin{aligned}\tilde{\beta}_1 &= \frac{\sum_{i=1}^{n}(z_i - \overline{z})(\beta_0 + \beta_1 x_i + u_i)}{s_{zx}} \\ &= \frac{\beta_0\sum_{i=1}^{n}(z_i - \overline{z}) + \beta_1 s_{zx} + \sum_{i=1}^{n}(z_i - \overline{z})u_i}{s_{zx}} \\ &= \beta_1 + \frac{\sum_{i=1}^{n}(z_i - \overline{z})u_i}{s_{zx}}\end{aligned}$$

where we use the fact that $\sum_{i=1}^{n}(z_i - \overline{z}) = 0$ always. Now s_{zx} is a function of the z_i and x_i and the expected value of each u_i is zero conditional on all z_i and x_i in the sample. Therefore, conditional on these values,

$$\mathrm{E}(\tilde{\beta}_1) = \beta_1 + \frac{\sum_{i=1}^{n}(z_i - \overline{z})\mathrm{E}(u_i)}{s_{zx}} = \beta_1$$

because $\mathrm{E}(u_i) = 0$ for all i.

(ii) From the fourth equation in part (i) we have (again conditional on the z_i and x_i in the sample),

$$\mathrm{Var}(\tilde{\beta}_1) = \frac{\mathrm{Var}\left[\sum_{i=1}^{n}(z_i - \bar{z})u_i\right]}{s_{zx}^2} = \frac{\sum_{i=1}^{n}(z_i - \bar{z})^2 \mathrm{Var}(u_i)}{s_{zx}^2}$$

$$= \sigma^2 \frac{\sum_{i=1}^{n}(z_i - \bar{z})^2}{s_{zx}^2}$$

because of the homoskedasticity assumption [Var(u_i) = σ^2 for all i]. Given the definition of s_{zx}, this is what we wanted to show.

(iii) We know that Var($\hat{\beta}_1$) = σ^2/[$\sum_{i=1}^{n}(x_i - \bar{x})^2$]. Now we can rearrange the inequality in the hint, drop $\bar{x}$ from the sample covariance, and cancel n^{-1} everywhere, to get [$\sum_{i=1}^{n}(z_i - \bar{z})^2$]/$s_{zx}^2$ ≥ 1/[$\sum_{i=1}^{n}(x_i - \bar{x})^2$]. When we multiply through by σ^2 we get Var($\tilde{\beta}_1$) ≥ Var($\hat{\beta}_1$), which is what we wanted to show.

3.15 (i) The degrees of freedom of the first regression is $n - k - 1 = 353 - 1 - 1 = 351$.The degrees of freedom of the second regression is $n - k - 1 = 353 - 2 - 1 = 350$. The standard error is smaller than the simple regression equation because one more explanatory variable is included in the second regression. The SSR falls from 326.196 to 198.475 when another explanatory variable is added, and the degrees of freedom also falls by one, which affects the standard error.

(ii) Yes, there is a positive moderate correlation between *years* and *rbisyr*.
VIF$_{years}$ = $\frac{1}{1-R_{years}^2}$ = $\frac{1}{1-0.597}$ = 2.48139; from this value, we can say that there is little collinearity between *years* and *rbisyr*.

(iii) The standard error for the coefficient on *years* in the multiple regression is:

$$\mathrm{se}(\hat{\beta}_{years}) = \hat{\sigma} \big/ [\mathrm{SST}(1 - R_{years}^2]^{1/2}$$

The standard error is smaller than the simple regression equation because one more explanatory variable is included in the second regression. The SSR falls from 326.196 to 198.475 when another explanatory variable is added. The degrees of freedom also falls by one, which affects the standard error. Therefore, the standard error for the coefficient of years in the multiple regression is smaller than its simple regression.

SOLUTIONS TO COMPUTER EXERCISES

C3.1 (i) Probably β_2 > 0, as more income typically means better nutrition for the mother and better prenatal care.

(ii) On the one hand, an increase in income generally increases the consumption of a food, and *cigs* and *faminc* could be positively correlated. On the other, family incomes are also higher for families with more education, and more education and cigarette smoking tend to be negatively correlated. The sample correlation between *cigs* and *faminc* is about –.173, indicating a negative correlation.

