CHAPTER 1

TEACHING NOTES

You have substantial latitude about what to emphasize in Chapter 1. I find it useful to talk about the economics of crime example (Example 1.1) and the wage example (Example 1.2) so that students see, at the outset, that econometrics is linked to economic reasoning, even if the economics is not complicated theory.

I like to familiarize students with the important data structures that empirical economists use, focusing primarily on cross-sectional and time series data sets, as these are what I cover in a first-semester course. It is probably a good idea to mention the growing importance of data sets that have both a cross-sectional and time dimension.

I spend almost an entire lecture talking about the problems inherent in drawing causal inferences in the social sciences. I do this mostly through the agricultural yield, return to education, and crime examples. These examples also contrast experimental and nonexperimental (observational) data. Students studying business and finance tend to find the term structure of interest rates example more relevant, although the issue there is testing the implication of a simple theory, as opposed to inferring causality. I have found that spending time talking about these examples, in place of a formal review of probability and statistics, is more successful (and more enjoyable for the students and me).

SOLUTIONS TO PROBLEMS

1.1 It does not make sense to pose the question in terms of causality. Economists would assume that students choose a mix of studying and working (and other activities, such as attending class, leisure, and sleeping) based on rational behavior, such as maximizing utility subject to the constraint that there are only 168 hours in a week. We can then use statistical methods to measure the association between studying and working, including regression analysis that we cover starting in Chapter 2. But we would not be claiming that one variable "causes" the other. They are both choice variables of the student.

1.2 (i) Ideally, we could randomly assign students to classes of different sizes. That is, each student is assigned a different class size without regard to any student characteristics such as ability and family background. For reasons we will see in Chapter 2, we would like substantial variation in class sizes (subject, of course, to ethical considerations and resource constraints).

(ii) A negative correlation means that larger class size is associated with lower performance. We might find a negative correlation because larger class size actually hurts performance. However, with observational data, there are other reasons we might find a negative relationship. For example, children from more affluent families might be more likely to attend schools with smaller class sizes, and affluent children generally score better on standardized tests. Another possibility is that, within a school, a principal might assign the better students to smaller classes. Or, some parents might insist their children are in the smaller classes, and these same parents tend to be more involved in their children's education.

(iii) Given the potential for confounding factors – some of which are listed in (ii) – finding a negative correlation would not be strong evidence that smaller class sizes actually lead to better performance. Some way of controlling for the confounding factors is needed, and this is the subject of multiple regression analysis.

1.3 (i) Here is one way to pose the question: If two firms, say *A* and *B*, are identical in all respects except that firm *A* supplies job training one hour per worker more than firm *B*, by how much would firm *A*'s output differ from firm *B*'s?

(ii) Firms are likely to choose job training depending on the characteristics of workers. Some observed characteristics are years of schooling, years in the workforce, and experience in a particular job. Firms might even discriminate based on age, gender, or race. Perhaps firms choose to offer training to more or less able workers, where "ability" might be difficult to quantify but where a manager has some idea about the relative abilities of different employees. Moreover, different kinds of workers might be attracted to firms that offer more job training on average, and this might not be evident to employers.

(iii) The amount of capital and technology available to workers would also affect output. So, two firms with exactly the same kinds of employees would generally have different outputs if they use different amounts of capital or technology. The quality of managers would also have an effect.

(iv) No, unless the amount of training is randomly assigned. The many factors listed in parts (ii) and (iii) can contribute to finding a positive correlation between *output* and *training* even if job training does not improve worker productivity.

SOLUTIONS TO COMPUTER EXERCISES

C1.1 (i) The average of *educ* is about 12.6 years. There are two people reporting zero years of education, and 19 people reporting 18 years of education.

(ii) The average of *wage* is about $5.90, which seems low in the year 2008.

(iii) Using Table B-60 in the 2004 *Economic Report of the President*, the CPI was 56.9 in 1976 and 184.0 in 2003.

(iv) To convert 1976 dollars into 2003 dollars, we use the ratio of the CPIs, which is $184/56.9 \approx 3.23$. Therefore, the average hourly wage in 2003 dollars is roughly $3.23(\$5.90) \approx \$19.06$, which is a reasonable figure.

(v) The sample contains 252 women (the number of observations with *female* = 1) and 274 men.

C1.2 (i) There are 1,388 observations in the sample. Tabulating the variable *cigs* shows that 212 women have *cigs* > 0.

(ii) The average of *cigs* is about 2.09, but this includes the 1,176 women who did not smoke. Reporting just the average masks the fact that almost 85 percent of the women did not smoke. It makes more sense to say that the "typical" woman does not smoke during pregnancy; indeed, the median number of cigarettes smoked is zero.

(iii) The average of *cigs* over the women with *cigs* > 0 is about 13.7. Of course this is much higher than the average over the entire sample because we are excluding 1,176 zeros.

(iv) The average of *fatheduc* is about 13.2. There are 196 observations with a missing value for *fatheduc*, and those observations are necessarily excluded in computing the average.

(v) The average and standard deviation of *faminc* are about 29.027 and 18.739, respectively, but *faminc* is measured in thousands of dollars. So, in dollars, the average and standard deviation are $29,027 and $18,739.

C1.3 (i) The largest is 100, the smallest is 0.

(ii) 38 out of 1,823, or about 2.1 percent of the sample.

(iii) 17

(iv) The average of *math4* is about 71.9 and the average of *read4* is about 60.1. So, at least in 2001, the reading test was harder to pass.

(v) The sample correlation between *math4* and *read4* is about .843, which is a very high degree of (linear) association. Not surprisingly, schools that have high pass rates on one test have a strong tendency to have high pass rates on the other test.

(vi) The average of *exppp* is about $5,194.87. The standard deviation is $1,091.89, which shows rather wide variation in spending per pupil. [The minimum is $1,206.88 and the maximum is $11,957.64.]

C1.4 (i) 185/445 ≈ .416 is the fraction of men receiving job training, or about 41.6%.

(ii) For men receiving job training, the average of *re78* is about 6.35, or $6,350. For men not receiving job training, the average of *re78* is about 4.55, or $4,550. The difference is $1,800, which is very large. On average, the men receiving the job training had earnings about 40% higher than those not receiving training.

(iii) About 24.3% of the men who received training were unemployed in 1978; the figure is 35.4% for men not receiving training. This, too, is a big difference.

(iv) The differences in earnings and unemployment rates suggest the training program had strong, positive effects. Our conclusions about economic significance would be stronger if we could also establish statistical significance (which is done in Computer Exercise C9.10 in Chapter 9).

CHAPTER 2

TEACHING NOTES

This is the chapter where I expect students to follow most, if not all, of the algebraic derivations. In class I like to derive at least the unbiasedness of the OLS slope coefficient, and usually I derive the variance. At a minimum, I talk about the factors affecting the variance. To simplify the notation, after I emphasize the assumptions in the population model, and assume random sampling, I just condition on the values of the explanatory variables in the sample. Technically, this is justified by random sampling because, for example, $E(u_i|x_1,x_2,\ldots,x_n) = E(u_i|x_i)$ by independent sampling. I find that students are able to focus on the key assumption SLR.4 and subsequently take my word about how conditioning on the independent variables in the sample is harmless. (If you prefer, the appendix to Chapter 3 does the conditioning argument carefully.) Because statistical inference is no more difficult in multiple regression than in simple regression, I postpone inference until Chapter 4. (This reduces redundancy and allows you to focus on the interpretive differences between simple and multiple regression.)

You might notice how, compared with most other texts, I use relatively few assumptions to derive the unbiasedness of the OLS slope estimator, followed by the formula for its variance. This is because I do not introduce redundant or unnecessary assumptions. For example, once SLR.4 is assumed, nothing further about the relationship between *u* and *x* is needed to obtain the unbiasedness of OLS under random sampling.

SOLUTIONS TO PROBLEMS

2.1 In the equation $y = \beta_0 + \beta_1 x + u$, add and subtract α_0 from the right hand side to get $y = (\alpha_0 + \beta_0) + \beta_1 x + (u - \alpha_0)$. Call the new error $e = u - \alpha_0$, so that $E(e) = 0$. The new intercept is $\alpha_0 + \beta_0$, but the slope is still β_1.

2.2 (i) Let $y_i = GPA_i$, $x_i = ACT_i$, and $n = 8$. Then $\bar{x} = 25.875$, $\bar{y} = 3.2125$, $\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = 5.8125$, and $\sum_{i=1}^{n}(x_i - \bar{x})^2 = 56.875$. From equation (2.9), we obtain the slope as $\hat{\beta}_1 = 5.8125/56.875 \approx .1022$, rounded to four places after the decimal. From (2.17), $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \approx 3.2125 - (.1022)25.875 \approx .5681$. So we can write

$$\widehat{GPA} = .5681 + .1022\, ACT$$
$$n = 8.$$

The intercept does not have a useful interpretation because *ACT* is not close to zero for the population of interest. If *ACT* is 5 points higher, $\widehat{GPA}$ increases by .1022(5) = .511.

(ii) The fitted values and residuals — rounded to four decimal places — are given along with the observation number *i* and *GPA* in the following table:

i	*GPA*	$\widehat{GPA}$	$\hat{u}$
1	2.8	2.7143	.0857
2	3.4	3.0209	.3791
3	3.0	3.2253	–.2253
4	3.5	3.3275	.1725
5	3.6	3.5319	.0681
6	3.0	3.1231	–.1231
7	2.7	3.1231	–.4231
8	3.7	3.6341	.0659

You can verify that the residuals, as reported in the table, sum to –.0002, which is pretty close to zero given the inherent rounding error.

(iii) When $ACT = 20$, $\widehat{GPA} = .5681 + .1022(20) \approx 2.61$.

(iv) The sum of squared residuals, $\sum_{i=1}^{n} \hat{u}_i^2$, is about .4347 (rounded to four decimal places), and the total sum of squares, $\sum_{i=1}^{n} (y_i - \bar{y})^2$, is about 1.0288. So the *R*-squared from the regression is

$$R^2 = 1 - \text{SSR/SST} \approx 1 - (.4347/1.0288) \approx .577.$$

Therefore, about 57.7% of the variation in *GPA* is explained by *ACT* in this small sample of students.

2.3 (i) Income, age, and family background (such as number of siblings) are just a few possibilities. It seems that each of these could be correlated with years of education. (Income and education are probably positively correlated; age and education may be negatively correlated because women in more recent cohorts have, on average, more education; and number of siblings and education are probably negatively correlated.)

(ii) Not if the factors we listed in part (i) are correlated with *educ*. Because we would like to hold these factors fixed, they are part of the error term. But if *u* is correlated with *educ* then E(*u*|*educ*) $\neq$ 0, and so SLR.4 fails.

2.4 (i) We would want to randomly assign the number of hours in the preparation course so that *hours* is independent of other factors that affect performance on the SAT. Then, we would collect information on SAT score for each student in the experiment, yielding a data set $\{(sat_i, hours_i): i = 1, ..., n\}$, where *n* is the number of students we can afford to have in the study. From equation (2.7), we should try to get as much variation in $hours_i$ as is feasible.

(ii) Here are three factors: innate ability, family income, and general health on the day of the exam. If we think students with higher native intelligence think they do not need to prepare for the SAT, then ability and *hours* will be negatively correlated. Family income would probably be positively correlated with *hours*, because higher income families can more easily afford preparation courses. Ruling out chronic health problems, health on the day of the exam should be roughly uncorrelated with hours spent in a preparation course.

(iii) If preparation courses are effective, β_1 should be positive: other factors equal, an increase in *hours* should increase *sat*.

(iv) The intercept, β_0, has a useful interpretation in this example: because E(*u*) = 0, β_0 is the average SAT score for students in the population with *hours* = 0.

2.5 (i) When we condition on *inc* in computing an expectation, $\sqrt{inc}$ becomes a constant. So E(*u*|*inc*) = E($\sqrt{inc} \cdot e$|*inc*) = $\sqrt{inc} \cdot$ E(*e*|*inc*) = $\sqrt{inc} \cdot 0$ because E(*e*|*inc*) = E(*e*) = 0.

(ii) Again, when we condition on *inc* in computing a variance, $\sqrt{inc}$ becomes a constant. So Var(*u*|*inc*) = Var($\sqrt{inc} \cdot e$|*inc*) = ($\sqrt{inc}$)^{2}Var(*e*|*inc*) = $\sigma_e^2 inc$ because Var(*e*|*inc*) = σ_e^2.

(iii) Families with low incomes do not have much discretion about spending; typically, a low-income family must spend on food, clothing, housing, and other necessities. Higher income people have more discretion, and some might choose more consumption while others more saving. This discretion suggests wider variability in saving among higher income families.

2.6 (i) This derivation is essentially done in equation (2.52), once $(1/\text{SST}_x)$ is brought inside the summation (which is valid because SST_x does not depend on *i*). Then, just define $w_i = d_i / \text{SST}_x$.

(ii) Because $\text{Cov}(\hat{\beta}_1, \bar{u}) = \text{E}[(\hat{\beta}_1 - \beta_1)\bar{u}]$, we show that the latter is zero. But, from part (i), $\text{E}[(\hat{\beta}_1 - \beta_1)\bar{u}] = \text{E}\left[\left(\sum_{i=1}^n w_i u_i\right)\bar{u}\right] = \sum_{i=1}^n w_i \text{E}(u_i \bar{u})$. Because the u_i are pairwise uncorrelated (they are independent), $\text{E}(u_i \bar{u}) = \text{E}(u_i^2 / n) = \sigma^2 / n$ (because $\text{E}(u_i u_h) = 0, \ i \neq h$). Therefore, $\sum_{i=1}^n w_i \text{E}(u_i \bar{u}) = \sum_{i=1}^n w_i (\sigma^2 / n) = (\sigma^2 / n) \sum_{i=1}^n w_i = 0.$

(iii) The formula for the OLS intercept is $\hat{\beta}_0 = \bar{y} - \hat{\beta} \bar{x}$ and, plugging in $\bar{y} = \beta_0 + \beta_1 \bar{x} + \bar{u}$ gives $\hat{\beta}_0 = (\beta_0 + \beta_1 \bar{x} + \bar{u}) - \hat{\beta}_1 \bar{x} = \beta_0 + \bar{u} - (\hat{\beta}_1 - \beta_1)\bar{x}.$

(iv) Because $\hat{\beta}_1$ and $\bar{u}$ are uncorrelated,

$$\text{Var}(\hat{\beta}_0) = \text{Var}(\bar{u}) + \text{Var}(\hat{\beta}_1)\bar{x}^2 = \sigma^2 / n + (\sigma^2 / \text{SST}_x)\bar{x}^2 = \sigma^2 / n + \sigma^2 \bar{x}^2 / \text{SST}_x,$$

which is what we wanted to show.

(v) Using the hint and substitution gives $\text{Var}(\hat{\beta}_0) = \sigma^2[(\text{SST}_x / n) + \bar{x}^2] / \text{SST}_x$
$= \sigma^2 \left[\left(n^{-1} \sum_{i=1}^n x_i^2 - \bar{x}^2\right) + \bar{x}^2\right] / \text{SST}_x = \sigma^2 \left(n^{-1} \sum_{i=1}^n x_i^2\right) / \text{SST}_x.$

2.7 (i) Yes. If living closer to an incinerator depresses housing prices, then being farther away increases housing prices.

(ii) If the city chose to locate the incinerator in an area away from more expensive neighborhoods, then log(*dist*) is positively correlated with housing quality. This would violate SLR.4, and OLS estimation is biased.

(iii) Size of the house, number of bathrooms, size of the lot, age of the home, and quality of the neighborhood (including school quality), are just a handful of factors. As mentioned in part (ii), these could certainly be correlated with *dist* [and log(*dist*)].

2.8 (i) We follow the hint, noting that $\overline{c_1 y} = c_1 \bar{y}$ (the sample average of $c_1 y_i$ is c_1 times the sample average of y_i) and $\overline{c_2 x} = c_2 \bar{x}$. When we regress $c_1 y_i$ on $c_2 x_i$ (including an intercept) we use equation (2.19) to obtain the slope:

$$\tilde{\beta}_1 = \frac{\sum_{i=1}^{n}(c_2 x_i - c_2 x)(c_1 \bar{y}_i - c_1 \bar{y})}{\sum_{i=1}^{n}(c_2 x_i - c_2 \bar{x})^2} = \frac{\sum_{i=1}^{n} c_1 c_2 (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} c_2^2 (x_i - \bar{x})^2}$$

$$= \frac{c_1}{c_2} \cdot \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{c_1}{c_2}\hat{\beta}_1.$$

From (2.17), we obtain the intercept as $\tilde{\beta}_0 = (c_1 \bar{y}) - \tilde{\beta}_1 (c_2 \bar{x}) = (c_1 \bar{y}) - [(c_1/c_2)\hat{\beta}_1](c_2 \bar{x}) = c_1(\bar{y} - \hat{\beta}_1 \bar{x}) = c_1 \hat{\beta}_0$) because the intercept from regressing y_i on x_i is $(\bar{y} - \hat{\beta}_1 \bar{x})$.

(ii) We use the same approach from part (i) along with the fact that $\overline{(c_1 + y)} = c_1 + \bar{y}$ and $\overline{(c_2 + x)} = c_2 + \bar{x}$. Therefore, $(c_1 + y_i) - \overline{(c_1 + y)} = (c_1 + y_i) - (c_1 + \bar{y}) = y_i - \bar{y}$ and $(c_2 + x_i) - \overline{(c_2 + x)} = x_i - \bar{x}$. So c_1 and c_2 entirely drop out of the slope formula for the regression of $(c_1 + y_i)$ on $(c_2 + x_i)$, and $\tilde{\beta}_1 = \hat{\beta}_1$. The intercept is $\tilde{\beta}_0 = \overline{(c_1 + y)} - \tilde{\beta}_1 \overline{(c_2 + x)} = (c_1 + \bar{y}) - \hat{\beta}_1 (c_2 + \bar{x}) = (\bar{y} - \hat{\beta}_1 \bar{x}) + c_1 - c_2 \hat{\beta}_1 = \hat{\beta}_0 + c_1 - c_2 \hat{\beta}_1$, which is what we wanted to show.

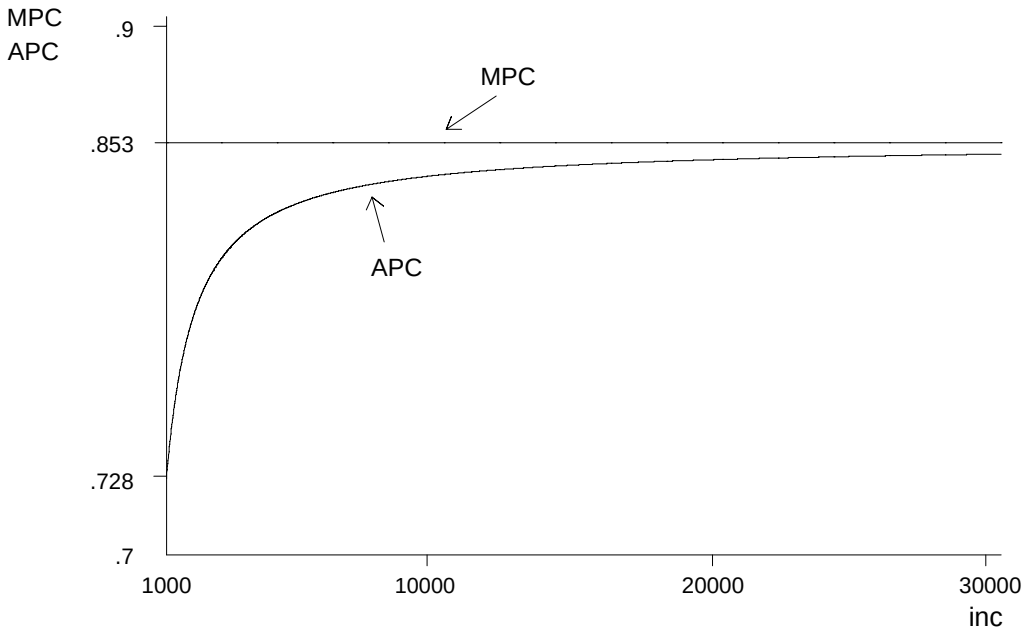
(iii) We can simply apply part (ii) because $\log(c_1 y_i) = \log(c_1) + \log(y_i)$. In other words, replace c_1 with $\log(c_1)$, y_i with $\log(y_i)$, and set $c_2 = 0$.

