Answers to Text Exercises

Chapter One: An Overview of Regression Analysis

1-3. (a) Positive, (b) negative, (c) positive, (d) negative, (e) ambiguous, (f) negative.

1-4. (a) Customers number 3, 4, and 20; no.
(b) Weight is determined by more than just height.
(c) People who decide to play the weight-guessing game may feel they have a weight that is hard to guess.

1-5. (a) The coefficients in the new equation are not the same as those estimated in the previous equation because the sample is different. When the sample changes, so too can the estimated coefficients. In particular, the constant term can change substantially between samples; in our research for this exercise, for instance, we found one sample that had a negative intercept (and a very steep slope).
(b) Equation 1.21 has the steeper slope (6.38 > 4.30) while Equation 1.24 has the greater intercept (125.1 > 103.4). They intersect at 9.23 inches above 5 feet (162.3 pounds).
(c) Equation 1.24 misguesses by more than 10 pounds on exactly the same three observations that Equation 1.21 does, but the sum of the squared residuals is greater for Equation 1.24 than for Equation 1.21. This is not a surprise, because the coefficients of Equation 1.21 were calculated using these data.
(d) If it were our last day on the job, we'd probably use an equation that we'd calculate from both equations by taking the mean, or by taking an average weighted by sample size, of the two.

1-6. (a) The coefficient of Li represents the change in the percentage chance of making a putt when the length of the putt increases by 1 foot. In this case, the percentage chance of making the putt decreases by 4.1 for each foot longer the putt is.
(b) The equations are identical. To convert one to the other, note that $\hat{P}_i = P_i - e_i$, which is true because $e_i = P_i - \hat{P}_i$ (or more generally, $e_i = Y_i - \hat{Y}_i$).
(c) 42.6 percent, yes; 79.5 percent, no (too low); −18.9 percent, no (negative!).
(d) One problem is that the theoretical relationship between the length of the putt and the percentage of putts made is almost surely nonlinear in the variables; we'll discuss models appropriate to this problem in Chapter 7. A second problem is that the actual dependent variable is limited by zero and one but the regression estimate is not; we'll discuss models appropriate to this problem in Chapter 13.

1-7. (a) The estimated slope coefficient of 3.62 represents the change in the size of a house (in square feet) given a one thousand dollar increase in the price of the house. The estimated intercept of −290 is the value of SIZE when PRICE equals zero. The estimated intercept is negative, but because the estimate includes the constant value of any omitted variables, any measurement errors, and/or an incorrect functional form, students should not attach any importance to the negative sign.

(b) No. All we have shown is that a statistical relationship exists between the price of a house and its size.

(c) The new slope coefficient would be 0.00362 (or 3.62/1000), but nothing else would change.

1-8. (a) β_Y is the change in the S caused by a one-unit increase in Y, holding G constant and β_G is the change in S caused by a one-unit increase in G, holding Y constant.

(b) +, −

(c) Yes. Richer states spend at least some of their extra money on education, but states with rapidly growing student populations find it difficult to increase spending at the same rate as the student population, causing spending per student to fall, especially if you hold the wealth of the state constant.

(d) $\hat{S}_i = -183 + 0.1422Y_i - 59.26G_i$. Note that $59.26 \times 10 = 5926 \times 0.10$, so nothing in the equation has changed except the scale of the coefficient of G.

1-9. (a) 2.29 is the estimated constant term, and it is an estimate of the gift when the alum has no income and no calls were made to that alum. 0.001 is an estimate of the slope coefficient of INCOME, and it tells us how much the gift would be likely to increase when the alum's income increases by a dollar, holding constant the number of calls to that alum. 4.62 is an estimate of the slope coefficient of CALLS, and it tells us how much the gift would be likely to increase if the college made one more call to the alum, holding constant the alum's income. The signs of the estimated slope coefficients are as expected, but we typically do not develop hypotheses involving constant terms.

(b) Once we estimate the equation, the left-hand variable now is the estimated value of the dependent variable because the right-hand side of the equation also consists of estimated coefficients (in all but one case multiplied by independent variables).

(c) An error term is unobservable and couldn't be included in an *estimated* equation from which we actually calculate a $\hat{Y}$. If a student rewords the question to ask why a *residual* isn't included, then most students should be able to figure out the answer if you remind them that $e = Y - \hat{Y}$.

(d) The right-hand side of the equation would become 2.29 + 1.0 INCOME + 4.62 CALLS. Nothing in the equation has changed except the scale of the coefficient of INCOME.

(e) Many good possibilities exist. However, students should be warned not to include a lagged dependent variable (as tempting as that may seem) until they've read Chapter 12 on time-series models.