(iii) The regressions without and with *faminc* are

$$\widehat{bwght} = 119.77 - .514\, cigs$$

$$n = 1{,}388,\ \ R^2 = .023$$

and

$$\widehat{bwght} = 116.97 - .463\, cigs + .093\ \ faminc$$

$$n = 1{,}388,\ \ R^2 = .030.$$

The effect of cigarette smoking is slightly smaller when *faminc* is added to the regression, but the difference is not great. This is due to the fact that *cigs* and *faminc* are not very correlated, and the coefficient on *faminc* is practically small. (The variable *faminc* is measured in thousands, so $10,000 more in 1988 income increases predicted birth weight by only .93 ounces.)

C3.3 (i) The constant elasticity equation is

$$\widehat{\log(salary)} = 4.62 + .162\log(sales) + .107\log(mktval)$$

$$n = 177,\ \ R^2 = .299.$$

(ii) We cannot include profits in logarithmic form because profits are negative for nine of the companies in the sample. When we add it in levels form, we get

$$\widehat{\log(salary)} = 4.69 + .161\log(sales) + .098\log(mktval) + .000036\, profits$$

$$n = 177, R^2 = .299.$$

The coefficient on *profits* is very small. Here, *profits* are measured in millions, so if profits increase by $1 billion, which means $\Delta profits$ = 1,000 – a huge change – predicted salary increases by about only 3.6%. However, remember that we are holding sales and market value fixed.

Together, these variables (and we could drop *profits* without losing anything) explain almost 30% of the sample variation in log(*salary*). This is certainly not “most” of the variation.

(iii) Adding *ceoten* to the equation gives

$$\widehat{\log(salary)} = 4.56 + .162\log(sales) + .102\log(mktval) + .000029\,profits + .012ceoten$$

$$n = 177,\ \ R^2 = .318.$$

This means that one more year as *CEO* increases predicted salary by about 1.2%.

(iv) The sample correlation between log(*mktval*) and *profits* is about .78, which is fairly high. As we know, this causes no bias in the OLS estimators, although it can cause their variances to be large. Given the fairly substantial correlation between market value and firm profits, it is not too surprising that the latter adds nothing to explaining CEO salaries. Also, *profits* is a short term measure of how the firm is doing, while *mktval* is based on past, current, and expected future profitability.

C3.5 The regression of *educ* on *exper* and *tenure* yields

$$educ = 13.57 - .074\ exper + .048\ tenure + \hat{r}_1.$$

$$n = 526,\ \ R^2 = .101.$$

Now, when we regress log(*wage*) on $\hat{r}_1$ we obtain

$$\widehat{\log(wage)} = 1.62 + .092\ \hat{r}_1$$

$$n = 526,\ \ R^2 = .207.$$

As expected, the coefficient on $\hat{r}_1$ in the second regression is identical to the coefficient on *educ* in equation (3.19). Notice that the *R*-squared from the above regression is below that in (3.19). In effect, the regression of log(*wage*) on $\hat{r}_1$ explains log(*wage*) using only the part of *educ* that is uncorrelated with *exper* and *tenure*; separate effects of *exper* and *tenure* are not included.

C3.7 (i) The results of the regression are

$$\widehat{math10} = -20.36\ +\ 6.23\ \log(expend)\ -\ .305\ lnchprg$$

$$n = 408,\ \ R^2 = .180.$$

The signs of the estimated slopes imply that more spending increases the pass rate (holding *lnchprg* fixed) and a higher poverty rate (proxied well by *lnchprg*) decreases the pass rate (holding spending fixed). These are what we expect.

(ii) As usual, the estimated intercept is the predicted value of the dependent variable when all regressors are set to zero. Setting *lnchprg* = 0 makes sense, as there are schools with low poverty rates. Setting log(*expend*) = 0 does not make sense, because it is the same as setting *expend* = 1, and spending is measured in dollars per student. Presumably this is well outside any sensible range. Not surprisingly, the prediction of a -20 pass rate is nonsensical.