(iv) Again, we can apply part (ii) with $c_1 = 0$ and replacing c_2 with $\log(c_2)$ and x_i with $\log(x_i)$. If $\hat{\beta}_0$ and $\hat{\beta}_1$ are the original intercept and slope, then $\tilde{\beta}_1 = \hat{\beta}_1$ and $\tilde{\beta}_0 = \hat{\beta}_0 - \log(c_2)\hat{\beta}_1$.

2.9 (i) The intercept implies that when *inc* = 0, *cons* is predicted to be negative \$124.84. This, of course, cannot be true, and reflects that fact that this consumption function might be a poor predictor of consumption at very low-income levels. On the other hand, on an annual basis, \$124.84 is not so far from zero.

(ii) Just plug 30,000 into the equation: $\widehat{cons} = -124.84 + .853(30{,}000) = 25{,}465.16$ dollars.

(iii) The MPC and the APC are shown in the following graph. Even though the intercept is negative, the smallest APC in the sample is positive. The graph starts at an annual income level of \$1,000 (in 1970 dollars).



2.10 (i) From equation (2.66),

$$\tilde{\beta}_1 = \left(\sum_{i=1}^{n} x_i y_i \right) / \left(\sum_{i=1}^{n} x_i^2 \right).$$

Plugging in $y_i = \beta_0 + \beta_1 x_i + u_i$ gives

$$\tilde{\beta}_1 = \left(\sum_{i=1}^{n} x_i (\beta_0 + \beta_1 x_i + u_i) \right) / \left(\sum_{i=1}^{n} x_i^2 \right).$$

After standard algebra, the numerator can be written as

$$\beta_0 \sum_{i=1}^{n} x_i + \beta_1 \sum_{i=1}^{n} x_i^2 + \sum_{i=1}^{n} x_i u_i \ .$$

Putting this over the denominator shows we can write $\tilde{\beta}_1$ as

$$\tilde{\beta}_1 = \beta_0 \left(\sum_{i=1}^{n} x_i \right) / \left(\sum_{i=1}^{n} x_i^2 \right) + \beta_1 + \left(\sum_{i=1}^{n} x_i u_i \right) / \left(\sum_{i=1}^{n} x_i^2 \right).$$

Conditional on the x_i, we have

$$\mathrm{E}(\tilde{\beta}_1) = \beta_0\left(\sum_{i=1}^{n} x_i\right) / \left(\sum_{i=1}^{n} x_i^2\right) + \beta_1$$

because $\mathrm{E}(u_i) = 0$ for all i. Therefore, the bias in $\tilde{\beta}_1$ is given by the first term in this equation. This bias is obviously zero when $\beta_0 = 0$. It is also zero when $\sum_{i=1}^{n} x_i = 0$, which is the same as $\bar{x} = 0$. In the latter case, regression through the origin is identical to regression with an intercept.

(ii) From the last expression for $\tilde{\beta}_1$ in part (i) we have, conditional on the x_i,

$$\mathrm{Var}(\tilde{\beta}_1) = \left(\sum_{i=1}^{n} x_i^2\right)^{-2} \mathrm{Var}\left(\sum_{i=1}^{n} x_i u_i\right) = \left(\sum_{i=1}^{n} x_i^2\right)^{-2} \left(\sum_{i=1}^{n} x_i^2 \mathrm{Var}(u_i)\right)$$

$$= \left(\sum_{i=1}^{n} x_i^2\right)^{-2} \left(\sigma^2 \sum_{i=1}^{n} x_i^2\right) = \sigma^2 / \left(\sum_{i=1}^{n} x_i^2\right).$$

(iii) From (2.57), $\mathrm{Var}(\hat{\beta}_1) = \sigma^2 / \left(\sum_{i=1}^{n} (x_i - \bar{x})^2\right)$. From the hint, $\sum_{i=1}^{n} x_i^2 \geq \sum_{i=1}^{n} (x_i - \bar{x})^2$, and so $\mathrm{Var}(\tilde{\beta}_1) \leq \mathrm{Var}(\hat{\beta}_1)$. A more direct way to see this is to write $\sum_{i=1}^{n} (x_i - \bar{x})^2 = \sum_{i=1}^{n} x_i^2 - n(\bar{x})^2$, which is less than $\sum_{i=1}^{n} x_i^2$ unless $\bar{x} = 0$.

(iv) For a given sample size, the bias in $\tilde{\beta}_1$ increases as $\bar{x}$ increases (holding the sum of the x_i^2 fixed). But as $\bar{x}$ increases, the variance of $\hat{\beta}_1$ increases relative to $\mathrm{Var}(\tilde{\beta}_1)$. The bias in $\tilde{\beta}_1$ is also small when β_0 is small. Therefore, whether we prefer $\tilde{\beta}_1$ or $\hat{\beta}_1$ on a mean squared error basis depends on the sizes of β_0, $\bar{x}$, and n (in addition to the size of $\sum_{i=1}^{n} x_i^2$).

2.11 (i) When *cigs* = 0, predicted birth weight is 119.77 ounces. When *cigs* = 20, $\widehat{bwght}$ = 109.49. This is about an 8.6% drop.

(ii) Not necessarily. There are many other factors that can affect birth weight, particularly overall health of the mother and quality of prenatal care. These could be correlated with cigarette smoking during birth. Also, something such as caffeine consumption can affect birth weight, and might also be correlated with cigarette smoking.

(iii) If we want a predicted *bwght* of 125, then *cigs* = (125 – 119.77)/(–.524) ≈ –10.18, or about –10 cigarettes! This is nonsense, of course, and it shows what happens when we are trying to predict something as complicated as birth weight with only a single explanatory variable. The largest predicted birth weight is necessarily 119.77. Yet almost 700 of the births in the sample had a birth weight higher than 119.77.

(iv) 1,176 out of 1,388 women did not smoke while pregnant, or about 84.7%. Because we are using only *cigs* to explain birth weight, we have only one predicted birth weight at *cigs* = 0. The predicted birth weight is necessarily roughly in the middle of the observed birth weights at *cigs* = 0, and so we will under predict high birth rates.

SOLUTIONS TO COMPUTER EXERCISES

C2.1 (i) The average *prate* is about 87.36 and the average *mrate* is about .732.

(ii) The estimated equation is

$$\widehat{prate} = 83.05 + 5.86\, mrate$$

$$n = 1{,}534,\ R^2 = .075.$$

(iii) The intercept implies that, even if *mrate* = 0, the predicted participation rate is 83.05 percent. The coefficient on *mrate* implies that a one-dollar increase in the match rate – a fairly large increase – is estimated to increase *prate* by 5.86 percentage points. This assumes, of course, that this change *prate* is possible (if, say, *prate* is already at 98, this interpretation makes no sense).

(iv) If we plug *mrate* = 3.5 into the equation we get $\hat{prate} = 83.05 + 5.86(3.5) = 103.59$. This is impossible, as we can have at most a 100 percent participation rate. This illustrates that, especially when dependent variables are bounded, a simple regression model can give strange predictions for extreme values of the independent variable. (In the sample of 1,534 firms, only 34 have *mrate* ≥ 3.5.)

(v) *mrate* explains about 7.5% of the variation in *prate*. This is not much, and suggests that many other factors influence 401(k) plan participation rates.

C2.2 (i) Average salary is about 865.864, which means $865,864 because *salary* is in thousands of dollars. Average *ceoten* is about 7.95.

(ii) There are five CEOs with *ceoten* = 0. The longest tenure is 37 years.

(iii) The estimated equation is

$$\widehat{\log(salary)} = 6.51 + .0097\, ceoten$$

$$n = 177,\ R^2 = .013.$$

We obtain the approximate percentage change in *salary* given $\Delta ceoten = 1$ by multiplying the coefficient on *ceoten* by 100, 100(.0097) = .97%. Therefore, one more year as CEO is predicted to increase salary by almost 1%.

C2.3 (i) The estimated equation is

$$\widehat{sleep} = 3{,}586.4 - .151\ totwrk$$

$$n = 706,\ R^2 = .103.$$

The intercept implies that the estimated amount of sleep per week for someone who does not work is 3,586.4 minutes, or about 59.77 hours. This comes to about 8.5 hours per night.

(ii) If someone works two more hours per week then $\Delta totwrk = 120$ (because *totwrk* is measured in minutes), and so $\Delta\widehat{sleep} = -.151(120) = -18.12$ minutes. This is only a few minutes a night. If someone were to work one more hour on each of five working days, $\Delta\widehat{sleep} = -.151(300) = -45.3$ minutes, or about five minutes a night.

C2.4 (i) Average salary is about \$957.95 and average IQ is about 101.28. The sample standard deviation of IQ is about 15.05, which is pretty close to the population value of 15.

(ii) This calls for a level-level model:

$$\widehat{wage} = 116.99 + 8.30\ IQ$$

$$n = 935,\ R^2 = .096.$$

An increase in *IQ* of 15 increases predicted monthly salary by 8.30(15) = \$124.50 (in 1980 dollars). *IQ* score does not even explain 10% of the variation in *wage*.

(iii) This calls for a log-level model:

$$\widehat{\log(wage)} = 5.89 + .0088\ IQ$$

$$n = 935,\ R^2 = .099.$$

If $\Delta IQ = 15$ then $\Delta\widehat{\log(wage)} = .0088(15) = .132$, which is the (approximate) proportionate change in predicted wage. The percentage increase is therefore approximately 13.2.

C2.5 (i) The constant elasticity model is a log-log model:

$$\log(rd) = \beta_0 + \beta_1 \log(sales) + u,$$

where β_1 is the elasticity of *rd* with respect to *sales*.

(ii) The estimated equation is

$$\widehat{\log(rd)} = -4.105 + 1.076 \log(sales)$$

$$n = 32,\ R^2 = .910.$$

The estimated elasticity of *rd* with respect to *sales* is 1.076, which is just above one. A one percent increase in *sales* is estimated to increase *rd* by about 1.08%.

C2.6 (i) It seems plausible that another dollar of spending has a larger effect for low-spending schools than for high-spending schools. At low-spending schools, more money can go toward purchasing more books, computers, and for hiring better qualified teachers. At high levels of spending, we would expend little, if any, effect because the high-spending schools already have high-quality teachers, nice facilities, plenty of books, and so on.

(ii) If we take changes, as usual, we obtain

$$\Delta math10 = \beta_1 \Delta \log(expend) \approx (\beta_1/100)(\%\Delta expend),$$

just as in the second row of Table 2.3. So, if $\%\Delta expend = 10$, $\Delta math10 = \beta_1/10$.

(iii) The regression results are

$$\widehat{math10} = -69.34 + 11.16 \log(expend)$$
$$n = 408, \quad R^2 = .0297$$

(iv) If *expend* increases by 10 percent, $\widehat{math10}$ increases by about 1.1 percentage points. This is not a huge effect, but it is not trivial for low-spending schools, where a 10 percent increase in spending might be a fairly small dollar amount.

(v) In this data set, the largest value of *math10* is 66.7, which is not especially close to 100. In fact, the largest fitted values is only about 30.2.

C2.7 (i) The average gift is about 7.44 Dutch guilders. Out of 4,268 respondents, 2,561 did not give a gift, or about 60 percent.

(ii) The average mailings per year is about 2.05. The minimum value is .25 (which presumably means that someone has been on the mailing list for at least four years) and the maximum value is 3.5.

(iii) The estimated equation is

$$\widehat{gift} = 2.01 + 2.65\ mailsyear$$
$$n = 4{,}268, \quad R^2 = .0138$$

(iv) The slope coefficient from part (iii) means that each mailing per year is associated with – perhaps even "causes" – an estimated 2.65 additional guilders, on average. Therefore, if each mailing costs one guilder, the expected profit from each mailing is estimated to be 1.65 guilders. This is only the average, however. Some mailings generate no contributions, or a contribution less than the mailing cost; other mailings generated much more than the mailing cost.

(v) Because the smallest *mailsyear* in the sample is .25, the smallest predicted value of *gifts* is $2.01 + 2.65(.25) \approx 2.67$. Even if we look at the overall population, where some people have received no mailings, the smallest predicted value is about two. So, with this estimated equation, we never predict zero charitable gifts.

CHAPTER 3

TEACHING NOTES

For undergraduates, I do not work through most of the derivations in this chapter, at least not in detail. Rather, I focus on interpreting the assumptions, which mostly concern the population. Other than random sampling, the only assumption that involves more than population considerations is the assumption about no perfect collinearity, where the possibility of perfect collinearity in the sample (even if it does not occur in the population) should be touched on. The more important issue is perfect collinearity in the population, but this is fairly easy to dispense with via examples. These come from my experiences with the kinds of model specification issues that beginners have trouble with.

The comparison of simple and multiple regression estimates – based on the particular sample at hand, as opposed to their statistical properties – usually makes a strong impression. Sometimes I do not bother with the "partialling out" interpretation of multiple regression.

As far as statistical properties, notice how I treat the problem of including an irrelevant variable: no separate derivation is needed, as the result follows form Theorem 3.1.

I do like to derive the omitted variable bias in the simple case. This is not much more difficult than showing unbiasedness of OLS in the simple regression case under the first four Gauss-Markov assumptions. It is important to get the students thinking about this problem early on, and before too many additional (unnecessary) assumptions have been introduced.

I have intentionally kept the discussion of multicollinearity to a minimum. This partly indicates my bias, but it also reflects reality. It is, of course, very important for students to understand the potential consequences of having highly correlated independent variables. But this is often beyond our control, except that we can ask less of our multiple regression analysis. If two or more explanatory variables are highly correlated in the sample, we should not expect to precisely estimate their ceteris paribus effects in the population.

I find extensive treatments of multicollinearity, where one "tests" or somehow "solves" the multicollinearity problem, to be misleading, at best. Even the organization of some texts gives the impression that imperfect multicollinearity is somehow a violation of the Gauss-Markov assumptions: they include multicollinearity in a chapter or part of the book devoted to "violation of the basic assumptions," or something like that. I have noticed that master's students who have had some undergraduate econometrics are often confused on the multicollinearity issue. It is very important that students not confuse multicollinearity among the included explanatory variables in a regression model with the bias caused by omitting an important variable.

I do not prove the Gauss-Markov theorem. Instead, I emphasize its implications. Sometimes, and certainly for advanced beginners, I put a special case of Problem 3.12 on a midterm exam, where I make a particular choice for the function $g(x)$. Rather than have the students directly

compare the variances, they should appeal to the Gauss-Markov theorem for the superiority of OLS over any other linear, unbiased estimator.

SOLUTIONS TO PROBLEMS

3.1 (i) Yes. Because of budget constraints, it makes sense that, the more siblings there are in a family, the less education any one child in the family has. To find the increase in the number of siblings that reduces predicted education by one year, we solve $1 = .094(\Delta sibs)$, so $\Delta sibs = 1/.094 \approx 10.6$.

(ii) Holding *sibs* and *feduc* fixed, one more year of mother's education implies .131 years more of predicted education. So if a mother has four more years of education, her son is predicted to have about a half a year (.524) more years of education.

(iii) Since the number of siblings is the same, but *meduc* and *feduc* are both different, the coefficients on *meduc* and *feduc* both need to be accounted for. The predicted difference in education between B and A is .131(4) + .210(4) = 1.364.

3.2 (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 is 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) A larger rank for a law school means that the school has less prestige; this lowers starting salaries. For example, a rank of 100 means there are 99 schools thought to be better.

(ii) $\beta_1 > 0$, $\beta_2 > 0$. Both *LSAT* and *GPA* are measures of the quality of the entering class. No matter where better students attend law school, we expect them to earn more, on average. β_3, $\beta_4 > 0$. The number of volumes in the law library and the tuition cost are both measures of the school quality. (Cost is less obvious than library volumes, but should reflect quality of the faculty, physical plant, and so on.)

(iii) This is just the coefficient on *GPA*, multiplied by 100: 24.8%.

(iv) This is an elasticity: a one percent increase in library volumes implies a .095% increase in predicted median starting salary, other things equal.

(v) It is definitely better to attend a law school with a lower rank. If law school A has a ranking 20 less than law school B, the predicted difference in starting salary is 100(.0033)(20) = 6.6% higher for law school A.

3.4 (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: Δ*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 Conditioning on the outcomes of the explanatory variables, we have $E(\hat{\theta}_1) = E(\hat{\beta}_1 + \hat{\beta}_2) = E(\hat{\beta}_1) + E(\hat{\beta}_2) = \beta_1 + \beta_2 = \theta_1$.

3.6 (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 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 We can use Table 3.2. By definition, $\beta_2 > 0$, and by assumption, Corr(x_1,x_2) < 0. Therefore, there is a negative bias in $\tilde{\beta}_1$: E($\tilde{\beta}_1$) < β_1. This means that, on average across different random samples, the simple regression estimator underestimates the effect of the training program. It is even possible that E($\tilde{\beta}_1$) is negative even though $\beta_1 > 0$.

3.8 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) Because x_1 is highly correlated with x_2 and x_3, and these latter variables have large partial effects on *y*, the simple and multiple regression coefficients on x_1 can differ by large amounts. We have not done this case explicitly, but given equation (3.46) and the discussion with a single omitted variable, the intuition is pretty straightforward.

(ii) Here we would expect $\tilde{\beta}_1$ and $\hat{\beta}_1$ to be similar (subject, of course, to what we mean by "almost uncorrelated"). The amount of correlation between x_2 and x_3 does not directly effect the multiple regression estimate on x_1 if x_1 is essentially uncorrelated with x_2 and x_3.

(iii) In this case we are (unnecessarily) introducing multicollinearity into the regression: x_2 and x_3 have small partial effects on *y* and yet x_2 and x_3 are highly correlated with x_1. Adding x_2 and x_3 like increases the standard error of the coefficient on x_1 substantially, so se($\hat{\beta}_1$) is likely to be much larger than se($\tilde{\beta}_1$).

(iv) In this case, adding x_2 and x_3 will decrease the residual variance without causing much collinearity (because x_1 is almost uncorrelated with x_2 and x_3), so we should see se($\hat{\beta}_1$) smaller than se($\tilde{\beta}_1$). The amount of correlation between x_2 and x_3 does not directly affect se($\hat{\beta}_1$).

3.10 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

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

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 $\hat{r}_{i1}$.

3.11 (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 $\text{Corr}(x_1,x_2) < 0$, 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.

3.12 (i) For notational simplicity, define $s_{zx} = \sum_{i=1}^{n} (z_i - \bar{z}) x_i$; 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

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

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

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

where we use the fact that $\sum_{i=1}^{n}(z_i - \bar{z}) = 0$ 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,

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

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 [$\mathrm{Var}(u_i) = \sigma^2$ for all i]. Given the definition of s_{zx}, this is what we wanted to show.

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

3.13 (i) The shares, by definition, add to one. If we do not omit one of the shares then the equation would suffer from perfect multicollinearity. The parameters would not have a ceteris paribus interpretation, as it is impossible to change one share while holding *all* of the other shares fixed.

(ii) Because each share is a proportion (and can be at most one, when all other shares are zero), it makes little sense to increase $share_p$ by one unit. If $share_p$ increases by .01 – which is equivalent to a one percentage point increase in the share of property taxes in total revenue – holding $share_I$, $share_S$, and the *other factors* fixed, then *growth* increases by $\beta_1(.01)$. With the other shares fixed, the excluded share, $share_F$, must fall by .01 when $share_p$ increases by .01.

SOLUTIONS TO COMPUTER EXERCISES

C3.1 (i) Probably $\beta_2 > 0$, 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 good, 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.)

C3.2 (i) The estimated equation is

$$\widehat{price} = -19.32 + .128\, sqrft + 15.20\, bdrms$$

$$n = 88,\ R^2 = .632$$

(ii) Holding square footage constant, $\Delta\widehat{price}$ = 15.20 $\Delta bdrms$, and so $\widehat{price}$ increases by 15.20, which means $15,200.