1-10. (a) 17.08: A $1 billion increase in GDP will be associated with an increase of $17.08 in the average price of a new house. 12.928: Technically, the constant term equals the value of the dependent variable when all the independent variables equal zero, but in this case, such a definition has little economic meaning. As we'll learn in Chapters 4 and 7, estimates of the constant term should not be relied on for inference.

(b) It doesn't matter what letters we use as symbols for the dependent and independent variables.

(c) You could measure both P_t and Y_t in real terms by dividing each observation by the GDP deflator (or the CPI) for that year (and multiplying by 100).

(d) The price of houses is determined by the forces of supply and demand, and we won't discuss the estimation of simultaneous equations until Chapter 14. In a demand-oriented sense, GDP is probably measuring buying power, which is better represented by disposable income. In a supply-oriented sense, GDP might be standing for costs like wages and price of materials.

(e) No. In an annual time-series equation, the independent variables should be from different years, so GDP in year t makes sense. In a cross-sectional equation, the independent variables should represent different entities (in this case, different houses) in the same time period, so GDP would be identical for every observation. Instead, an independent variable should measure an attribute of the *i*th house.

1-11. (a) The error term is the theoretical, unobservable difference between the true (population) regression line and the observed point. The residual is the measured difference between the observed point and the estimated regression line.

(b)

Y_i	2	6	3	8	5	4
X_i	1	4	2	5	3	4
e_i	0.20	0.24	–0.12	0.92	0.56	–1.76
ε_i	0.50	0.00	0.00	0.50	0.50	–2.00

1-12. (a) β_2 represents the impact on the wage of the *i*th worker of a 1-year increase in the education of the *i*th worker, holding constant that worker's experience and gender.

(b) β_3 represents the impact on the wage of the *i*th worker of being male instead of female, holding constant that worker's experience and education.

(c) There are two ways of defining such a dummy variable. You could define $COLOR_i = 1$ if the *i*th worker is a person of color and 0 otherwise, or you could define $COLOR_1 = 1$ if the *i*th worker is not a person of color and 0 otherwise. (The actual name you use for the variable doesn't have to be "COLOR." You could choose any variable name as long as it didn't conflict with the other variable names in the equation.)

(d) We'd favor adding a measure of the quality of the worker to this equation, and answer iv, the number of employee of the month awards won, is the best measure of quality in this group. As tempting as it might be to add the average wage in the field, it would be the same for each employee in the sample and thus wouldn't provide any useful information.

1-13. (a) On one level, the answer is yes, because the coefficient of HOT is 59 times the size of the coefficient of EASE. However, there surely are some important variables that have been omitted from this equation, and it would be risky to draw conclusions when important variables have been left out. For example, if HOT teachers happen to be more effective communicators than EASY teachers, then the coefficient of HOT would pick up the impact of the omitted variable to the extent that the two variables were correlated. We'll address this topic (omitted variable bias) in more detail in Chapter 6.

(c) Yes. Besides the already-mentioned ability to communicate, other possible variables would include knowledge of the field, enthusiasm, organization, etc.

(d) Our guess is that the coefficient of HOT would decrease in size quite a bit. The coefficient of EASE already is extremely low, so it probably wouldn't change much.

Chapter Two: Ordinary Least Squares

2-3. (a) 71.
(b) 84.
(c) 213, yes.
(d) 155, yes

2-4. (a) The squares are "least" in the sense that they are being minimized.
(b) If $R^2 = 0$, then RSS = TSS, and ESS = 0. If R^2 is calculated as ESS/TSS, then it cannot be negative. If R^2 is calculated as 1 − RSS/TSS, however, then it can be negative if RSS > TSS, which can happen if $\hat{Y}$ is a *worse* predictor of Y than $\bar{Y}$ (possible only with a non-OLS estimator or if the constant term is omitted).
(c) Positive.
(d) We prefer Model T because it has estimated signs that meet expectations and also because it includes an important variable (assuming that interest rates are nominal) that Model A omits. A higher R^2 does not *automatically* mean that an equation is preferred.

2-5. (a) Yes. The new coefficient represents the impact of HEIGHT on WEIGHT, holding MAIL constant, while the original coefficient did not hold MAIL constant. We'd expect the estimated coefficient to change (even if only slightly) because of this new constraint.
(b) One weakness of R^2 is that adding a variable will usually decrease (and will never increase) the summed squared residuals no matter how nonsensical the variable is. As a result, adding a nonsensical variable will usually increase (and will never decrease) R^2.
(c) $\bar{R}^2$ is adjusted for degrees of freedom and R^2 isn't, so it's completely possible that the two measures could move in opposite directions when a variable is added to an equation.
(d) The coefficient is indeed equal to zero in theory, but in any given sample the observed values for MAIL may provide some minor explanatory power beyond that provided by HEIGHT. As a result, it's typical to get a nonzero estimated coefficient even for the most nonsensical of variables.