(iii) The simple regression results are

$$\widehat{math10} = -69.34 \ + \ 11.16 \ \log(expend)$$

$$n \ = \ 408, \quad R^2 \ = \ .030.$$

and the estimated spending effect is larger than it was in part (i) – almost double.

(iv) The sample correlation between *lexpend* and *lnchprg* is about $-.19$, which means that, on average, high schools with poorer students spent less per student. This makes sense, especially in 1993 in Michigan, where school funding was essentially determined by local property tax collections.

(v) We can use equation (3.23). Because Corr(x_1,x_2) < 0, which means $\tilde{\delta}_1 < 0$, and $\hat{\beta}_2 < 0$, the simple regression estimate, $\tilde{\beta}_1$, is larger than the multiple regression estimate, $\hat{\beta}_1$. Intuitively, failing to account for the poverty rate leads to an overestimate of the effect of spending.

C3.9 (i) The estimated equation is

$$\widehat{gift} = -4.55 \ + \ 2.17 \ mailsyear + .0059 \ giftlast + \ 15.36 \ propresp$$
$$n = 4{,}268, \quad R^2 = .0834.$$

The *R*-squared is now about .083, compared with about .014 for the simple regression case. Therefore, the variables *giftlast* and *propresp* help to explain significantly more variation in *gifts* in the sample (although still just over eight percent).

(ii) Holding *giftlast* and *propresp* fixed, one more mailing per year is estimated to increase *gifts* by 2.17 guilders. The simple regression estimate is 2.65, so the multiple regression estimate is somewhat smaller. Remember, the simple regression estimate holds no other factors fixed.

(iii) Because *propresp* is a proportion, it makes little sense to increase it by one. Such an increase can happen only if *propresp* goes from zero to one. Instead, consider a .10 increase in *propresp*, which means a 10 percentage point increase. Then, *gift* is estimated to be 15.36(.1) ≈ 1.54 guilders higher.

(iv) The estimated equation is

$$\widehat{gift} = -7.33 \ + \ 1.20 \ mailsyear - .261 \ giftlast + \ 16.20 \ propresp + .527 \ avggift$$
$$n = 4{,}268, \quad R^2 = .2005$$

After controlling for the average past gift level, the effect of mailings becomes even smaller: 1.20 guilders, or less than half the effect estimated by simple regression.

(v) After controlling for the average of past gifts – which we can view as measuring the "typical" generosity of the person and is positively related to the current gift level – we find that the current gift amount is negatively related to the most recent gift. A negative relationship makes some sense, as people might follow a large donation with a smaller one.

C3.11 (i) The regression results are:

$$\widehat{math}4 = 96.7704 - 0.8328 pctsgle.$$

The percentage of children not in the married-couples families has a negative impact on percentage of satisfactory level of 4th grade math. The effect of single parenthood seem small. If, say, *pctsgle* increases by .10 (ten percentage points), the percentage of satisfactory level of 4th grade math is estimated to decrease by .08328 percentage, which is a small effect.

(ii) The estimated regression results are:

$$\widehat{math}4 = 51.723 - 0.1996 pctsgle - 0.3964 free + 3.5601 lmedinc.$$

The coefficient of *pctsgle* has negatively increased from -0.8328 to -0.1996. This means that, as the percentage of children not in married couples increases, the percentage of satisfactory level of 4th grade math decreases.

(iii) The sample correlation between *lmedinc* and *free* is -0.74. This is the expected relationship because as the median income increases, the eligibility of the free lunch decreases.

(iv) No, because high correlations among the variables *lmedinc* and *free* do not make it more difficult to determine the causal effect of single parenthood on student performance.

(v) $\text{VIF}_{pctsgle} = \frac{1}{1-R^2} = \frac{1}{1-0.3795} = 1.6116.$
$\text{VIF}_{free} = \frac{1}{1-R^2} = \frac{1}{1-0.4455} = 1.8034.$
$\text{VIF}_{lmedinc} = \frac{1}{1-R^2} = \frac{1}{1-0.3212} = 1.4732.$

By comparing the three variables, it is very clear that the variable *free* has the highest VIF.
No, this knowledge does not affect the model to study the causal effect of single parenthood on math performance.