(iii) Now $\Delta\widehat{price}$ = .128 $\Delta sqrft$ + 15.20 $\Delta bdrms$ = .128(140) + 15.20 = 33.12, or $33,120. Because the size of the house is increasing, this is a much larger effect than in (ii).

(iv) About 63.2%.

(v) The predicted price is –19.32 + .128(2,438) + 15.20(4) = 353.544, or $353,544.

(vi) From part (v), the estimated value of the home based only on square footage and number of bedrooms is $353,544. The actual selling price was $300,000, which suggests the buyer underpaid by some margin. But, of course, there are many other features of a house (some that we cannot even measure) that affect price, and we have not controlled for these.

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

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

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

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.

C3.4 (i) The minimum, maximum, and average values for these three variables are given in the table below:

Variable	Average	Minimum	Maximum
atndrte	81.71	6.25	100
priGPA	2.59	.86	3.93
ACT	22.51	13	32

(ii) The estimated equation is

$$\widehat{atndrte} = 75.70 + 17.26\ priGPA - 1.72\ ACT$$

$$n = 680, \quad R^2 = .291.$$

The intercept means that, for a student whose prior GPA is zero and ACT score is zero, the predicted attendance rate is 75.7%. But this is clearly not an interesting segment of the population. (In fact, there are no students in the college population with *priGPA* = 0 and *ACT* = 0, or with values even close to zero.)

(iii) The coefficient on *priGPA* means that, if a student's prior GPA is one point higher (say, from 2.0 to 3.0), the attendance rate is about 17.3 percentage points higher. This holds *ACT* fixed. The negative coefficient on *ACT* is, perhaps initially a bit surprising. Five more points on the *ACT* is predicted to lower attendance by 8.6 percentage points at a given level of *priGPA*. As *priGPA* measures performance in college (and, at least partially, could reflect, past attendance rates), while *ACT* is a measure of potential in college, it appears that students that had more promise (which could mean more innate ability) think they can get by with missing lectures.

(iv) We have $\widehat{atndrte} = 75.70 + 17.267(3.65) - 1.72(20) \approx 104.3$. Of course, a student cannot have higher than a 100% attendance rate. Getting predictions like this is always possible when using regression methods for dependent variables with natural upper or lower bounds. In practice, we would predict a 100% attendance rate for this student. (In fact, this student had an actual attendance rate of 87.5%.)

(v) The difference in predicted attendance rates for A and B is $17.26(3.1 - 2.1) - (21 - 26) = 25.86$.

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

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

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

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

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

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

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.

C3.6 (i) The slope coefficient from the regression *IQ* on *educ* is (rounded to five decimal places) $\tilde{\delta}_1 = 3.53383$.

(ii) The slope coefficient from log(*wage*) on *educ* is $\tilde{\beta}_1 = .05984$.

(iii) The slope coefficients from log(*wage*) on *educ* and *IQ* are $\hat{\beta}_1 = .03912$ and $\hat{\beta}_2 = .00586$, respectively.

(iv) We have $\hat{\beta}_1 + \tilde{\delta}_1\hat{\beta}_2 = .03912 + 3.53383(.00586) \approx .05983$, which is very close to .05984; the small difference is due to rounding error.

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

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

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

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.

C3.8 (i) The average of *prpblck* is .113 with standard deviation .182; the average of *income* is 47,053.78 with standard deviation 13,179.29. It is evident that *prpblck* is a proportion and that *income* is measured in dollars.

(ii) The results from the OLS regression are

$$\widehat{psoda} = .956 + .115\, prpblck + .0000016\, income$$

$$n = 401, \quad R^2 = .064.$$

If, say, *prpblck* increases by .10 (ten percentage points), the price of soda is estimated to increase by .0115 dollars, or about 1.2 cents. While this does not seem large, there are communities with no black population and others that are almost all black, in which case the difference in *psoda* is estimated to be almost 11.5 cents.

(iii) The simple regression estimate on *prpblck* is .065, so the simple regression estimate is actually lower. This is because *prpblck* and *income* are negatively correlated (-.43) and *income* has a positive coefficient in the multiple regression.

(iv) To get a constant elasticity, income should be in logarithmic form. I estimate the constant elasticity model:

$$\widehat{\log(psoda)} = -.794 + .122\ prpblck + .077 \log(income)$$

$$n = 401, \quad R^2 = .068.$$

If *prpblck* increases by .20, log(*psoda*) is estimated to increase by .20(.122) = .0244, or about 2.44 percent.

(v) $\hat{\beta}_{prpblck}$ falls to about .073 when *prppov* is added to the regression.

(vi) The correlation is about $-.84$, which makes sense because poverty rates are determined by income (but not directly in terms of median income).

(vii) There is no argument that they are highly correlated, but we are using them simply as controls to determine if the is price discrimination against blacks. In order to isolate the pure discrimination effect, we need to control for as many measures of income as we can; including both variables makes sense.

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.

CHAPTER 4

TEACHING NOTES

At the start of this chapter is good time to remind students that a specific error distribution played no role in the results of Chapter 3. That is because only the first two moments were derived under the full set of Gauss-Markov assumptions. Nevertheless, normality is needed to obtain exact normal sampling distributions (conditional on the explanatory variables). I emphasize that the full set of CLM assumptions are used in this chapter, but that in Chapter 5 we relax the normality assumption and still perform approximately valid inference. One could argue that the classical linear model results could be skipped entirely, and that only large-sample analysis is needed. But, from a practical perspective, students still need to know where the *t* distribution comes from because virtually all regression packages report *t* statistics and obtain *p*-values off of the *t* distribution. I then find it very easy to cover Chapter 5 quickly, by just saying we can drop normality and still use *t* statistics and the associated *p*-values as being approximately valid. Besides, occasionally students will have to analyze smaller data sets, especially if they do their own small surveys for a term project.

It is crucial to emphasize that we test hypotheses about unknown population parameters. I tell my students that they will be punished if they write something like H_0: $\hat{\beta}_1 = 0$ on an exam or, even worse, H_0: .632 = 0.

One useful feature of Chapter 4 is its illustration of how to rewrite a population model so that it contains the parameter of interest in testing a single restriction. I find this is easier, both theoretically and practically, than computing variances that can, in some cases, depend on numerous covariance terms. The example of testing equality of the return to two- and four-year colleges illustrates the basic method, and shows that the respecified model can have a useful interpretation. Of course, some statistical packages now provide a standard error for linear combinations of estimates with a simple command, and that should be taught, too.

One can use an *F* test for single linear restrictions on multiple parameters, but this is less transparent than a *t* test and does not immediately produce the standard error needed for a confidence interval or for testing a one-sided alternative. The trick of rewriting the population model is useful in several instances, including obtaining confidence intervals for predictions in Chapter 6, as well as for obtaining confidence intervals for marginal effects in models with interactions (also in Chapter 6).

The major league baseball player salary example illustrates the difference between individual and joint significance when explanatory variables (*rbisyr* and *hrunsyr* in this case) are highly correlated. I tend to emphasize the *R*-squared form of the *F* statistic because, in practice, it is applicable a large percentage of the time, and it is much more readily computed. I do regret that this example is biased toward students in countries where baseball is played. Still, it is one of the better examples of multicollinearity that I have come across, and students of all backgrounds seem to get the point.

SOLUTIONS TO PROBLEMS

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

(ii) The proportionate effect on $\widehat{salary}$ 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).

4.2 (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) 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

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

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.

4.4 (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.

4.5 (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.

4.6 (i) We need to compute the *F* statistic for the overall significance of the regression with $n = 142$ and $k = 4$: $F = [.0395/(1 – .0395)](137/4) \approx 1.41$. 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 $[.0330/(1 – .0330)](137/4) \approx 1.17$, which is even lower than in part (i). None of the *t* statistics is significant at a reasonable level.

(iii) We probably should not use the logs, as the logarithm is not defined for firms that have zero for *dkr* or *eps*. Therefore, we would lose some firms in the regression.

(iv) 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.

4.7 (i) $.412 \pm 1.96(.094)$, 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.

4.8 (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 $F = [(.113 – .103)/(1 – .113)](702/2) \approx 3.96$. 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.

4.9 (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.

4.10 (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

$$y = \beta_0 + (\theta_1 + 3\beta_2)x_1 + \beta_2 x_2 + \beta_3 x_3 + u$$

$$= \beta_0 + \theta_1 x_1 + \beta_2 (3x_1 + x_2) + \beta_3 x_3 + u.$$

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.

4.11 (i) Holding *profmarg* fixed, $\Delta\widehat{rdintens} = .321\,\Delta\log(sales) = (.321/100)[100\cdot\Delta\log(sales)] \approx .00321(\%\Delta sales)$. Therefore, if $\%\Delta sales = 10$, $\Delta\widehat{rdintens} \approx .032$, 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.

SOLUTIONS TO COMPUTER EXERCISES

C4.1 (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: $\beta_2 = -\beta_1$.

C4.2 (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

$$\widehat{\log(salary)} = \underset{(0.53)}{8.34} + \underset{(.0040)}{.0047\, LSAT} + \underset{(.090)}{.248\, GPA} + \underset{(.033)}{.095 \log(libvol)} + \underset{(.032)}{.038 \log(cost)} - \underset{(.0003)}{.0033\, rank}$$

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

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.

C4.3 (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

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

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%).

C4.4 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 $F = [(.0387 - .0348)/(1 - .0387)](1{,}185/2) \approx 2.40$, which yields *p*-value $\approx .091$ 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 $\approx .242$.

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

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

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

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

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

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

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.

C4.6 (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.

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:

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

$n = 6{,}763,\ R^2 = .223$

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.

C4.8 (i) There are 2,017 single people in the sample of 9,275.

(ii) The estimated equation is

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

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

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 84 of the text.

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

$$\widehat{\log(psoda)} = \underset{(0.29)}{-1.46} + \underset{(.031)}{.073\ prpblck} + \underset{(.027)}{.137\log(income)} + \underset{(.133)}{.380\ prppov}$$

$n = 401,\ R^2 = .087$

The p-value for testing H_0: $\beta_1 = 0$ against the two-sided alternative is about .018, so that we reject H_0 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

$$\widehat{\log(psoda)} = \underset{(.29)}{-.84} + \underset{(.029)}{.098\ prpblck} - \underset{(.038)}{.053\log(income)} + \underset{(.134)}{.052\ prppov} + \underset{(.018)}{.121\log(hseval)}$$

$n = 401,\ R^2 = .184$

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.

C4.10 (i) Using the 1,848 observations, the simple regression estimate of β_{bs} is about $-.795$. The 95% confidence interval runs from -1.088 to $-.502$, which includes -1. Therefore, at the 5% level, we cannot reject that $H_0 : \beta_{bs} = -1$ against the two-sided alternative.

(ii) When *lenrol* and *lstaff* are added to the regression, the coefficient on *bs* becomes about –.605; it is now statistically different from one, as the 95% CI is from about –.818 to –.392. The situation is very similar to that in Table 4.1, where the simple regression estimate is –.825 and the multiple regression estimate (with the logs of enrollment and staff included) is –.605. (It is a coincidence that the two multiple regression estimates are the same, as the data set in Table 4.1 is for an earlier year at the high school level.)

(iii) The standard error of the simple regression estimate is about .150, and that for the multiple regression estimate is about .109. When we add extra explanatory variables, two factors work in opposite directions on the standard errors. Multicollinearity – in this case, correlation between *bs* and the two variables *lenrol* and *lstaff* works to increase the multiple regression standard error. Working to reduce the standard error of $\hat{\beta}_{bs}$ is the smaller error variance when *lenrol* and *lstaff* are included in the regression; in effect, they are taken out of the simple regression error term. In this particular example, the multicollinearity is modest compared with the reduction in the error variance. In fact, the standard error of the regression goes from .231 for simple regression to .168 in the multiple regression. (Another way to summarize the drop in the error variance is to note that the *R*-squared goes from a very small .0151 for the simple regression to .4882 for multiple regression.) Of course, ahead of time we cannot know which effect will dominate, but we can certainly compare the standard errors after running both regressions.

(iv) The variable *lstaff* is the log of the number of staff per 1,000 students. As *lstaff* increases, there are more teachers per student. We can associate this with smaller class sizes, which are generally desirable from a teacher's perspective. It appears that, all else equal, teachers are willing to take less in salary to have smaller class sizes. The elasticity of *salary* with respect to *staff* is about –.714, which seems quite large: a ten percent increase in staff size (holding enrollment fixed) is associated with a 7.14 percent lower salary.

(v) When *lunch* is added to the regression, its coefficient is about –.00076, with $t = -4.69$. Therefore, other factors fixed (*bs*, *lenrol*, and *lstaff*), a hire poverty rate is associated with *lower* teacher salaries. In this data set, the average value of *lunch* is about 36.3 with standard deviation of 25.4. Therefore, a one standard deviation increase in *lunch* is associated with a change in *lsalary* of about $-.00076(25.4) \approx -.019$, or almost two percent lower. Certainly there is no evidence that teachers are compensated for teaching disadvantaged children.

(vi) Yes, the pattern obtained using ELEM94_95.RAW is very similar to that in Table 4.1, and the magnitudes are reasonably close, too. The largest estimate (in absolute value) is the simple regression estimate, and the absolute value declines as more explanatory variables are added. The final regressions in the two cases are not the same, because we do not control for *lunch* in Table 4.1, and graduation and dropout rates are not relevant for elementary school children.

CHAPTER 5

TEACHING NOTES

Chapter 5 is short, but it is conceptually more difficult than the earlier chapters, primarily because it requires some knowledge of asymptotic properties of estimators. In class, I give a brief, heuristic description of consistency and asymptotic normality before stating the consistency and asymptotic normality of OLS. (Conveniently, the same assumptions that work for finite sample analysis work for asymptotic analysis.) More advanced students can follow the proof of consistency of the slope coefficient in the bivariate regression case. Section E.4 contains a full matrix treatment of asymptotic analysis appropriate for a master's level course.

An explicit illustration of what happens to standard errors as the sample size grows emphasizes the importance of having a larger sample. I do not usually cover the *LM* statistic in a first-semester course, and I only briefly mention the asymptotic efficiency result. Without full use of matrix algebra combined with limit theorems for vectors and matrices, it is very difficult to prove asymptotic efficiency of OLS.

I think the conclusions of this chapter are important for students to know, even though they may not fully grasp the details. On exams I usually include true-false type questions, with explanation, to test the students' understanding of asymptotics. [For example: "In large samples we do not have to worry about omitted variable bias." (False). Or "Even if the error term is not normally distributed, in large samples we can still compute approximately valid confidence intervals under the Gauss-Markov assumptions." (True).]

SOLUTIONS TO PROBLEMS

5.1 Write y = $\beta_0 + \beta_1 x + u$, and take the expected value: E(y) = $\beta_0 + \beta_1$E(x) + E(u), or $\mu_y = \beta_0 + \beta_1 \mu_x$, since E($u$) = 0, where μ_y = E(y) and μ_x = E(x). We can rewrite this as $\beta_0 = \mu_y - \beta_1 \mu_x$. Now, $\tilde{\beta}_0 = \bar{y} - \tilde{\beta}_1 \bar{x}$. Taking the plim of this we have plim($\tilde{\beta}_0$) = plim($\bar{y} - \tilde{\beta}_1 \bar{x}$) = plim($\bar{y}$) – plim($\tilde{\beta}_1$)·plim($\bar{x}$) = $\mu_y - \beta_1 \mu_x$, where we use the fact that plim($\bar{y}$) = μ_y and plim($\bar{x}$) = μ_x by the law of large numbers, and plim($\tilde{\beta}_1$) = β_1. We have also used the parts of the Property PLIM.2 from Appendix C.

5.2 The variable *cigs* has nothing close to a normal distribution in the population. Most people do not smoke, so *cigs* = 0 for over half of the population. A normally distributed random variable takes on no particular value with positive probability. Further, the distribution of *cigs* is skewed, whereas a normal random variable must be symmetric about its mean.

5.3 A higher tolerance of risk means more willingness to invest in the stock market, so $\beta_2 > 0$. By assumption, *funds* and *risktol* are positively correlated. Now we use equation (5.5), where $\delta_1 > 0$: plim($\tilde{\beta}_1$) = $\beta_1 + \beta_2 \delta_1 > \beta_1$, so $\tilde{\beta}_1$ has a positive inconsistency (asymptotic bias). This makes sense: if we omit *risktol* from the regression and it is positively correlated with *funds*, some of the estimated effect of *funds* is actually due to the effect of *risktol*.

5.4 Write y = $\beta_0 + \beta_1 x_1 + u$, and take the expected value: E(y) = $\beta_0 + \beta_1$E(x_1) + E(u), or $\mu_y = \beta_0 + \beta_1 \mu_x$ since E(u) = 0, where μ_y = E(y) and μ_x = E(x_1). We can rewrite this as $\beta_0 = \mu_y - \beta_1 \mu_x$. Now, $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}_1$. Taking the plim of this we have plim($\hat{\beta}_0$) = plim($\bar{y} - \hat{\beta}_1 \bar{x}_1$) = plim($\bar{y}$) – plim($\hat{\beta}_1$)·plim($\bar{x}_1$) = $\mu_y - \beta_1 \mu_x$, where we use the fact that plim($\bar{y}$) = μ_y and plim($\bar{x}_1$) = μ_x by the law of large numbers, and plim($\hat{\beta}_1$) = β_1. We have also used the parts of Property PLIM.2 from Appendix C.

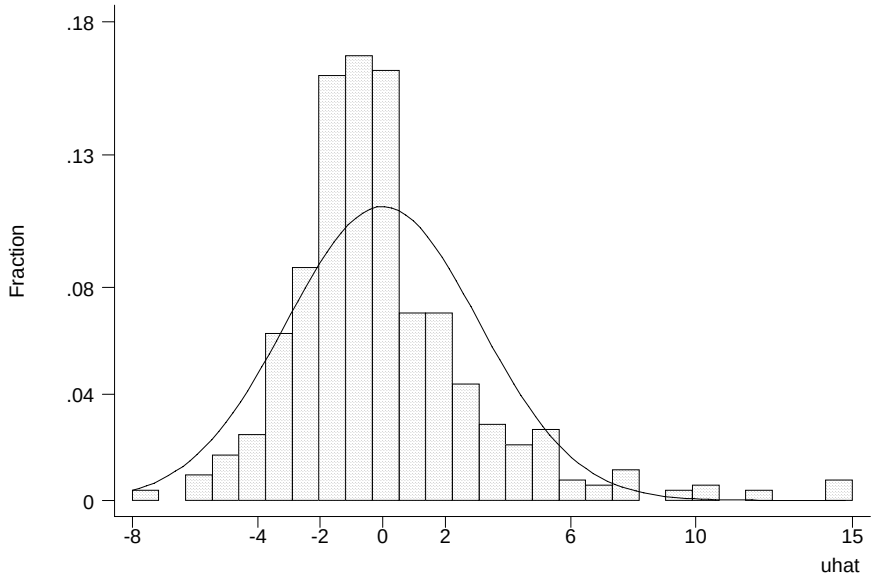
SOLUTIONS TO COMPUTER EXERCISES

C5.1 (i) The estimated equation is

$$\widehat{wage} = -2.87 + .599\ educ + .022\ exper + .169\ tenure$$
$$(0.73)\ (.051)\qquad (.012)\qquad (.022)$$

$n = 526$, $R^2 = .306$, $\hat{\sigma} = 3.085$.

Below is a histogram of the 526 residual, $\hat{u}_i$, $i = 1, 2, ..., 526$. The histogram uses 27 bins, which is suggested by the formula in the Stata manual for 526 observations. For comparison, the normal distribution that provides the best fit to the histogram is also plotted.

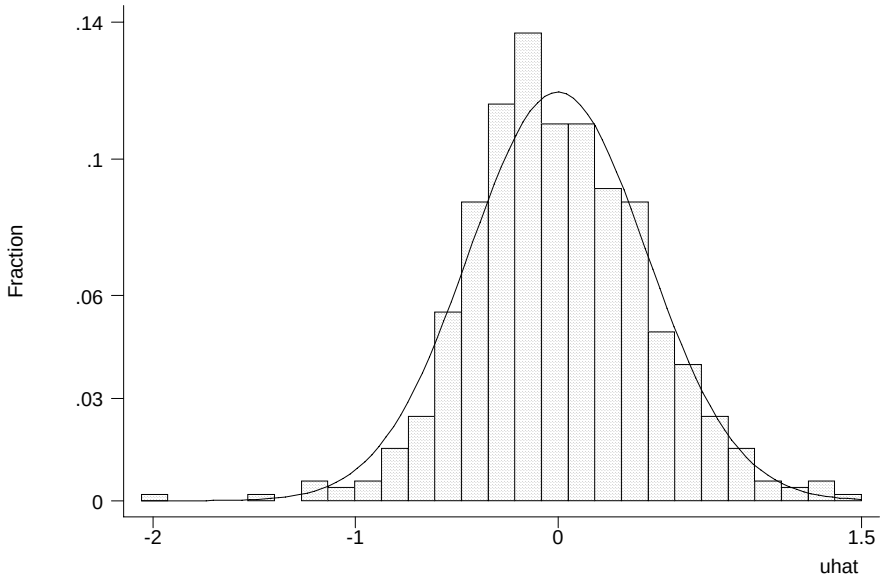


(ii) With log(*wage*) as the dependent variable the estimated equation is

$$\widehat{\log(wage)} = \underset{(.104)}{.284} + \underset{(.007)}{.092}\,educ + \underset{(.0017)}{.0041}\,exper + \underset{(.003)}{.022}\,tenure$$

$n = 526, \quad R^2 = .316, \quad \hat{\sigma} = .441.$

The histogram for the residuals from this equation, with the best-fitting normal distribution overlaid, is given below:



(iii) The residuals from the log(*wage*) regression appear to be more normally distributed. Certainly the histogram in part (ii) fits under its comparable normal density better than in part (i), and the histogram for the *wage* residuals is notably skewed to the left. In the *wage* regression there are some very large residuals (roughly equal to 15) that lie almost five estimated standard deviations ($\hat{\sigma}$ = 3.085) from the mean of the residuals, which is identically zero, of course. Residuals far from zero does not appear to be nearly as much of a problem in the log(*wage*) regression.

C5.2 (i) The regression with all 4,137 observations is

$$\widehat{colgpa} = \underset{}{1.392} \underset{(0.072)}{- .01352\ hsperc} \underset{(.00055)}{+ .00148\ sat} \qquad (.00007)$$

$n = 4{,}137, \quad R^2 = .273.$

(ii) Using only the first 2,070 observations gives

$$\widehat{colgpa} = 1.436 - .01275\, hsperc + .00147\, sat$$
$$(0.098)\quad (.00072)\qquad (.00009)$$

$n = 2{,}070, \quad R^2 = .283.$

(iii) The ratio of the standard error using 2,070 observations to that using 4,137 observations is about 1.31. From (5.10) we compute $\sqrt{(4{,}137/2{,}070)} \approx 1.41$, which is somewhat above the ratio of the actual standard errors.

C5.3 We first run the regression *colgpa* on *cigs*, *parity*, and *faminc* using only the 1,191 observations with nonmissing observations on *motheduc* and *fatheduc*. After obtaining these residuals, $\tilde{u}_i$, these are regressed on $cigs_i$, $parity_i$, $faminc_i$, $motheduc_i$, and $fatheduc_i$, where, of course, we can only use the 1,197 observations with nonmissing values for both *motheduc* and *fatheduc*. The *R*-squared from this regression, R_u^2, is about .0024. With 1,191 observations, the chi-square statistic is (1,191)(.0024) $\approx$ 2.86. The *p*-value from the χ_2^2 distribution is about .239, which is very close to .242, the *p*-value for the comparable *F* test.

C5.4 (i) The measure of skewness for *inc* is about 1.86. When we use log(*inc*), the skewness measure is about .360. Therefore, there is much less skewness in log of income, which means *inc* is less likely to be normally distributed. (In fact, the skewness in income distributions is a well-documented fact across many countries and time periods.)

(ii) The skewness for *bwght* is about –.60. When we use log(*bwght*), the skewness measure is about –2.95. In this case, there is much more skewness after taking the natural log.

(iii) The example in part (ii) clearly shows that this statement cannot hold generally. It is possible to introduce skewness by taking the natural log. As an empirical matter, for many economic variables, particularly dollar values, taking the log often does help to reduce or eliminate skewness. But it does not *have* to.

(iv) For the purposes of regression analysis, we should be studying the *conditional* distributions; that is, the distributions of *y* and log(*y*) conditional on the explanatory variables, $x_1, \ldots, x_k$. If we think the mean is linear, as in Assumptions MLR.1 and MLR.3, then this is equivalent to studying the distribution of the population error, *u*. In fact, the skewness measure studied in this question often is applied to the residuals from and OLS regression.

CHAPTER 6

TEACHING NOTES

I cover most of Chapter 6, but not all of the material in great detail. I use the example in Table 6.1 to quickly run through the effects of data scaling on the important OLS statistics. (Students should already have a feel for the effects of data scaling on the coefficients, fitting values, and *R*-squared because it is covered in Chapter 2.) At most, I briefly mention beta coefficients; if students have a need for them, they can read this subsection.

The functional form material is important, and I spend some time on more complicated models involving logarithms, quadratics, and interactions. An important point for models with quadratics, and especially interactions, is that we need to evaluate the partial effect at interesting values of the explanatory variables. Often, zero is not an interesting value for an explanatory variable and is well outside the range in the sample. Using the methods from Chapter 4, it is easy to obtain confidence intervals for the effects at interesting *x* values.

As far as goodness-of-fit, I only introduce the adjusted *R*-squared, as I think using a slew of goodness-of-fit measures to choose a model can be confusing to novices (and does not reflect empirical practice). It is important to discuss how, if we fixate on a high *R*-squared, we may wind up with a model that has no interesting ceteris paribus interpretation.

I often have students and colleagues ask if there is a simple way to predict *y* when log(*y*) has been used as the dependent variable, and to obtain a goodness-of-fit measure for the log(*y*) model that can be compared with the usual *R*-squared obtained when *y* is the dependent variable. The methods described in Section 6.4 are easy to implement and, unlike other approaches, do not require normality.

The section on prediction and residual analysis contains several important topics, including constructing prediction intervals. It is useful to see how much wider the prediction intervals are than the confidence interval for the conditional mean. I usually discuss some of the residual-analysis examples, as they have real-world applicability.

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.

6.3 (i) Holding all other factors fixed we have

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

Dividing both sides by $\Delta educ$ gives the result. The sign of β_2 is not obvious, although $\beta_2 > 0$ if we think a child gets more out of another year of education the more highly educated are the child's parents.

(ii) We use the values *pareduc* = 32 and *pareduc* = 24 to interpret the coefficient on $educ \cdot pareduc$. The difference in the estimated return to education is .00078(32 – 24) = .0062, or about .62 percentage points.

(iii) When we add *pareduc* by itself, the coefficient on the interaction term is negative. The *t* statistic on $educ \cdot pareduc$ is about –1.33, which is not significant at the 10% level against a two-sided alternative. Note that the coefficient on *pareduc* is significant at the 5% level against a two-sided alternative. This provides a good example of how omitting a level effect (*pareduc* in this case) can lead to biased estimation of the interaction effect.

6.4 (i) The answer is not entirely obvious, but one must properly interpret the coefficient on *alcohol* in either case. If we include *attend*, then we are measuring the effect of alcohol consumption on college GPA, holding attendance fixed. Because attendance is likely to be an important mechanism through which drinking affects performance, we probably do not want to hold it fixed in the analysis. If we do include *attend*, then we interpret the estimate of $\beta_{alcohol}$ as measuring those effects on *colGPA* that are not due to attending class. (For example, we could be measuring the effects that drinking alcohol has on study time.) To get a total effect of alcohol consumption, we would leave *attend* out.

(ii) We would want to include *SAT* and *hsGPA* as controls, as these measure student abilities and motivation. Drinking behavior in college could be correlated with one's performance in high school and on standardized tests. Other factors, such as family background, would also be good controls.

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:

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

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

(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.

6.7 By definition of the OLS regression of c_0y_i on $c_1x_{i1}, \ldots, c_kx_{ik}$, $i = 2, \ldots, n$, the $\tilde{\beta}_j$ solve

$$\sum_{i=1}^{n}[(c_0y_i) - \tilde{\beta}_0 - \tilde{\beta}_1(c_1x_{i1}) - \ldots - \tilde{\beta}_k(c_kx_{ik})] = 0$$

$$\sum_{i=1}^{n}(c_1x_{i1})[(c_0y_i) - \tilde{\beta}_0 - \tilde{\beta}_1(c_1x_{i1}) - \ldots - \tilde{\beta}_k(c_kx_{ik})] = 0$$

$$\vdots$$

$$\sum_{i=1}^{n}(c_kx_{ik})[(c_0y_i) - \tilde{\beta}_0 - \tilde{\beta}_1(c_1x_{i1}) - \ldots - \tilde{\beta}_k(c_kx_{ik})] = 0.$$

[We obtain these from equations (3.13), where we plug in the scaled dependent and independent variables.] We now show that if $\tilde{\beta}_0 = c_0\hat{\beta}_0$ and $\tilde{\beta}_j = (c_0/c_j)\tilde{\beta}_j$, $j = 1,\ldots,k$, then these $k + 1$ first order conditions are satisfied, which proves the result because we know that the OLS estimates are the unique solutions to the FOCs (once we rule out perfect collinearity in the independent variables). Plugging in these guesses for the $\tilde{\beta}_j$ gives the expressions

$$\sum_{i=1}^{n}[(c_0y_i) - c_0\hat{\beta}_0 - (c_0/c_1)\hat{\beta}_1(c_1x_{i1}) - \ldots - (c_0/c_k)\hat{\beta}_k(c_kx_{ik})]$$

$$\sum_{i=1}^{n}(c_jx_{ij})[(c_0y_i) - c_0\hat{\beta}_0 - (c_0/c_1)\hat{\beta}_1(c_1x_{i1}) - \ldots - (c_0/c_k)\hat{\beta}_k(c_kx_{ik})],$$

for $j = 1,2,\ldots,k$. Simple cancellation shows we can write these equations as

$$\sum_{i=1}^{n}[(c_0y_i) - c_0\hat{\beta}_0 - c_0\hat{\beta}_1x_{i1} - \ldots - c_0\hat{\beta}_kx_{ik}]$$

and

$$\sum_{i=1}^{n}(c_jx_{ij})[(c_0y_i) - c_0\hat{\beta}_0 - c_0\hat{\beta}_1x_{i1} - \ldots - c_0\hat{\beta}_kx_{ik}], \; j = 1,2,\ldots,k$$

or, factoring out constants,

$$c_0\left(\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1x_{i1} - \ldots - \hat{\beta}_kx_{ik})\right)$$

and

$$c_0c_j\left(\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1x_{i1} - \ldots - \hat{\beta}_kx_{ik})\right), \; j = 1, 2, \ldots$$

But the terms multiplying c_0 and c_0c_j are identically zero by the first order conditions for the $\hat{\beta}_j$ since, by definition, they are obtained from the regression y_i on $x_{i1}, \ldots, x_{ik}$, $i = 1,2,..,n$. So we have shown that $\tilde{\beta}_0 = c_0\hat{\beta}_0$ and $\tilde{\beta}_j = (c_0/c_j)\ \hat{\beta}_j$, $j = 1, \ldots, k$ solve the requisite first order conditions.

6.8 The extended model has $df = 680 - 9 = 671$, and we are testing two restrictions. Therefore, $F = [(.232 - .229)/(1 - .232)](671/2) \approx 1.31$, which is well below the 10% critical value in the F distribution with 2 and ∞ df: $cv = 2.30$. Thus, $atndrte^2$ and $ACT \cdot atndrte$ are jointly insignificant. Because adding these terms complicates the model without statistical justification, we would not include them in the final model.

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.

C6.2 (i) The estimated equation is

$$\widehat{\log(wage)} = \underset{(.106)}{.128} + \underset{(.0075)}{.0904}\,educ + \underset{(.0052)}{.0410}\,exper - \underset{(.000116)}{.000714}\,exper^2$$

$n = 526,\ R^2 = .300,\ \bar{R}^2 = .296.$

(ii) The *t* statistic on $exper^2$ is about –6.16, which has a *p*-value of essentially zero. So *exper* is significant at the 1% level(and much smaller significance levels).

(iii) To estimate the return to the fifth year of experience, we start at *exper* = 4 and increase *exper* by one, so $\Delta exper = 1$:

$$\%\Delta\widehat{wage} \approx 100[.0410 - 2(.000714)4] \approx 3.53\%.$$

Similarly, for the 20th year of experience,

$$\%\Delta\widehat{wage} \approx 100[.0410 - 2(.000714)19] \approx 1.39\%$$

(iv) The turnaround point is about $.041/[2(.000714)] \approx 28.7$ years of experience. In the sample, there are 121 people with at least 29 years of experience. This is a fairly sizeable fraction of the sample.

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

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

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

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.

C6.4 (i) The estimated equation is

$$\widehat{sat} = \underset{(6.20)}{997.98} + \underset{(3.99)}{19.81\, hsize} - \underset{(0.55)}{2.13\, hsize^2}$$

$$n = 4{,}137, \quad R^2 = .0076.$$

The quadratic term is very statistically significant, with *t* statistic ≈ -3.87.

(ii) We want the value of *hsize*, say *hsize**, where $\hat{sat}$ reaches its maximum. This is the turning point in the parabola, which we calculate as *hsize** = 19.81/[2(2.13)] ≈ 4.65. Since *hsize* is in 100s, this means 465 students is the "optimal" class size. Of course, the very small *R*-squared shows that class size explains only a tiny amount of the variation in SAT score.

(iii) Only students who actually take the SAT exam appear in the sample, so it is not representative of all high school seniors. If the population of interest is all high school seniors, we need a random sample of such students who all took the same standardized exam.

(iv) With log(*sat*) as the dependent variable we get

$$\widehat{\log(sat)} = \underset{(0.006)}{6.896} + \underset{(.0040)}{.0196\ hsize} - \underset{(.00054)}{.00209\ hsize^2}$$

$$n = 4{,}137, \quad R^2 = .0078.$$

The optimal class size is now estimated as about 469, which is very close to what we obtained with the level-level model.

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

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

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

(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, …, 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.

C6.6 (i) For the model

$$voteA = \beta_0 + \beta_1 prtystrA + \beta_2 expendA + \beta_3 expendB + \beta_4 expendA \cdot expendB + u,$$

the ceteris paribus effect of *expendB* on *voteA* is obtained by taking changes and holding *prtystrA*, *expendA*, and *u* fixed:

$$\Delta voteA = \beta_3 \Delta expendB + \beta_4 expendA(\Delta expendB) = (\beta_3 + \beta_4 expendA)\Delta expendB,$$

or

$$\Delta voteA/\Delta expendB = \beta_3 + \beta_4 expendA.$$

We think $\beta_3 < 0$ if a ceteris paribus increase in spending by B lowers the share of the vote received by A. But the sign of β_4 is ambiguous: Is the effect of more spending by B smaller or larger for higher levels of spending by A?

(ii) The estimated equation is

$$\widehat{voteA} = \underset{(4.59)}{32.12} + \underset{(.088)}{.342\, prtystrA} + \underset{(.0050)}{.0383\, expendA} - \underset{(.0046)}{.0317\, expendB}$$

$$- \underset{(.0000072)}{.0000066\, expendA \cdot expendB}$$

$$n = 173, \quad R^2 = .571, \quad \bar{R}^2 = .561.$$

The interaction term is not statistically significant, as its *t* statistic is less than one in absolute value.

(iii) The average value of *expendA* is about 310.61, or \$310,610. If we set *expendA* at 300, which is close to the average value, we have

$$\Delta\widehat{voteA} = [-.0317 - .0000066 \cdot (300)]\Delta expendB \approx -.0337(\Delta expendB).$$

So, when $\Delta expendB = 100$, $\Delta\widehat{voteA} \approx -3.37$, which is a fairly large effect. (Note that, given the insignificance of the interaction term, we would be justified in leaving it out and reestimating the model. This would make the calculation easier.)

(iv) Now we have

$$\Delta\widehat{voteA} = (\hat{\beta}_2 + \hat{\beta}_4 expendB)\Delta expendA \approx .0376(\Delta expendA) = 3.76$$

when $\Delta expendA = 100$. This does make sense, and it is a nontrivial effect.

(v) When we replace the interaction term with *shareA* we obtain

$$\widehat{voteA} = \underset{(2.57)}{18.20} + \underset{(.050)}{.157}\, prtystrA - \underset{(.0028)}{.0067}\, expendA + \underset{(.0026)}{.0043}\, expendB + \underset{(.025)}{.494}\, shareA$$

$n = 173,\ R^2 = .868,\ \bar{R}^2 = .865.$

Notice how much higher the goodness-of-fit measures are as compared with the equation estimated in part (ii), and how significant *shareA* is. To obtain the partial effect of *expendB* on $\hat{vote}A$ we must compute the partial derivative. Generally, we have

$$\frac{\partial \widehat{voteA}}{\partial expendB} = \hat{\beta}_3 + \hat{\beta}_4 \left(\frac{\partial shareA}{\partial expendB} \right),$$

where *shareA* = 100[*expendA*/(*expendA* + *expendB*)]. Now

$$\frac{\partial shareA}{\partial expendB} = -100 \left(\frac{expendA}{(expendA + expendB)^2} \right).$$

Evaluated at *expendA* = 300 and *expendB* = 0, the partial derivative is $-100(300/300^2) = -1/3$, and therefore

$$\frac{\partial \widehat{voteA}}{\partial expendB} = \hat{\beta}_3 + \hat{\beta}_4 (1/3) = .0043 - .494/3 \approx -.164.$$

So $\hat{vote}A$ falls by .164 percentage points given the first thousand dollars of spending by candidate B, where A's spending is held fixed at 300 (or \$300,000). This is a fairly large effect, although it may not be the most typical scenario (because it is rare to have one candidate spend so much and another spend so little). The effect tapers off as *expendB* grows. For example, at *expendB* = 100, the effect of the thousand dollars of spending is only about $.0043 - .494(.188) \approx -.089$.

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 $\Delta priGPA$ 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

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

(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 $\hat{\theta}_2$.

C6.8 (i) The estimated equation (where *price* is in dollars) is

$$\begin{aligned} \widehat{price} &= -21{,}770.3 + 2.068\ lotsize + 122.78\ sqrft + 13{,}852.5\ bdrms \\ &\quad (29{,}475.0)\quad (0.642)\qquad\quad (13.24)\qquad\quad (9{,}010.1) \end{aligned}$$

$n = 88,\ \ R^2 = .672,\ \ \bar{R}^2 = .661,\ \ \hat{\sigma} = 59{,}833.$

The predicted price at *lotsize* = 10,000, *sqrft* = 2,300, and *bdrms* = 4 is about $336,714.

(ii) The regression is $price_i$ on $(lotsize_i - 10{,}000)$, $(sqrft_i - 2{,}300)$, and $(bdrms_i - 4)$. We want the intercept estimate and the associated 95% CI from this regression. The CI is approximately 336,706.7 ± 14,665, or about $322,042 to $351,372 when rounded to the nearest dollar.

(iii) We must use equation (6.36) to obtain the standard error of $\hat{e}^0$ and then use equation (6.37) (assuming that *price* is normally distributed). But from the regression in part (ii), $se(\hat{y}^0) \approx 7{,}374.5$ and $\hat{\sigma} \approx 59{,}833$. Therefore, $se(\hat{e}^0) \approx [(7{,}374.5)^2 + (59{,}833)^2]^{1/2} \approx 60{,}285.8$. Using 1.99 as the approximate 97.5th percentile in the t_{84} distribution gives the 95% CI for $price^0$, at the given values of the explanatory variables, as 336,706.7 ± 1.99(60,285.8) or, rounded to the nearest dollar, $216,738 to $456,675. This is a fairly wide prediction interval. But we have not used many factors to explain housing price. If we had more, we could, presumably, reduce the error standard deviation, and therefore $\hat{\sigma}$, to obtain a tighter prediction interval.

C6.9 (i) The estimated equation is

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

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

(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

$$\widehat{\log(wage)} = \underset{(.85)}{6.78} + \underset{(.007)}{.078}\,points + \underset{(.050)}{.218}\,exper - \underset{(.0028)}{.0071}\,exper^2 - \underset{(.035)}{.048}\,age - \underset{(.053)}{.040}\,coll$$

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

(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.

C6.10 (i) The estimated equation is

$$\widehat{\log(bwght)} = \underset{(.027)}{7.958} + \underset{(.0037)}{.0189}\,npvis - \underset{(.00012)}{.00043}\,npvis^2$$

$n = 1{,}764,\ R^2 = .0213,\ \bar{R}^2 = .0201$

The quadratic term is very significant as its *t* statistic is above 3.5 in absolute value.

(ii) The turning point calculation is by now familiar: $npvis^* = .0189/[2(.00043)] \approx 21.97$, or about 22. In the sample, 89 women had 22 or more prenatal visits.

(iii) While prenatal visits are a good thing for helping to prevent low birth weight, a woman's having many prenatal visits is a possible indicator of a pregnancy with difficulties. So it does make sense that the quadratic has a hump shape, provided we do not interpret the turnaround as implying that too many visits actually *causes* low birth weight.

(iv) With *mage* added in quadratic form, we get

$$\widehat{\log(bwght)} = \underset{(.137)}{7.584} + \underset{(.0037)}{.0180\, npvis} - \underset{(.00012)}{.00041\, npvis^2} + \underset{(.0093)}{.0254\, mage} - \underset{(.00015)}{.00041\, mage^2}$$

$n = 1{,}764$, $R^2 = .0256$, $\bar{R}^2 = .0234$

The birth weight is maximized at $mage \approx 31$. 746 women are at least 31 years old; 605 are at least 32.

(v) These variables explain on the order of 2.6% of the variation in log(*bwght*) (or even less according to $\bar{R}^2$), which is not very much.

(vi) If we regress *bwght* on *npvis*, $npvis^2$, *mage*, and $mage^2$, then $R^2 = .0192$. But remember, we cannot compare this directly with the *R*-squared reported in part (iv). Instead, we compute an *R*-squared for the log(*bwght*) model that can be compared with .0192. From Section 6.4, we compute the squared correlation between *bwght* and $\exp(\widehat{lbwght})$, where $\widehat{lbwght}$ denotes the fitted values from the log(*bwght*) model. The correlation is .1362, so its square is about .0186. Therefore, for explaining *bwght*, the model with *bwght* as the dependent variable actually fits slightly better (but nothing to make a big deal about).

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.

C6.12 (i) The youngest age is 25, and there are 99 people of this age in the sample with *fsize* = 1.

(ii) One literal interpretation is that β_2 is the increase in *nettfa* when *age* increases by one year, holding fixed *inc* and age^2. Of course, it makes no sense to change *age* while keeping age^2 fixed. Alternatively, because $\partial nettfa / \partial age = \beta_2 + 2\beta_3 age$, β_2 is the approximate increase in *nettfa* when *age* increases from zero to one. But in this application, the partial effect starting at *age* = 0 is not interesting; the sample represents single people at least 25 years old.

(iii) The OLS estimates are

$$\widehat{nettfa} = \underset{(15.28)}{-1.20} + \underset{(.060)}{.825\, inc} - \underset{(0.767)}{1.322\, age} + \underset{(.0090)}{.0256\, age^2}$$

$$n = 2{,}017,\ R^2 = .1229,\ \bar{R}^2 = .1216$$

Initially, the negative coefficient on *age* may seem counterintuitive. The estimated relationship is a U-shape, but, to make sense of it, we need to find the turning point in the quadratic. From equation (6.13), the estimated turning point is 1.322/[2(.0256)] ≈ 25.8. Interestingly, this is very close to the youngest age in the sample. In other words, starting at roughly *age* = 25, the relationship between *nettfa* and *age* is positive – as we might expect. So, in this case, the negative coefficient on *age* makes sense when we compute the partial effect.

(iv) I follow the hint, form the new regressor $(age - 25)^2$, and run the regression *nettfa* on *inc*, *age*, and $(age - 25)^2$. This changes the intercept (which we are not concerned with, anyway) and the coefficient on *age*, which is simply $\beta_2 + 2\beta_3(25)$ – the partial effect evaluated at *age* = 25 . The results are

$$\widehat{nettfa} = \underset{(9.97)}{-17.20} + \underset{(.060)}{.825\,inc} - \underset{(.767)}{.0437\,age} + \underset{(.0090)}{.0256\,(age-25)^2}$$

$$n = 2{,}017,\ R^2 = .1229,\ \bar{R}^2 = .1216$$

Therefore, the estimated partial effect starting at *age* = 25 is only –.044 and very statistically insignificant ($t = -.13$). The two-sided *p*-value is about .89.

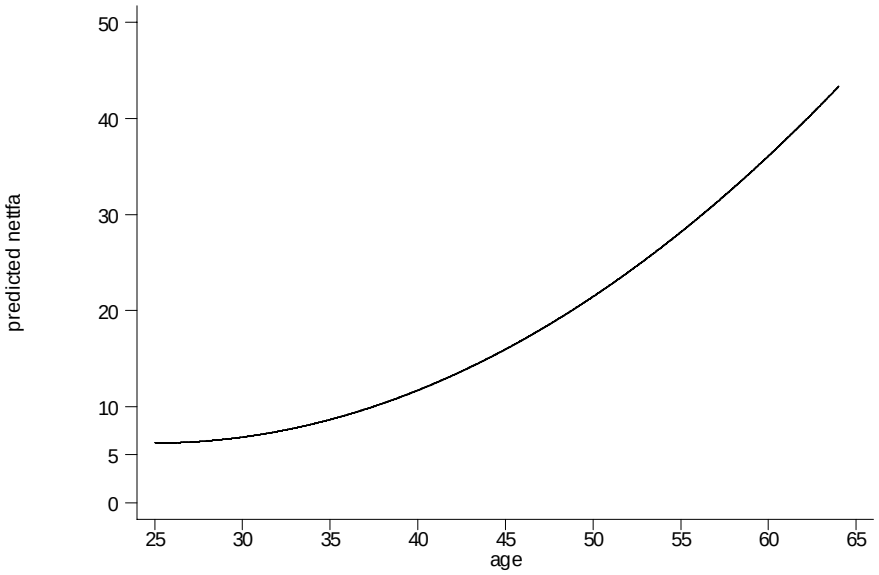
(v) If we drop *age* from the regression in part (iv) we get

$$\widehat{nettfa} = \underset{(2.18)}{-18.49} + \underset{(.060)}{.824\,inc} + \underset{(.0025)}{.0244\,(age-25)^2}$$

$$n = 2{,}017,\ R^2 = .1229,\ \bar{R}^2 = .1220$$

The adjusted *R*-squared is slightly higher when we drop *age*. But the real reason for dropping *age* is that its *t* statistic is quite small, and the model without it has a straightforward interpretation.

(vi) The graph of the relationship estimated in (v), with *inc* = 30, is



The slope of the relationship between $\widehat{nettfa}$ and *age* is clearly increasing. That is, there is an increasing marginal effect. The model is constructed so that the slope is zero at *age* = 25; from there, the slope increases.

(vii) When inc^2 is added to the regression in part (v) its coefficient is only –.00054 with t = –0.27. Thus, the linear relationship between *nettfa* and *inc* is not rejected, and we would exclude the squared income term.

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.

CHAPTER 7

TEACHING NOTES

This is a fairly standard chapter on using qualitative information in regression analysis, although I try to emphasize examples with policy relevance (and only cross-sectional applications are included.).

In allowing for different slopes, it is important, as in Chapter 6, to appropriately interpret the parameters and to decide whether they are of direct interest. For example, in the wage equation where the return to education is allowed to depend on gender, the coefficient on the female dummy variable is the wage differential between women and men at zero years of education. It is not surprising that we cannot estimate this very well, nor should we want to. In this particular example we would drop the interaction term because it is insignificant, but the issue of interpreting the parameters can arise in models where the interaction term is significant.

In discussing the Chow test, I think it is important to discuss testing for differences in slope coefficients after allowing for an intercept difference. In many applications, a significant Chow statistic simply indicates intercept differences. (See the example in Section 7.4 on student-athlete GPAs in the text.) From a practical perspective, it is important to know whether the partial effects differ across groups or whether a constant differential is sufficient.

I admit that an unconventional feature of this chapter is its introduction of the linear probability model. I cover the LPM here for several reasons. First, the LPM is being used more and more because it is easier to interpret than probit or logit models. Plus, once the proper parameter scalings are done for probit and logit, the estimated effects are often similar to the LPM partial effects near the mean or median values of the explanatory variables. The theoretical drawbacks of the LPM are often of secondary importance in practice. Computer Exercise C7.9 is a good one to illustrate that, even with over 9,000 observations, the LPM can deliver fitted values strictly between zero and one for all observations.

If the LPM is not covered, many students will never know about using econometrics to explain qualitative outcomes. This would be especially unfortunate for students who might need to read an article where an LPM is used, or who might want to estimate an LPM for a term paper or senior thesis. Once they are introduced to purpose and interpretation of the LPM, along with its shortcomings, they can tackle nonlinear models on their own or in a subsequent course.

A useful modification of the LPM estimated in equation (7.29) is to drop *kidsge6* (because it is not significant) and then define two dummy variables, one for *kidslt6* equal to one and the other for *kidslt6* at least two. These can be included in place of *kidslt6* (with no young children being the base group). This allows a diminishing marginal effect in an LPM. I was a bit surprised when a diminishing effect did not materialize.

SOLUTIONS TO PROBLEMS

7.1 (i) The coefficient on *male* is 87.75, so a man is estimated to sleep almost one and one-half hours more per week than a comparable woman. Further, $t_{male} = 87.75/34.33 \approx 2.56$, which is close to the 1% critical value against a two-sided alternative (about 2.58). Thus, the evidence for a gender differential is fairly strong.

(ii) The *t* statistic on *totwrk* is $-.163/.018 \approx -9.06$, which is very statistically significant. The coefficient implies that one more hour of work (60 minutes) is associated with $.163(60) \approx 9.8$ minutes less sleep.

(iii) To obtain R_r^2, the *R*-squared from the restricted regression, we need to estimate the model without *age* and age^2. When *age* and age^2 are both in the model, *age* has no effect only if the parameters on both terms are zero.

7.2 (i) If $\Delta cigs = 10$ then $\widehat{\Delta \log(bwght)} = -.0044(10) = -.044$, which means about a 4.4% lower birth weight.

(ii) A white child is estimated to weigh about 5.5% more, other factors in the first equation fixed. Further, $t_{white} \approx 4.23$, which is well above any commonly used critical value. Thus, the difference between white and nonwhite babies is also statistically significant.

(iii) If the mother has one more year of education, the child's birth weight is estimated to be .3% higher. This is not a huge effect, and the *t* statistic is only one, so it is not statistically significant.

(iv) The two regressions use different sets of observations. The second regression uses fewer observations because *motheduc* or *fatheduc* are missing for some observations. We would have to reestimate the first equation (and obtain the *R*-squared) using the same observations used to estimate the second equation.

7.3 (i) The *t* statistic on $hsize^2$ is over four in absolute value, so there is very strong evidence that it belongs in the equation. We obtain this by finding the turnaround point; this is the value of *hsize* that maximizes $\hat{sat}$ (other things fixed): $19.3/(2 \cdot 2.19) \approx 4.41$. Because *hsize* is measured in hundreds, the optimal size of graduating class is about 441.

(ii) This is given by the coefficient on *female* (since *black* = 0): nonblack females have SAT scores about 45 points lower than nonblack males. The *t* statistic is about –10.51, so the difference is very statistically significant. (The very large sample size certainly contributes to the statistical significance.)

(iii) Because *female* = 0, the coefficient on *black* implies that a black male has an estimated SAT score almost 170 points less than a comparable nonblack male. The *t* statistic is over 13 in absolute value, so we easily reject the hypothesis that there is no ceteris paribus difference.

(iv) We plug in *black* = 1, *female* = 1 for black females and *black* = 0 and *female* = 1 for nonblack females. The difference is therefore –169.81 + 62.31 = –107.50. Because the estimate depends on two coefficients, we cannot construct a *t* statistic from the information given. The easiest approach is to define dummy variables for three of the four race/gender categories and choose nonblack females as the base group. We can then obtain the *t* statistic we want as the coefficient on the black female dummy variable.

7.4 (i) The approximate difference is just the coefficient on *utility* times 100, or –28.3%. The *t* statistic is –.283/.099 ≈ –2.86, which is very statistically significant.

(ii) $100 \cdot [\exp(-.283) - 1) \approx -24.7\%$, and so the estimate is somewhat smaller in magnitude.

(iii) The proportionate difference is .181 – .158 = .023, or about 2.3%. One equation that can be estimated to obtain the standard error of this difference is

$$\log(salary) = \beta_0 + \beta_1 \log(sales) + \beta_2 roe + \delta_1 consprod + \delta_2 utility + \delta_3 trans + u,$$

where *trans* is a dummy variable for the transportation industry. Now, the base group is *finance*, and so the coefficient δ_1 directly measures the difference between the consumer products and finance industries, and we can use the *t* statistic on *consprod*.

7.5 (i) Following the hint, $\widehat{colGPA} = \hat{\beta}_0 + \hat{\delta}_0 (1 - noPC) + \hat{\beta}_1 hsGPA + \hat{\beta}_2 ACT = (\hat{\beta}_0 + \hat{\delta}_0) - \hat{\delta}_0 noPC + \hat{\beta}_1 hsGPA + \hat{\beta}_2 ACT$. For the specific estimates in equation (7.6), $\hat{\beta}_0 = 1.26$ and $\hat{\delta}_0 = .157$, so the new intercept is 1.26 + .157 = 1.417. The coefficient on *noPC* is –.157.

(ii) Nothing happens to the *R*-squared. Using *noPC* in place of *PC* is simply a different way of including the same information on *PC* ownership.

(iii) It makes no sense to include both dummy variables in the regression: we cannot hold *noPC* fixed while changing *PC*. We have only two groups based on *PC* ownership so, in addition to the overall intercept, we need only to include one dummy variable. If we try to include both along with an intercept we have perfect multicollinearity (the dummy variable trap).

7.6 In Section 3.3 – in particular, in the discussion surrounding Table 3.2 – we discussed how to determine the direction of bias in the OLS estimators when an important variable (ability, in this case) has been omitted from the regression. As we discussed there, Table 3.2 only strictly holds with a single explanatory variable included in the regression, but we often ignore the presence of other independent variables and use this table as a rough guide. (Or, we can use the results of Problem 3.10 for a more precise analysis.) If less able workers are more likely to receive training, then *train* and *u* are negatively correlated. If we ignore the presence of *educ* and *exper*, or at least assume that *train* and *u* are negatively correlated after netting out *educ* and *exper*, then we can use Table 3.2: the OLS estimator of β_1 (with ability in the error term) has a downward

bias. Because we think $\beta_1 \geq 0$, we are less likely to conclude that the training program was effective. Intuitively, this makes sense: if those chosen for training had not received training, they would have lowers wages, on average, than the control group.

7.7 (i) Write the population model underlying (7.29) as

$$inlf = \beta_0 + \beta_1 nwifeinc + \beta_2 educ + \beta_3 exper + \beta_4 exper^2 + \beta_5 age + \beta_6 kidslt6 + \beta_7 kidsage6 + u,$$

plug in $inlf = 1 - outlf$, and rearrange:

$$1 - outlf = \beta_0 + \beta_1 nwifeinc + \beta_2 educ + \beta_3 exper + \beta_4 exper^2 + \beta_5 age + \beta_6 kidslt6 + \beta_7 kidsage6 + u,$$

or

$$outlf = (1 - \beta_0) - \beta_1 nwifeinc - \beta_2 educ - \beta_3 exper - \beta_4 exper^2 - \beta_5 age - \beta_6 kidslt6 - \beta_7 kidsage6 - u,$$

The new error term, $-u$, has the same properties as u. From this we see that if we regress *outlf* on all of the independent variables in (7.29), the new intercept is 1 – .586 = .414 and each slope coefficient takes on the opposite sign from when *inlf* is the dependent variable. For example, the new coefficient on *educ* is –.038 while the new coefficient on *kidslt6* is .262.

(ii) The standard errors will not change. In the case of the slopes, changing the signs of the estimators does not change their variances, and therefore the standard errors are unchanged (but the *t* statistics change sign). Also, Var$(1 - \hat{\beta}_0)$ = Var$(\hat{\beta}_0)$, so the standard error of the intercept is the same as before.

(iii) We know that changing the units of measurement of independent variables, or entering qualitative information using different sets of dummy variables, does not change the *R*-squared. But here we are changing the *dependent* variable. Nevertheless, the *R*-squareds from the regressions are still the same. To see this, part (i) suggests that the squared residuals will be identical in the two regressions. For each *i* the error in the equation for $outlf_i$ is just the negative of the error in the other equation for $inlf_i$, and the same is true of the residuals. Therefore, the SSRs are the same. Further, in this case, the total sum of squares are the same. For *outlf* we have

$$\text{SST} = \sum_{i=1}^{n}(outlf_i - \overline{outlf})^2 = \sum_{i=1}^{n}[(1 - inlf_i) - (1 - \overline{inlf})]^2 = \sum_{i=1}^{n}(-inlf_i + \overline{inlf})^2 = \sum_{i=1}^{n}(inlf_i - \overline{inlf})^2 ,$$

which is the SST for *inlf*. Because $R^2 = 1 - \text{SSR/SST}$, the *R*-squared is the same in the two regressions.

7.8 (i) We want to have a constant semi-elasticity model, so a standard wage equation with marijuana usage included would be

$$\log(wage) = \beta_0 + \beta_1 usage + \beta_2 educ + \beta_3 exper + \beta_4 exper^2 + \beta_5 female + u.$$

Then $100 \cdot \beta_1$ is the approximate percentage change in *wage* when marijuana usage increases by one time per month.

(ii) We would add an interaction term in *female* and *usage*:

$$\log(wage) = \beta_0 + \beta_1 usage + \beta_2 educ + \beta_3 exper + \beta_4 exper^2 + \beta_5 female + \beta_6 female \cdot usage + u.$$

The null hypothesis that the effect of marijuana usage does not differ by gender is H_0: $\beta_6 = 0$.

(iii) We take the base group to be nonuser. Then we need dummy variables for the other three groups: *lghtuser*, *moduser*, and *hvyuser*. Assuming no interactive effect with gender, the model would be

$$\log(wage) = \beta_0 + \delta_1 lghtuser + \delta_2 moduser + \delta_3 hvyuser + \beta_2 educ + \beta_3 exper + \beta_4 exper^2 + \beta_5 female + u.$$

(iv) The null hypothesis is H_0: $\delta_1 = 0$, $\delta_2 = 0$, $\delta_3 = 0$, for a total of $q = 3$ restrictions. If n is the sample size, the *df* in the unrestricted model – the denominator *df* in the F distribution – is $n - 8$. So we would obtain the critical value from the $F_{q,n-8}$ distribution.

(v) The error term could contain factors, such as family background (including parental history of drug abuse) that could directly affect wages and also be correlated with marijuana usage. We are interested in the effects of a person's drug usage on his or her wage, so we would like to hold other confounding factors fixed. We could try to collect data on relevant background information.

7.9 (i) Plugging in $u = 0$ and $d = 1$ gives $f_1(z) = (\beta_0 + \delta_0) + (\beta_1 + \delta_1)z$.

(ii) Setting $f_0(z^*) = f_1(z^*)$ gives $\beta_0 + \beta_1 z^* = (\beta_0 + \delta_0) + (\beta_1 + \delta_1)z^*$ or $0 = \delta_0 + \delta_1 z^*$. Therefore, provided $\delta_1 \neq 0$, we have $z^* = -\delta_0 / \delta_1$. Clearly, z^* is positive if and only if δ_0 / δ_1 is negative, which means δ_0 and δ_1 must have opposite signs.

(iii) Using part (ii) we have $totcoll^* = .357/.030 = 11.9$ years.

(iv) The estimated years of college where women catch up to men is much too high to be practically relevant. While the estimated coefficient on $female \cdot totcoll$ shows that the gap is reduced at higher levels of college, it is never closed – not even close. In fact, at four years of college, the difference in predicted log wage is still $-.357 + .030(4) = -.237$, or about 21.1% less for women.

7.10 (i) Yes, simple regression does produce an unbiased estimator of the effect of the voucher program. Because participation was randomized, we can write

$$score = \beta_0 + \beta_1 voucher + u,$$

where *voucher* is independent of *u*, that is, all other factors affecting *score*. Therefore, the key assumption for unbiasedness of simple regression, Assumption SLR.3, is satisfied.

(ii) No, we do not need to control for background variables. In the equation from part (i), these are factors in the error term, *u*. But *voucher* was assigned to be independent of all factors, including the listed background variables.

(iii) We should include the background variables to reduce the sampling error of the estimated voucher effect. By pulling background variables out of the error term, we reduce the error variance – perhaps substantially. Further, we can be sure that multicollinearity is not a problem because the key variable of interest, *voucher*, is uncorrelated with all of the added explanatory variables. (This zero correlation will only be approximate in any random sample, but in large samples it should be very small.) The one case where we would not add these variables – or, at least, when there is no benefit from doing so – is when the background variables themselves have no affect on the test score. Given the list of background variables, this seems unlikely in the current application.

SOLUTIONS TO COMPUTER EXERCISES

C7.1 (i) The estimated equation is

$$\widehat{colGPA} = \underset{(0.34)}{1.26} + \underset{(.059)}{.152\,PC} + \underset{(.094)}{.450\,hsGPA} + \underset{(.0107)}{.0077\,ACT} - \underset{(.0603)}{.0038\,mothcoll} + \underset{(.0613)}{.0418\,fathcoll}$$

$$n = 141, \quad R^2 = .222.$$

The estimated effect of *PC* is hardly changed from equation (7.6), and it is still very significant, with $t_{pc} \approx 2.58$.

(ii) The *F* test for joint significance of *mothcoll* and *fathcoll*, with 2 and 135 *df*, is about .24 with *p*-value $\approx .78$; these variables are jointly very insignificant. It is not surprising the

estimates on the other coefficients do not change much when *mothcoll* and *fathcoll* are added to the regression.

(iii) When $hsGPA^2$ is added to the regression, its coefficient is about .337 and its *t* statistic is about 1.56. (The coefficient on *hsGPA* is about –1.803.) This is a borderline case. The quadratic in *hsGPA* has a U-shape, and it only turns up at about $hsGPA^* = 2.68$, which is hard to interpret. The coefficient of main interest, on *PC*, falls to about .140 but is still significant. Adding $hsGPA^2$ is a simple robustness check of the main finding.

C7.2 (i) The estimated equation is

$$\widehat{\log(wage)} = \underset{(0.11)}{5.40} + \underset{(.0063)}{.0654}\,educ + \underset{(.0032)}{.0140}\,exper + \underset{(.0025)}{.0117}\,tenure$$
$$+ \underset{(.039)}{.199}\,married - \underset{(.038)}{.188}\,black - \underset{(.026)}{.091}\,south + \underset{(.027)}{.184}\,urban$$

$n = 935,\ R^2 = .253.$

The coefficient on *black* implies that, at given levels of the other explanatory variables, black men earn about 18.8% less than nonblack men. The *t* statistic is about –4.95, and so it is very statistically significant.

(ii) The *F* statistic for joint significance of $exper^2$ and $tenure^2$, with 2 and 925 *df*, is about 1.49 with *p*-value $\approx$ .226. Because the *p*-value is above .20, these quadratics are jointly insignificant at the 20% level.

(iii) We add the interaction $black \cdot educ$ to the equation in part (i). The coefficient on the interaction is about –.0226 (se $\approx$ .0202). Therefore, the point estimate is that the return to another year of education is about 2.3 percentage points lower for black men than nonblack men. (The estimated return for nonblack men is about 6.7%.) This is nontrivial if it really reflects differences in the population. But the *t* statistic is only about 1.12 in absolute value, which is not enough to reject the null hypothesis that the return to education does not depend on race.

(iv) We choose the base group to be single, nonblack. Then we add dummy variables *marrnonblck*, *singblck*, and *marrblck* for the other three groups. The result is

$$\widehat{\log(wage)} = \underset{(0.11)}{5.40} + \underset{(.0063)}{.0655}\,educ + \underset{(.0032)}{.0141}\,exper + \underset{(.0025)}{.0117}\,tenure$$
$$- \underset{(.026)}{.092}\,south + \underset{(.027)}{.184}\,urban + \underset{(.043)}{.189}\,marrnonblck$$
$$- \underset{(.096)}{.241}\,singblck + \underset{(.0560)}{.0094}\,marrblck$$

$n = 935$, $R^2 = .253$.

We obtain the ceteris paribus differential between married blacks and married nonblacks by taking the difference of their coefficients: .0094 – .189 = –.1796, or about –.18. That is, a married black man earns about 18% less than a comparable, married nonblack man.

C7.3 (i) $H_0: \beta_{13} = 0$. Using the data in MLB1.RAW gives $\hat{\beta}_{13} \approx .254$, se($\hat{\beta}_{13}$) $\approx .131$. The *t* statistic is about 1.94, which gives a *p*-value against a two-sided alternative of just over .05. Therefore, we would reject H_0 at just about the 5% significance level. Controlling for the performance and experience variables, the estimated salary differential between catchers and outfielders is huge, on the order of $100 \cdot [\exp(.254) - 1] \approx 28.9\%$ [using equation (7.10)].

(ii) This is a joint null, $H_0: \beta_9 = 0, \beta_{10} = 0, \ldots, \beta_{13} = 0$. The *F* statistic, with 5 and 339 *df*, is about 1.78, and its *p*-value is about .117. Thus, we cannot reject H_0 at the 10% level.

(iii) Parts (i) and (ii) are roughly consistent. The evidence against the joint null in part (ii) is weaker because we are testing, along with the marginally significant *catcher*, several other insignificant variables (especially *thrdbase* and *shrtstop*, which has absolute *t* statistics well below one).

C7.4 (i) The two signs that are pretty clear are $\beta_3 < 0$ (because *hsperc* is defined so that the smaller the number the better the student) and $\beta_4 > 0$. The effect of size of graduating class is not clear. It is also unclear whether males and females have systematically different GPAs. We may think that $\beta_6 < 0$, that is, athletes do worse than other students with comparable characteristics. But remember, we are controlling for ability to some degree with *hsperc* and *sat*.

(ii) The estimated equation is

$$\widehat{colgpa} = \underset{(0.079)}{1.241} - \underset{(.0164)}{.0569}\,hsize + \underset{(.00225)}{.00468}\,hsize^2 - \underset{(.0006)}{.0132}\,hsperc$$
$$+ \underset{(.00007)}{.00165}\,sat + \underset{(.018)}{.155}\,female + \underset{(.042)}{.169}\,athlete$$

$n = 4{,}137$, $R^2 = .293$.

Holding other factors fixed, an athlete is predicted to have a GPA about .169 points *higher* than a nonathlete. The *t* statistic .169/.042 $\approx$ 4.02, which is very significant.

(iii) With *sat* dropped from the model, the coefficient on *athlete* becomes about .0054 (se $\approx$ .0448), which is practically and statistically not different from zero. This happens because we do not control for SAT scores, and athletes score lower on average than nonathletes. Part (ii) shows that, once we account for SAT differences, athletes do better than nonathletes. Even if we do not control for SAT score, there is no difference.

(iv) To facilitate testing the hypothesis that there is no difference between women athletes and women nonathletes, we should choose one of these as the base group. We choose female nonathletes. The estimated equation is

$$\widehat{colgpa} = \underset{(0.076)}{1.396} - \underset{(.0164)}{.0568\, hsize} + \underset{(.00225)}{.00467\, hsize^2} - \underset{(.0006)}{.0132\, hsperc}$$

$$+ \underset{(.00007)}{.00165\, sat} + \underset{(.084)}{.175\, femath} + \underset{(.049)}{.013\, maleath} - \underset{(.018)}{.155\, malenonath}$$

$n = 4{,}137, \quad R^2 = .293.$

The coefficient on *femath* = *female*$\cdot$*athlete* shows that *colgpa* is predicted to be about .175 points higher for a female athlete than a female nonathlete, other variables in the equation fixed. The hypothesis that there is no difference between female athletes and female nonathletes is testing by using the *t* statistic on *femath*. In this case, $t = 2.08$, which is statistically significant at the 5% level against a two-sided alternative.

(v) Whether we add the interaction *female*$\cdot$*sat* to the equation in part (ii) or part (iv), the outcome is practically the same. For example, when *female*$\cdot$*sat* is added to the equation in part (ii), its coefficient is about .000051 and its *t* statistic is about .40. There is very little evidence that the effect of *sat* differs by gender.

C7.5 The estimated equation is

$$\widehat{\log(salary)} = \underset{(0.29)}{4.30} + \underset{(.034)}{.288 \log(sales)} + \underset{(.0040)}{.0167\, roe} - \underset{(.109)}{.226\, rosneg}$$

$n = 209, \quad R^2 = .297, \quad \bar{R}^2 = .286.$

The coefficient on *rosneg* implies that if the CEO's firm had a negative return on its stock over the 1988 to 1990 period, the CEO salary was predicted to be about 22.6% lower, for given levels of *sales* and *roe*. The *t* statistic is about –2.07, which is significant at the 5% level against a two-sided alternative.

C7.6 (i) The estimated equation for men is

$$\widehat{sleep} = \underset{(310.0)}{3{,}648.2} - \underset{(.024)}{.182\, totwrk} - \underset{(7.41)}{13.05\, educ} + \underset{(14.32)}{7.16\, age} - \underset{(.1684)}{.0448\, age^2} + \underset{(59.02)}{60.38\, yngkid}$$

$n = 400, \quad R^2 = .156$

and the estimated equation for women is

$$\widehat{sleep} = \underset{(384.9)}{4{,}238.7} - \underset{(.028)}{.140\, totwrk} - \underset{(9.59)}{10.21\, educ} - \underset{(18.53)}{30.36\, age} - \underset{(.223)}{.368\, age^2} - \underset{(93.19)}{118.28\, yngkid}$$

$n = 306, \quad R^2 = .098.$

There are certainly notable differences in the point estimates. For example, having a young child in the household leads to less sleep for women (about two hours a week) while men are estimated to sleep about an hour more. The quadratic in *age* is a hump-shape for men but a U-shape for women. The intercepts for men and women are also notably different.

(ii) The F statistic (with 6 and 694 df) is about 2.12 with p-value $\approx$ .05, and so we reject the null that the sleep equations are the same at the 5% level.

(iii) If we leave the coefficient on *male* unspecified under H_0, and test only the five interaction terms, $male \cdot totwrk$, $male \cdot educ$, $male \cdot age$, $male \cdot age^2$, and $male \cdot yngkid$, the F statistic (with 5 and 694 df) is about 1.26 and p-value $\approx$ .28.

(iv) The outcome of the test in part (iii) shows that, once an intercept difference is allowed, there is not strong evidence of slope differences between men and women. This is one of those cases where the practically important differences in estimates for women and men in part (i) do not translate into statistically significant differences. We need a larger sample size to confidently determine whether there are differences in slopes. For the purposes of studying the sleep-work tradeoff, the original model with *male* added as an explanatory variable seems sufficient.

C7.7 (i) When *educ* = 12.5, the approximate proportionate difference in estimated *wage* between women and men is –.227 – .0056(12.5) = –.297. When *educ* = 0, the difference is –.227. So the differential at 12.5 years of education is about 7 percentage points greater.

(ii) We can write the model underlying (7.18) as

$$\begin{aligned}\log(wage) &= \beta_0 + \delta_0\, female + \beta_1 educ + \delta_1\, female \cdot educ + other\ factors \\ &= \beta_0 + (\delta_0 + 12.5\delta_1)\, female + \beta_1 educ + \delta_1\, female \cdot (educ - 12.5) \\ &\quad + other\ factors \\ &\equiv \beta_0 + \theta_0\, female + \beta_1 educ + \delta_1\, female \cdot (educ - 12.5) + other\ factors,\end{aligned}$$

where $\theta_0 \equiv \delta_0 + 12.5\delta_1$ is the gender differential at 12.5 years of education. When we run this regression we obtain about –.294 as the coefficient on *female* (which differs from –.297 due to rounding error). Its standard error is about .036.

(iii) The *t* statistic on *female* from part (ii) is about –8.17, which is very significant. This is because we are estimating the gender differential at a reasonable number of years of education, 12.5, which is close to the average. In equation (7.18), the coefficient on *female* is the gender differential when *educ* = 0. There are no people of either gender with close to zero years of education, and so we cannot hope – nor do we want to – to estimate the gender differential at *educ* = 0.

C7.8 (i) If the appropriate factors have been controlled for, $\beta_1 > 0$ signals discrimination against minorities: a white person has a greater chance of having a loan approved, other relevant factors fixed.

(ii) The simple regression results are

$$\widehat{approve} = \underset{(.018)}{.708} + \underset{(.020)}{.201}\, white$$

$$n = 1{,}989, \quad R^2 = .049.$$

The coefficient on *white* means that, in the sample of 1,989 loan applications, an application submitted by a white application was 20.1% more likely to be approved than that of a nonwhite applicant. This is a practically large difference and the *t* statistic is about 10. (We have a large sample size, so standard errors are pretty small.)

(iii) When we add the other explanatory variables as controls, we obtain $\hat{\beta}_1 \approx .129$, $se(\hat{\beta}_1) \approx .020$. The coefficient has fallen by some margin because we are now controlling for factors that should affect loan approval rates, and some of these clearly differ by race. (On average, white people have financial characteristics – such as higher incomes and stronger credit histories – that make them better loan risks.) But the race effect is still strong and very significant (*t* statistic ≈ 6.45).

(iv) When we add the interaction *white*·*obrat* to the regression, its coefficient and *t* statistic are about .0081 and 3.53, respectively. Therefore, there is an interactive effect: a white applicant is penalized less than a nonwhite applicant for having other obligations as a larger percent of income.

(v) The trick should be familiar by now. Replace *white*·*obrat* with *white*·(*obrat* – 32); the coefficient on *white* is now the race differential when *obrat* = 32. We obtain about .113 and se ≈ .020. So the 95% confidence interval is about .113 ± 1.96(.020) or about .074 to .152. Clearly, this interval excludes zero, so at the average *obrat* there is evidence of discrimination (or, at least loan approval rates that differ by race for some other reason that is not captured by the control variables).

C7.9 (i) About .392, or 39.2%.

(ii) The estimated equation is

$$\widehat{e401k} = \underset{(.081)}{-.506} + \underset{(.0006)}{.0124}\,inc - \underset{(.000005)}{.000062}\,inc^2 + \underset{(.0039)}{.0265}\,age - \underset{(.00005)}{.00031}\,age^2 - \underset{(.0121)}{.0035}\,male$$

$n = 9{,}275$, $R^2 = .094$.

(iii) 401(k) eligibility clearly depends on income and age in part (ii). Each of the four terms involving *inc* and *age* have very significant *t* statistics. On the other hand, once income and age are controlled for, there seems to be no difference in eligibility by gender. The coefficient on *male* is very small – at given income and age, males are estimated to have a .0035 lower probability of being 401(k) eligible – and it has a very small *t* statistic.

(iv) Somewhat surprisingly, out of 9,275 fitted values, none is outside the interval [0,1]. The smallest fitted value is about .030 and the largest is about .697. This means one theoretical problem with the LPM – the possibility of generating silly probability estimates – does not materialize in this application.

(v) Using the given rule, 2,460 families are predicted to be eligible for a 401(k) plan.

(vi) Of the 5,638 families actually ineligible for a 401(k) plan, about 81.7 are correctly predicted not to be eligible. Of the 3,637 families actually eligible, only 39.3 percent are correctly predicted to be eligible.

(vii) The overall percent correctly predicted is a weighted average of the two percentages obtained in part (vi). As we saw there, the model does a good job of predicting when a family is ineligible. Unfortunately, it does less well – predicting correctly less than 40% of the time – in predicting that a family is eligible for a 401(k).

(viii) The estimated equation is

$$\widehat{e401k} = \underset{(.081)}{-.502} + \underset{(.0006)}{.0123\, inc} - \underset{(.000005)}{.000061\, inc^2} + \underset{(.0039)}{.0265\, age} - \underset{(.00005)}{.00031\, age^2}$$

$$\underset{(.0121)}{-.0038\, male} + \underset{(.0122)}{.0198\, pira}$$

$n = 9{,}275, \quad R^2 = .095.$

The coefficient on *pira* means that, other things equal, IRA ownership is associated with about a .02 higher probability of being eligible for a 401(k) plan. However, the *t* statistic is only about 1.62, which gives a two-sided *p*-value = .105. So *pira* is not significant at the 10% level against a two-sided alternative.

C7.10 (i) The estimated equation is

$$\widehat{points} = \underset{(1.18)}{4.76} + \underset{(.33)}{1.28\, exper} - \underset{(.024)}{.072\, exper^2} + \underset{(1.00)}{2.31\, guard} + \underset{(1.00)}{1.54\, forward}$$

$n = 269, \quad R^2 = .091, \quad \overline{R}^2 = .077.$

(ii) Including all three position dummy variables would be redundant, and result in the dummy variable trap. Each player falls into one of the three categories, and the overall intercept is the intercept for centers.

(iii) A guard is estimated to score about 2.3 points more per game, holding experience fixed. The *t* statistic is 2.31, so the difference is statistically different from zero at the 5% level, against a two-sided alternative.

(iv) When *marr* is added to the regression, its coefficient is about .584 (se = .740). Therefore, a married player is estimated to score just over half a point more per game (experience and position held fixed), but the estimate is not statistically different from zero (*p*-value = .43). So, based on points per game, we cannot conclude married players are more productive.

(v) Adding the terms $marr \cdot exper$ and $marr \cdot exper^2$ leads to complicated signs on the three terms involving *marr*. The *F* test for their joint significance, with 3 and 261 *df*, gives $F = 1.44$ and *p*-value = .23. Therefore, there is not very strong evidence that marital status has any partial effect on points scored.

(vi) If in the regression from part (iv) we use *assists* as the dependent variable, the coefficient on *marr* becomes .322 (se = .222). Therefore, holding experience and position fixed, a married man has almost one-third more assist per game. The *p*-value against a two-sided alternative is about .15, which is stronger, but not overwhelming, evidence that married men are more productive when it comes to assists.

C7.11 (i) The average is 19.072, the standard deviation is 63.964, the smallest value is –502.302, and the largest value is 1,536.798. Remember, these are in thousands of dollars.

(ii) This can be easily done by regressing *nettfa* on *e401k* and doing a *t* test on $\hat{\beta}_{e401k}$; the estimate is the average difference in *nettfa* for those eligible for a 401(k) and those not eligible. Using the 9,275 observations gives $\hat{\beta}_{e401k} = 18.858$ and $t_{e401k} = 14.01$. Therefore, we strongly reject the null hypothesis that there is no difference in the averages. The coefficient implies that, on average, a family eligible for a 401(k) plan has $18,858 more on net total financial assets.

(iii) The equation estimated by OLS is

$$\widehat{nettfa} = \underset{(9.96)}{23.09} + \underset{(1.277)}{9.705}\, e401k - \underset{(.075)}{.278}\, inc + \underset{(.0006)}{.0103}\, inc^2 - \underset{(.483)}{1.972}\, age + \underset{(.0055)}{.0348}\, age^2$$

$n = 9{,}275,\ R^2 = .202$

Now, holding income and age fixed, a 401(k)-eligible family is estimated to have $9,705 more in wealth than a non-eligible family. This is just more than half of what is obtained by simply comparing averages.

(iv) Only the interaction $e401k\cdot(age - 41)$ is significant. Its coefficient is .654 ($t = 4.98$). It shows that the effect of 401(k) eligibility on financial wealth increases with age. Another way to think about it is that *age* has a stronger positive effect on *nettfa* for those with 401(k) eligibility. The coefficient on $e401k\cdot(age - 41)^2$ is –.0038 (*t* statistic = –.33), so we could drop this term.

(v) The effect of e401k in part (iii) is the same for all ages, 9.705. For the regression in part (iv), the coefficient on *e401k* from part (iv) is about 9.960, which is the effect at the average age, *age* = 41. Including the interactions increases the estimated effect of *e401k*, but only by $255. If we evaluate the effect in part (iv) at a wide range of ages, we would see more dramatic differences.

(vi) I chose *fsize1* as the base group. The estimated equation is

$$\widehat{nettfa} = \underset{(10.12)}{16.34} + \underset{(1.278)}{9.455}\, e401k - \underset{(.075)}{.240}\, inc + \underset{(.0006)}{.0100}\, inc^2 - \underset{(.483)}{1.495}\, age + \underset{(.0055)}{.0290}\, age^2$$

$$- \underset{(1.818)}{.859}\, fsize2 - \underset{(1.877)}{4.665}\, fsize3 - \underset{(1.868)}{6.314}\, fsize4 - \underset{(2.101)}{7.361}\, fsize5$$

$n = 9{,}275,\ R^2 = .204,\ SSR = 30{,}215{,}207.5$

The *F* statistic for joint significance of the four family size dummies is about 5.44. With 4 and 9,265 *df*, this gives *p*-value = .0002. So the family size dummies are jointly significant.

(vii) The SSR for the restricted model is from part (vi): $SSR_r = 30{,}215{,}207.5$. The SSR for the unrestricted model is obtained by adding the SSRs for the five separate family size regressions. I get $SSR_{ur} = 29{,}985{,}400$. The Chow statistic is $F = [(30{,}215{,}207.5 - 29{,}985{,}400)/29{,}985{,}400]*(9245/20) \approx 3.54$. With 20 and 9,245 *df*, the *p*-value is essentially zero. In this case, there is strong evidence that the slopes change across family size. Allowing for intercept changes alone is not sufficient. (If you look at the individual regressions, you will see that the signs on the income variables actually change across family size.)

C7.12 (i) For women, the fraction rated as having above average looks is about .33; for men, it is .29. The proportion of women rated as having below average looks is only .135; for men, it is even lower at about .117.

(ii) The difference is about .04, that is, the percent rated as having above average looks is about four percentage points higher for women than men. A simple way to test whether the difference is statistically significant is to run a simple regression of *abvavg* on *female* and do a *t* test (which is asymptotically valid). The *t* statistic is about 1.48 with two-sided *p*-value = .14. Therefore, there is not strong evidence against the null that the population fractions are the same, but there is some evidence.

(iii) The regression for men is

$$\widehat{\log(wage)} = \underset{(0.024)}{1.884} - \underset{(.060)}{.199}\, belavg - \underset{(.042)}{.044}\, abvavg$$

$$n = 824 \quad R^2 = .013$$

and the regression for women is

$$\widehat{\log(wage)} = \underset{(0.034)}{1.309} - \underset{(.076)}{.138}\, belavg + \underset{(.055)}{.034}\, abvavg$$

$$n = 436 \quad R^2 = .011.$$

Using the standard approximation, a man with below average looks earns almost 20% less than a man of average looks, and a woman with below average looks earns about 13.8% less than a woman with average looks. (The more accurate estimates are about 18% and 12.9%, respectively.) The null hypothesis H_0: $\beta_1 = 0$ against H_1: $\beta_1 < 0$ means that the null is that people with below average looks earn the same, on average, as people with average looks; the alternative is that people with below average looks earn less than people with average looks (in the population). The one-sided *p*-value for men is .0005 and for women it is .036. We reject H_0 more strongly for men because the estimate is larger in magnitude and the estimate has less sampling variation (as measured by the standard error).

(iv) Women with above average looks are estimated to earn about 3.4% more, on average, than women with average looks. But the one-sided *p*-value is .272, and this provides very little evidence against H_0: $\beta_2 = 0$.

(v) Given the number of added controls, with many of them very statistically significant, the coefficients on the looks variables do not change by much. For men, the coefficient on *belavg* becomes –.143 ($t = -2.80$) and the coefficient on *abvavg* becomes –.001 ($t = -.03$). For women, the changes in magnitude are similar: the coefficient on *belavg* becomes –.115 ($t = -1.75$) and the coefficient on *abvavg* becomes .058 ($t = 1.18$). In both cases, the estimates on *belavg* move closer to zero but are still reasonably large.

C7.13 (i) 412/660 ≈ .624.

(ii) The OLS estimates of the LPM are

$$\widehat{ecobuy} = \underset{(.165)}{.424} - \underset{(.109)}{.803\, ecoprc} + \underset{(.132)}{.719\, regprc} + \underset{(.00053)}{.00055\, faminc} + \underset{(.013)}{.024\, hhsize}$$

$$+ \underset{(.008)}{.025\, educ} - \underset{(.00125)}{.00050\, age}$$

$n = 660$, $R^2 = .110$

If *ecoprc* increases by, say, 10 cents (.10), then the probability of buying eco-labeled apples falls by about .080. If *regprc* increases by 10 cents, the probability of buying eco-labeled apples increases by about .072. (Of course, we are assuming that the probabilities are not close to the boundaries of zero and one, respectively.)

(iii) The *F* test, with 4 and 653 *df*, is 4.43, with *p*-value = .0015. Thus, based on the usual *F* test, the four non-price variables are jointly very significant. Of the four variables, *educ* appears to have the most important effect. For example, a difference of four years of education implies an increase of .025(4) = .10 in the estimated probability of buying eco-labeled apples. This suggests that more highly educated people are more open to buying produce that is environmentally friendly, which is perhaps expected. Household size (*hhsize*) also has an effect. Comparing a couple with two children to one that has no children – other factors equal – the couple with two children has a .048 higher probability of buying eco-labeled apples.

(iv) The model with log(*faminc*) fits the data slightly better: the *R*-squared increases to about .112. (We would not expect a large increase in *R*-squared from a simple change in the functional form.) The coefficient on log(*faminc*) is about .045 ($t = 1.55$). If log(*faminc*) increases by .10, which means roughly a 10% increase in *faminc*, then P(*ecobuy* = 1) is estimated to increase by about .0045, a pretty small effect.

(v) The fitted probabilities range from about .185 to 1.051, so none are negative. There are two fitted probabilities above 1, which is not a source of concern with 660 observations.

(vi) Using the standard prediction rule – predict one when $\widehat{ecobuy}_i \geq .5$ and zero otherwise – gives the fraction correctly predicted for *ecobuy* = 0 as 102/248 ≈ .411, so about 41.1%. For *ecobuy* = 1, the fraction correctly predicted is 340/412 ≈ .825, or 82.5%. With the usual prediction rule, the model does a much better job predicting the decision to buy eco-labeled apples. (The overall percent correctly predicted is about 67%.)

C7.14 (i) The estimated LPM is

$$\widehat{respond} = \underset{(.009)}{.282} + \underset{(.015)}{.344\ resplast} + \underset{(.00009)}{.00015\ avggift}$$

$n = 4{,}268,\ R^2 = .110$

Holding the average gift fixed, the probability of a current response is estimated to be .344 higher if the person responded most recently.

(ii) Once we control for responding most recently, the effect of of *avggift* is very small. Even if *avggift* is 100 guilders more (the mean is about 18.2 with standard deviation 78.7), the probability of responding this period is only .015 higher. Plus, the *t* statistic has a two-sided *p*-value of about .09, so it is only marginally statistically significant.

(iii) The coefficient on *propresp* is about .747 (standard error = .034). If *propresp* increases by .1 (for example, from .4 to .5), the probability of responding is about .075 higher.

(iv) When *propresp* is added to the regression, the coefficient on *resplast* falls to about .095 (although it is still very statistically significant). This makes sense, because the relationship between responding currently and responding most recently should be weaker once the average response is controlled for. Certainly *resplast* and *propresp* are positively correlated.

(v) The coefficient on *mailsyear* is about .062 (*t* = 6.18). This is a reasonably large effect: each new mailing is estimated to increase the probability of responding by .062. Unfortunately, we do not know how the charitable organization determines the mailings sent. To the extent that it depends only on past gift giving, as controlled for by the average gift, the most recent response, and the response rate, the estimate could be a good (consistent) estimate of the causal effect. But if mailings are determined by other factors that are necessarily in the error term – such as income – then the estimate would be systematically biased. If, say, more mailings are sent to people with higher incomes, and higher income people are more likely to respond, then the regression that omits income produces an upward bias for the *mailsyear* coefficient.

CHAPTER 8

TEACHING NOTES

This is a good place to remind students that homoskedasticity played no role in showing that OLS is unbiased for the parameters in the regression equation. In addition, you probably should mention that there is nothing wrong with the *R*-squared or adjusted *R*-squared as goodness-of-fit measures. The key is that these are estimates of the population *R*-squared, $1 - [\text{Var}(u)/\text{Var}(y)]$, where the variances are the *unconditional* variances in the population. The usual *R*-squared, and the adjusted version, consistently estimate the population *R*-squared whether or not $\text{Var}(u|\mathbf{x}) = \text{Var}(y|\mathbf{x})$ depends on $\mathbf{x}$. Of course, heteroskedasticity causes the usual standard errors, *t* statistics, and *F* statistics to be invalid, even in large samples, with or without normality.

By explicitly stating the homoskedasticity assumption as conditional on the explanatory variables that appear in the conditional mean, it is clear that only heteroskedasticity that depends on the explanatory variables in the model affects the validity of standard errors and test statistics. The version of the Breusch-Pagan test in the text, and the White test, are ideally suited for detecting forms of heteroskedasticity that invalidate inference obtained under homoskedasticity. If heteroskedasticity depends on an exogenous variable that does not also appear in the mean equation, this can be exploited in weighted least squares for efficiency, but only rarely is such a variable available. One case where such a variable is available is when an individual-level equation has been aggregated. I discuss this case in the text but I rarely have time to teach it.

As I mention in the text, other traditional tests for heteroskedasticity, such as the Park and Glejser tests, do not directly test what we want, or add too many assumptions under the null. The Goldfeld-Quandt test only works when there is a natural way to order the data based on one independent variable. This is rare in practice, especially for cross-sectional applications.

Some argue that weighted least squares estimation is a relic, and is no longer necessary given the availability of heteroskedasticity-robust standard errors and test statistics. While I am sympathetic to this argument, it presumes that we do not care much about efficiency. Even in large samples, the OLS estimates may not be precise enough to learn much about the population parameters. With substantial heteroskedasticity we might do better with weighted least squares, even if the weighting function is misspecified. As discussed in the text on pages 287-288, one can, and probably should, compute robust standard errors after weighted least squares. For asymptotic efficiency comparisons, these would be directly comparable to the heteroskedasiticity-robust standard errors for OLS.

Weighted least squares estimation of the LPM is a nice example of feasible GLS, at least when all fitted values are in the unit interval. Interestingly, in the LPM examples in the text and the LPM computer exercises, the heteroskedasticity-robust standard errors often differ by only small amounts from the usual standard errors. However, in a couple of cases the differences are notable, as in Computer Exercise C8.7.

SOLUTIONS TO PROBLEMS

8.1 Parts (ii) and (iii). The homoskedasticity assumption played no role in Chapter 5 in showing that OLS is consistent. But we know that heteroskedasticity causes statistical inference based on the usual *t* and *F* statistics to be invalid, even in large samples. As heteroskedasticity is a violation of the Gauss-Markov assumptions, OLS is no longer BLUE.

8.2 With $\text{Var}(u|inc,price,educ,female) = \sigma^2 inc^2$, $h(\mathbf{x}) = inc^2$, where $h(\mathbf{x})$ is the heteroskedasticity function defined in equation (8.21). Therefore, $\sqrt{h(\mathbf{x})} = inc$, and so the transformed equation is obtained by dividing the original equation by *inc*:

$$\frac{beer}{inc} = \beta_0(1/inc) + \beta_1 + \beta_2(price/inc) + \beta_3(educ/inc) + \beta_4(female/inc) + (u/inc).$$

Notice that β_1, which is the slope on *inc* in the original model, is now a constant in the transformed equation. This is simply a consequence of the form of the heteroskedasticity and the functional forms of the explanatory variables in the original equation.

8.3 False. The unbiasedness of WLS and OLS hinges crucially on Assumption MLR.4, and, as we know from Chapter 4, this assumption is often violated when an important variable is omitted. When MLR.4 does not hold, both WLS and OLS are biased. Without specific information on how the omitted variable is correlated with the included explanatory variables, it is not possible to determine which estimator has a small bias. It is possible that WLS would have more bias than OLS or less bias. Because we cannot know, we should not claim to use WLS in order to solve "biases" associated with OLS.

8.4 (i) These coefficients have the anticipated signs. If a student takes courses where grades are, on average, higher – as reflected by higher *crsgpa* – then his/her grades will be higher. The better the student has been in the past – as measured by *cumgpa* – the better the student does (on average) in the current semester. Finally, *tothrs* is a measure of experience, and its coefficient indicates an increasing return to experience.

The *t* statistic for *crsgpa* is very large, over five using the usual standard error (which is the largest of the two). Using the robust standard error for *cumgpa*, its *t* statistic is about 2.61, which is also significant at the 5% level. The *t* statistic for *tothrs* is only about 1.17 using either standard error, so it is not significant at the 5% level.

(ii) This is easiest to see without other explanatory variables in the model. If *crsgpa* were the only explanatory variable, H_0: $\beta_{crsgpa} = 1$ means that, without any information about the student, the best predictor of term GPA is the average GPA in the students' courses; this holds essentially by definition. (The intercept would be zero in this case.) With additional explanatory variables it is not necessarily true that $\beta_{crsgpa} = 1$ because *crsgpa* could be correlated with characteristics of the student. (For example, perhaps the courses students take are influenced by ability – as measured by test scores – and past college performance.) But it is still interesting to test this hypothesis.

The *t* statistic using the usual standard error is $t = (.900 - 1)/.175 \approx -.57$; using the heteroskedasticity-robust standard error gives $t \approx -.60$. In either case we fail to reject H_0: $\beta_{crsgpa} = 1$ at any reasonable significance level, certainly including 5%.

(iii) The in-season effect is given by the coefficient on *season*, which implies that, other things equal, an athlete's GPA is about .16 points lower when his/her sport is competing. The *t* statistic using the usual standard error is about –1.60, while that using the robust standard error is about –1.96. Against a two-sided alternative, the *t* statistic using the robust standard error is just significant at the 5% level (the standard normal critical value is 1.96), while using the usual standard error, the *t* statistic is not quite significant at the 10% level ($cv \approx 1.65$). So the standard error used makes a difference in this case. This example is somewhat unusual, as the robust standard error is more often the larger of the two.

8.5 (i) No. For each coefficient, the usual standard errors and the heteroskedasticity-robust ones are practically very similar.

(ii) The effect is –.029(4) = –.116, so the probability of smoking falls by about .116.

(iii) As usual, we compute the turning point in the quadratic: $.020/[2(.00026)] \approx 38.46$, so about 38 and one-half years.

(iv) Holding other factors in the equation fixed, a person in a state with restaurant smoking restrictions has a .101 lower chance of smoking. This is similar to the effect of having four more years of education.

(v) We just plug the values of the independent variables into the OLS regression line:

$$\widehat{smokes} = .656 - .069 \cdot \log(67.44) + .012 \cdot \log(6,500) - .029(16) + .020(77) - .00026(77^2) \approx .0052.$$

Thus, the estimated probability of smoking for this person is close to zero. (In fact, this person is not a smoker, so the equation predicts well for this particular observation.)

8.6 (i) The proposed test is a hybrid of the BP and White tests. There are $k + 1$ regressors, each original explanatory variable and the squared fitted values. So, the number of restrictions tested is $k + 1$, and this is the numerator *df*. The denominator *df* is $n - (k + 2) = n - k - 2$.

(ii) For the BP test, this is easy: the hybrid test has an extra regressor, $\hat{y}^2$, and so the *R*-squared will be no less for the hybrid test than for the BP test. For the special case of the White test, the argument is a bit more subtle. In regression (8.20), the fitted values are a linear function of the regressors (where, of course, the coefficients in the linear function are the OLS estimates). So, we are putting a restriction on how the original explanatory variables appear in the regression. This means that the *R*-squared from (8.20) will be no greater than the *R*-squared from the hybrid regression.

(iii) No. The *F* statistic for joint significance of the regressors depends on $R^2_{\hat{u}^2}/(1 - R^2_{\hat{u}^2})$, and it is true that this ratio increases as $R^2_{\hat{u}^2}$ increases. But, the *F* statistic also depends on the *df*, and

the *df* are different among all three tests: the BP test, the special case of the White test, and the hybrid test. So we do not know which test will deliver the smallest *p*-value.

(iv) As discussed in part (ii), the OLS fitted values are a linear combination of the original regressors. Because those regressors appear in the hybrid test, adding the OLS fitted values is redundant; perfect collinearity would result.

8.7 (i) This follows from the simple fact that, for uncorrelated random variables, the variance of the sum is the sum of the variances: $\text{Var}(f_i + v_{i,e}) = \text{Var}(f_i) + \text{Var}(v_{i,e}) = \sigma_f^2 + \sigma_v^2$.

(ii) We compute the covariance between any two of the composite errors as

$$\begin{aligned}\text{Cov}(u_{i,e}, u_{i,g}) &= \text{Cov}(f_i + v_{i,e},\, f_i + v_{i,g}) = \text{Cov}(f_i,\, f_i) + \text{Cov}(f_i, v_{i,g}) + \text{Cov}(v_{i,e},\, f_i) + \text{Cov}(v_{i,e}, v_{i,g}) \\ &= \text{Var}(f_i) + 0 + 0 + 0 = \sigma_f^2,\end{aligned}$$

where we use the fact that the covariance of a random variable with itself is its variance and the assumptions that $f_i, v_{i,e},$ and $v_{i,g}$ are pairwise uncorrelated.

(iii) This is most easily solved by writing

$$m_i^{-1}\sum_{e=1}^{m_i} u_{i,e} = m_i^{-1}\sum_{e=1}^{m_i}(f_i + u_{i,e}) = f_i + m_i^{-1}\sum_{e=1}^{m_i} v_{i,e}.$$

Now, by assumption, f_i is uncorrelated with each term in the last sum; therefore, f_i is uncorrelated with $m_i^{-1}\sum_{e=1}^{m_i} v_{i,e}$. It follows that

$$\begin{aligned}\text{Var}\left(f_i + m_i^{-1}\sum_{e=1}^{m_i} v_{i,e}\right) &= \text{Var}(f_i) + \text{Var}\left(m_i^{-1}\sum_{e=1}^{m_i} v_{i,e}\right) \\ &= \sigma_f^2 + \sigma_v^2 / m_i,\end{aligned}$$

where we use the fact that the variance of an average of m_i uncorrelated random variables with common variance (σ_v^2 in this case) is simply the common variance divided by m_i – the usual formula for a sample average from a random sample.

(iv) The standard weighting ignores the variance of the firm effect, σ_f^2. Thus, the (incorrect) weight function used is $1/h_i = m_i$. A valid weighting function is obtained by writing the variance from (iii) as $\text{Var}(\bar{u}_i) = \sigma_f^2[1 + (\sigma_v^2/\sigma_f^2)/m_i] = \sigma_f^2 h_i$. But obtaining the proper weights requires us to know (or be able to estimate) the ratio σ_v^2/σ_f^2. Estimation is possible, but we do not discuss that here. In any event, the usual weight is incorrect. When the m_i are large or the ratio σ_v^2/σ_f^2 is small – so that the firm effect is more important than the individual-specific effect – the correct weights are close to being constant. Thus, attaching large weights to large firms may be quite inappropriate.

SOLUTIONS TO COMPUTER EXERCISES

C8.1 (i) Given the equation

$$sleep = \beta_0 + \beta_1 totwrk + \beta_2 educ + \beta_3 age + \beta_4 age^2 + \beta_5 yngkid + \beta_6 male + u,$$

the assumption that the variance of *u* given all explanatory variables depends only on gender is

$$Var(u \mid totwrk, educ, age, yngkid, male) = Var(u \mid male) = \delta_0 + \delta_1 male$$

Then the variance for women is simply δ_0 and that for men is $\delta_0 + \delta_1$; the difference in variances is δ_1.

(ii) After estimating the above equation by OLS, we regress $\hat{u}_i^2$ on $male_i$, $i = 1,2, \ldots,706$ (including, of course, an intercept). We can write the results as

$$\begin{aligned} \hat{u}^2 \;=\; & 189{,}359.2 \;-\; 28{,}849.6\, male \;+\; residual \\ & (20{,}546.4) \quad (27{,}296.5) \end{aligned}$$

$$n = 706, \quad R^2 = .0016.$$

Because the coefficient on *male* is negative, the estimated variance is higher for women.

(iii) No. The *t* statistic on *male* is only about –1.06, which is not significant at even the 20% level against a two-sided alternative.

C8.2 (i) The estimated equation with both sets of standard errors (heteroskedasticity-robust standard errors in parentheses) is

$$\begin{aligned} \widehat{price} = \; & -21.77 \;+\; .00207\, lotsize \;+\; .123\, sqrft \;+\; 13.85\, bdrms \\ & (29.48) \quad (.00064) \quad (.013) \quad (9.01) \\ & [36.28] \quad [.00122] \quad [.017] \quad [8.28] \end{aligned}$$

$$n = 88, \quad R^2 = .672.$$

The robust standard error on *lotsize* is almost twice as large as the usual standard error, making *lotsize* much less significant (the *t* statistic falls from about 3.23 to 1.70). The *t* statistic on *sqrft* also falls, but it is still very significant. The variable *bdrms* actually becomes somewhat more significant, but it is still barely significant. The most important change is in the significance of *lotsize*.

(ii) For the log-log model,

$$\widehat{\log(price)} = \underset{(0.65)\ [0.76]}{5.61} + \underset{(.038)\ [.041]}{.168\log(lotsize)} + \underset{(.093)\ [.101]}{.700\log(sqrft)} + \underset{(.028)\ [.030]}{.037\,bdrms}$$

$n = 88,\ R^2 = .643.$

Here, the heteroskedasticity-robust standard error is always slightly greater than the corresponding usual standard error, but the differences are relatively small. In particular, log(*lotsize*) and log(*sqrft*) still have very large *t* statistics, and the *t* statistic on *bdrms* is not significant at the 5% level against a one-sided alternative using either standard error.

(iii) As we discussed in Section 6.2, using the logarithmic transformation of the dependent variable often mitigates, if not entirely eliminates, heteroskedasticity. This is certainly the case here, as no important conclusions in the model for log(*price*) depend on the choice of standard error. (We have also transformed two of the independent variables to make the model of the constant elasticity variety in *lotsize* and *sqrft*.)

C8.3 After estimating equation (8.18), we obtain the squared OLS residuals $\hat{u}^2$. The full-blown White test is based on the *R*-squared from the auxiliary regression (with an intercept),

$$\hat{u}^2 \text{ on } llotsize,\ lsqrft,\ bdrms,\ llotsize^2,\ lsqrft^2,\ bdrms^2,$$
$$llotsize \cdot lsqrft,\ llotsize \cdot bdrms, \text{ and } lsqrft \cdot bdrms,$$

where "*l* " in front of *lotsize* and *sqrft* denotes the natural log. [See equation (8.19).] With 88 observations the *n*-*R*-squared version of the White statistic is $88(.109) \approx 9.59$, and this is the outcome of an (approximately) χ^2_9 random variable. The *p*-value is about .385, which provides little evidence against the homoskedasticity assumption.

C8.4 (i) The estimated equation is

$$\widehat{voteA} = \underset{(4.74)}{37.66} + \underset{(.071)}{.252\,prtystrA} + \underset{(1.407)}{3.793\,democA} + \underset{(0.392)}{5.779\log(expendA)} - \underset{(0.397)}{6.238\log(expendB)} + \hat{u}$$

$n = 173,\ R^2 = .801,\ \bar{R}^2 = .796.$

You can convince yourself that regressing the $\hat{u}_i$ on all of the explanatory variables yields an *R*-squared of zero, although it might not be exactly zero in your computer output due to rounding error. Remember, OLS works by choosing the estimates, $\hat{\beta}_j$, such that the residuals are uncorrelated in the sample with each independent variable (and the residuals have a zero sample average, too).

(ii) The B-P test entails regressing the $\hat{u}_i^2$ on the independent variables in part (i). The F statistic for joint significant (with 4 and 168 *df*) is about 2.33 with p-value $\approx .058$. Therefore, there is some evidence of heteroskedasticity, but not quite at the 5% level.

(iii) Now we regress $\hat{u}_i^2$ on $\widehat{voteA}_i$ and $(\widehat{voteA}_i)^2$, where the $\widehat{voteA}_i$ are the OLS fitted values from part (i). The F test, with 2 and 170 *df*, is about 2.79 with p-value $\approx .065$. This is slightly less evidence of heteroskedasticity than provided by the B-P test, but the conclusion is very similar.

C8.5 (i) By regressing *sprdcvr* on an intercept only we obtain $\hat{\mu} \approx .515$ se $\approx .021$). The asymptotic t statistic for H_0: $\mu = .5$ is $(.515 - .5)/.021 \approx .71$, which is not significant at the 10% level, or even the 20% level.

(ii) 35 games were played on a neutral court.

(iii) The estimated LPM is

$$\widehat{sprdcvr} = \underset{(.045)}{.490} + \underset{(.050)}{.035}\, favhome + \underset{(.095)}{.118}\, neutral - \underset{(.050)}{.023}\, fav25 + \underset{(.092)}{.018}\, und25$$

$n = 553$, $R^2 = .0034$.

The variable *neutral* has by far the largest effect – if the game is played on a neutral court, the probability that the spread is covered is estimated to be about .12 higher – and, except for the intercept, its t statistic is the only t statistic greater than one in absolute value (about 1.24).

(iv) Under H_0: $\beta_1 = \beta_2 = \beta_3 = \beta_4 = 0$, the response probability does not depend on any explanatory variables, which means neither the mean nor the variance depends on the explanatory variables. [See equation (8.38).]

(v) The F statistic for joint significance, with 4 and 548 *df*, is about .47 with p-value $\approx .76$. There is essentially no evidence against H_0.

(vi) Based on these variables, it is not possible to predict whether the spread will be covered. The explanatory power is very low, and the explanatory variables are jointly very insignificant. The coefficient on *neutral* may indicate something is going on with games played on a neutral court, but we would not want to bet money on it unless it could be confirmed with a separate, larger sample.

C8.6 (i) The estimates are given in equation (7.31). Rounded to four decimal places, the smallest fitted value is .0066 and the largest fitted value is .5577.

(ii) The estimated heteroskedasticity function for each observation i is $\hat{h}_i = \widehat{arr86}_i(1 - \widehat{arr86}_i)$, which is strictly between zero and one because $0 < \widehat{arr86}_i < 1$ for all i.

The weights for WLS are $1/\hat{h}_i$. To show the WLS estimate of each parameter, we report the WLS results using the same equation format as for OLS:

$$\widehat{arr86} = .448 - .168\,pcnv + .0054\,avgsen - .0018\,tottime - .025\,ptime86 - .045\,qemp86$$
$$(.018)\ (.019)\qquad (.0051)\qquad (.0033)\qquad (.003)\qquad (.005)$$

$n = 2{,}725,\ R^2 = .0744.$

The coefficients on the significant explanatory variables are very similar to the OLS estimates. The WLS standard errors on the slope coefficients are generally lower than the nonrobust OLS standard errors. A proper comparison would be with the robust OLS standard errors.

(iii) After WLS estimation, the *F* statistic for joint significance of *avgsen* and *tottime*, with 2 and 2,719 *df*, is about .88 with *p*-value $\approx$ .41. They are not close to being jointly significant at the 5% level. If your econometrics package has a command for WLS and a test command for joint hypotheses, the *F* statistic and *p*-value are easy to obtain. Alternatively, you can obtain the restricted *R*-squared using the same weights as in part (ii) and dropping *avgsen* and *tottime* from the WLS estimation. (The unrestricted *R*-squared is .0744.)

C8.7 (i) The heteroskedasticity-robust standard error for $\hat{\beta}_{white} \approx .129$ is about .026, which is notably higher than the nonrobust standard error (about .020). The heteroskedasticity-robust 95% confidence interval is about .078 to .179, while the nonrobust CI is, of course, narrower, about .090 to .168. The robust CI still excludes the value zero by some margin.

(ii) There are no fitted values less than zero, but there are 231 greater than one. Unless we do something to those fitted values, we cannot directly apply WLS, as $\hat{h}_i$ will be negative in 231 cases.

C8.8 (i) The equation estimated by OLS is

$$\widehat{colGPA} = 1.36 + .412\,hsGPA + .013\,ACT - .071\,skipped + .124\,PC$$
$$(.33)\quad (.092)\qquad (.010)\qquad (.026)\qquad (.057)$$

$n = 141,\ R^2 = .259,\ \bar{R}^2 = .238$

(ii) The *F* statistic obtained for the White test is about 3.58. With 2 and 138 *df*, this gives *p*-value $\approx$ .031. So, at the 5% level, we conclude there is evidence of heteroskedasticity in the errors of the *colGPA* equation. (As an aside, note that the *t* statistics for each of the terms is very small, and we could have simply dropped the quadratic term without losing anything of value.)

(iii) In fact, the smallest fitted value from the regression in part (ii) is about .027, while the largest is about .165. Using these fitted values as the $\hat{h}_i$ in a weighted least squares regression gives the following:

$$\widehat{colGPA} = \underset{(.30)}{1.40} + \underset{(.083)}{.402\, hsGPA} + \underset{(.010)}{.013\, ACT} - \underset{(.022)}{.076\, skipped} + \underset{(.056)}{.126\, PC}$$

$n = 141,\ R^2 = .306,\ \bar{R}^2 = .286$

There is very little difference in the estimated coefficient on *PC*, and the OLS *t* statistic and WLS *t* statistic are also very close. Note that we have used the usual OLS standard error, even though it would be more appropriate to use the heteroskedasticity-robust form (since we have evidence of heteroskedasticity). The *R*-squared in the weighted least squares estimation is larger than that from the OLS regression in part (i), but, remember, these are not comparable.

(iv) With robust standard errors – that is, with standard errors that are robust to misspecifying the function $h(\mathbf{x})$ – the equation is

$$\widehat{colGPA} = \underset{(.31)}{1.40} + \underset{(.086)}{.402\, hsGPA} + \underset{(.010)}{.013\, ACT} - \underset{(.021)}{.076\, skipped} + \underset{(.059)}{.126\, PC}$$

$n = 141,\ R^2 = .306,\ \bar{R}^2 = .286$

The robust standard errors do not differ by much from those in part (iii); in most cases, they are slightly higher, but all explanatory variables that were statistically significant before are still statistically significant. But the confidence interval for β_{PC} is a bit wider.

C8.9 (i) I now get $R^2 = .0527$, but the other estimates seem okay.

(ii) One way to ensure that the unweighted residuals are being provided is to compare them with the OLS residuals. They will not be the same, of course, but they should not be wildly different.

(iii) The *R*-squared from the regression $\breve{u}_i^2$ on $\breve{y}_i,\ \breve{y}_i^2,\ i = 1, ..., 807$ is about .027. We use this as $R^2_{\hat{u}^2}$ in equation (8.15) but with $k = 2$. This gives $F = 11.15$, and so the *p*-value is essentially zero.

(iv) The substantial heteroskedasticity found in part (iii) shows that the feasible GLS procedure described on page 279 does not, in fact, eliminate the heteroskedasticity. Therefore, the usual standard errors, *t* statistics, and *F* statistics reported with weighted least squares are not valid, even asymptotically.

(v) Weighted least squares estimation with robust standard errors gives

$$\widehat{cigs} = 5.64 + 1.30 \log(income) - 2.94 \log(cigpric) - .463\, educ$$

(37.31) (.54) (8.97) (.149)

+ .482 *age* – .0056 age^2 – 3.46 *restaurn*
(.115) (.0012) (.72)

$n = 807$, $R^2 = .1134$

The substantial differences in standard errors compared with equation (8.36) further indicate that our proposed correction for heteroskedasticity did not fully solve the heteroskedasticity problem. With the exception of *restaurn*, all standard errors got notably bigger; for example, the standard error for log(*cigpric*) doubled. All variables that were statistically significant with the nonrobust standard errors remain significant, but the confidence intervals are much wider in several cases.

[Instructor's Note: You can also do this exercise with regression (8.34) used in place of (8.32). This gives a somewhat larger estimated income effect.]

C8.10 (i) In the following equation, estimated by OLS, the usual standard errors are in (·) and the heteroskedasticity-robust standard errors are in [·]:

$$\widehat{e401k} = \underset{\substack{(.081)\\[.079]}}{-.506} + \underset{\substack{(.0006)\\[.0006]}}{.0124\ inc} - \underset{\substack{(.000005)\\[.000005]}}{.000062\ inc^2} + \underset{\substack{(.0039)\\[.0038]}}{.0265\ age} - \underset{\substack{(.00005)\\[.00004]}}{.00031\ age^2} - \underset{\substack{(.0121)\\[.0121]}}{.0035\ male}$$

$n = 9{,}275$, $R^2 = .094$.

There are no important differences; if anything, the robust standard errors are smaller.

(ii) This is a general claim. Since $\text{Var}(y|\mathbf{x}) = p(\mathbf{x})[1-p(\mathbf{x})]$, we can write $\text{E}(u^2 \mid \mathbf{x}) = p(\mathbf{x}) - [p(\mathbf{x})]^2$. Written in error form, $u^2 = p(\mathbf{x}) - [p(\mathbf{x})]^2 + v$. In other words, we can write this as a regression model $u^2 = \delta_0 + \delta_1 p(\mathbf{x}) + \delta_2 [p(\mathbf{x})]^2 + v$, with the restrictions $\delta_0 = 0$, $\delta_1 = 1$, and $\delta_2 = -1$. Remember that, for the LPM, the fitted values, $\hat{y}_i$, are estimates of $p(\mathbf{x}_i) = \beta_0 + \beta_1 x_{i1} + ... + \beta_k x_{ik}$. So, when we run the regression $\hat{u}_i^2$ on $\hat{y}_i, \hat{y}_i^2$ (including an intercept), the intercept estimates should be close to zero, the coefficient on $\hat{y}_i$ should be close to one, and the coefficient on $\hat{y}_i^2$ should be close to –1.

(iii) The White *F* statistic is about 310.32, which is very significant. The coefficient on $\widehat{e401k}$ is about 1.010, the coefficient on $(\widehat{e401k})^2$ is about –.970, and the intercept is about -.009. These estimates are quite close to what we expect to find from the theory in part (ii).

(iv) The smallest fitted value is about .030 and the largest is about .697. The WLS estimates of the LPM are

$$\widehat{e401k} = \underset{(.076)}{-.488} + \underset{(.0005)}{.0126\, inc} - \underset{(.000004)}{.000062\, inc^2} + \underset{(.0037)}{.0255\, age} - \underset{(.00004)}{.00030\, age^2} - \underset{(.0117)}{.0055\, male}$$

$n = 9{,}275, \quad R^2 = .108.$

There are no important differences with the OLS estimates. The largest relative change is in the coefficient on *male*, but this variable is very insignificant using either estimation method.

C8.11 (i) The usual OLS standard errors are in (·), the heteroskedasticity-robust standard errors are in [·]:

$$\widehat{nettfa} = \underset{\substack{(2.82)\\[3.23]}}{-17.20} + \underset{\substack{(.080)\\[.098]}}{.628\, inc} + \underset{\substack{(.0026)\\[.0044]}}{.0251\,(age-25)^2} + \underset{\substack{(2.04)\\[2.06]}}{2.54\, male}$$

$$- \underset{\substack{(4.40)\\[6.25]}}{3.83\, e401k} + \underset{\substack{(.124)\\[.220]}}{.343\, e401k\cdot inc}$$

$n = 2{,}017, R^2 = .131$

Although the usual OLS *t* statistic on the interaction term is about 2.8, the heteroskedasticity-robust *t* statistic is just under 1.6. Therefore, using OLS, we must conclude the interaction term is only marginally significant. But the coefficient is nontrivial: it implies a much more sensitive relationship between financial wealth and income for those eligible for a 401(k) plan.

(ii) The WLS estimates, with usual WLS standard errors in (·) and the robust ones in [·], are

$$\widehat{nettfa} = \underset{\substack{(2.27)\\[2.53]}}{-14.09} + \underset{\substack{(.084)\\[.091]}}{.619\, inc} + \underset{\substack{(.0019)\\[.0026]}}{.0175\,(age-25)^2} + \underset{\substack{(1.56)\\[1.31]}}{1.78\, male}$$

$$- \underset{\substack{(3.66)\\[3.51]}}{2.17\, e401k} + \underset{\substack{(.130)\\[.160]}}{.295\, e401k\cdot inc}$$

$n = 2{,}017, R^2 = .114$

The robust *t* statistic is about 1.84, and so the interaction term is marginally significant (two-sided *p*-value is about .066).

(iii) The coefficient on *e*401*k* literally gives the estimated difference in financial wealth at *inc* = 0, which obviously is not interesting. It is not surprising that it is not statistically different from zero; we obviously cannot hope to estimate the difference at *inc* = 0, nor do we care to.

(iv) When we replace *e*401*k*·*inc* with *e*401*k*·(*inc* – 30), the coefficient on *e*401*k* becomes 6.68 (robust *t* = 3.20). Now, this coefficient is the estimated difference in *nettfa* between those with and without 401(k) eligibility at roughly the average income, $30,000. Naturally, we can estimate this much more precisely, and its magnitude ($6,680) makes sense.

C8.12 (i) The estimated equation is

$$\widehat{math4} = \underset{(19.96)}{\underset{[23.09]}{91.93}} - \underset{(.015)}{\underset{[.017]}{.449\,lunch}} - \underset{(0.94)}{\underset{[1.13]}{5.40\,lenroll}} + \underset{(2.10)}{\underset{[2.35]}{3.52\,lexppp}}$$

$n = 1{,}692$, $R^2 = .373$

The heteroskedasticity-robust standard errors are somewhat larger, in all cases, than the usual OLS standard errors. The robust *t* statistic on *lexppp* is about 1.50, which raises further doubt about whether performance is linked to spending.

(ii) The value of the *F* statistic is 132.7, which gives a *p*-value of zero to at least four decimal places. Therefore, there is strong evidence of heteroskedasticity.

(iii) The equation estimated by WLS is

$$\widehat{math4} = \underset{(16.51)}{50.48} - \underset{(.015)}{.449\,lunch} - \underset{(0.84)}{2.65\,lenroll} + \underset{(1.69)}{6.47\,lexppp}$$

$n = 1{,}692$, $R^2 = .360$

where the usual WLS standard errors are in (·). The OLS and WLS coefficients on *lunch* are the same to three decimal places, but the other coefficients differ in practically important ways. The most important is that the WLS coefficient on *lexppp* is much larger than the OLS coefficient. Now, a 10 percent increase in spending (so *lexppp* increases by .1) is associated with roughly a .65 percentage point increase in the math pass rate. The WLS *t* statistic is much larger, too: $t \approx 3.83$.

(iv) Because our model of heteroskedasticity might be wrong, it is a good idea to compute the robust standard errors for WLS. On the key variable *lexppp*, the robust standard error is about 1.82, which is somewhat higher than the usual WLS standard error. The robust standard error on *lenroll* is also somewhat higher, 1.05. That on *lunch* is slightly lower: .014 to three decimal places. For *lexppp*, the robust *t* statistic is about 3.55, which is still very statistically significant.

(v) WLS is more precise: its robust standard error is 1.82, compare with the robust standard error for OLS of 2.35. Of course, the *t* for WLS is much larger partly because the coefficient estimate is much larger. The lower standard error has an effect, too.