2-6. (a) Positive; both going to class and doing problem sets should improve a student's grade.
(b) Yes.
(c) $0.04 \times 1.74 > 0.02 \times 0.60$, so going to class pays off more.
(d) $0.02 \times 1.74 < 0.10 \times 0.60$, so doing problem sets pays off more. Since the units of variables can differ dramatically, coefficient size does not measure importance. (If all variables are measured identically in a properly specified equation, then the size of the coefficient is indeed one measure of importance.)
(e) An R^2 of 0.33 means that a third of the variation of student grades around their mean can be explained by attendance at lectures and the completion of problem sets. This might seem low to many beginning econometricians, but in fact it's either about right or perhaps even a bit higher than we might have expected.
(f) The most likely variable to add to this equation is the *i*th student's GPA or some other measure of student ability. We'd expect both R^2 and $\bar{R}^2$ to rise.

2-7. (a) Even though the fit in Equation A is better, most researchers would prefer Equation B because the signs of the estimated coefficients are as would be expected. In addition, X_4 is a theoretically sound variable for a campus track, while X_3 seems poorly specified because an especially hot *or* cold day would discourage fitness runners.

(b) The coefficient of an independent variable tells us the impact of a one-unit increase in that variable on the dependent variable holding constant the other explanatory variables in the equation. If we change the other variables in the equation, we're holding different variables constant, and so the $\hat{\beta}$ has a different meaning.

2-8. (a) Yes.

(b) At first glance, perhaps, but see below.

(c) Three dissertations, since $(978 \times 3) = \$2934 > (204 \times 2 + 36 \times 2) = \$480 > (\$460 \times 1) = \$460$

(d) The coefficient of D seems to be too high; perhaps it is absorbing the impact of an independent variable that has been omitted from the regression. For example, students may choose a dissertation adviser on the basis of reputation, a variable not in the equation.

2-9. As we'll learn in Chapters 6 and 7, there's a lot more to specifying an equation than maximizing $\bar{R}^2$.

2-10. (a) V_i: positive.

H_i: negative (although some would argue that in a world of perfect information, drivers would take fewer risks if they knew the state had few hospitals).

C_i: ambiguous because a high rate of driving citations could indicate risky driving (raising fatalities) *or* zealous police citation policies (reducing risky driving and therefore fatalities).

(b) No, because the coefficient differences are small and the data will differ from year to year. We'd be more concerned if the coefficients differed by orders of magnitude or changed sign.

(c) Since the equation for the second year has similar degrees of freedom and a much lower R^2, no calculation is needed to know that the equation for the first year has a higher R^2. Just to be sure, we calculated R^2 and obtained 0.652 for the first year and 0.565 for the second year.

2-11. (a) It might seem that the higher the percentage body fat, the higher the weight, holding constant height, but muscle weighs more than fat, so it's possible that a lean, highly muscled man could weigh more than a less well-conditioned man of the same height.

(b) We prefer Equation 1.24 because we don't think F belongs in the equation on theoretical grounds. The meaning of the coefficient of X changes in that F now is held constant.

(c) The fact that $\bar{R}^2$ drops when the percentage body fat is introduced to the equation strengthens our preference for Equation 1.24.

(d) This is subtle, but since 0.28 times 12.0 equals 3.36, we have reason to believe that the impact of bodyfat on weight (holding constant height) is very small indeed. That is, moving from average bodyfat to *no* bodyfat would lower your weight by only 3.36 pounds.

2-12. (a) $\partial\Sigma(e_i^2)/\partial\hat{\beta} = 2\Sigma(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i)(-1)$

$\partial\Sigma(e_i^2)/\partial\hat{\beta}_1 = 2\Sigma(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i)(\hat{\beta}_1 X_i)$

(b) $0 = -2\Sigma(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i)$

$0 = 2\hat{\beta}_1\Sigma(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i)(X_i)$ or, rearranging:

$\Sigma Y_{i=} n\hat{\beta}_0 + \hat{\beta}_1\Sigma X_i$

$\Sigma Y_i X_{i=} \hat{\beta}_0\Sigma X_i + \hat{\beta}_1 X_i^2$

These are the normal equations.

(c) To get $\hat{\beta}_1$, solve the first normal equation for $\hat{\beta}_0$, obtaining $\hat{\beta}_0 = (\Sigma Y_i - \beta_1\Sigma X_i)/N$ and substitute this value in for $\hat{\beta}_0$ where it appears in the second normal equation, obtaining $\Sigma Y_i X_i = (\Sigma Y_i - \hat{\beta}_1\Sigma X_i)(\Sigma X_i)/N + \hat{\beta}_1 X_i^2$, which becomes $\hat{\beta}_1 = (N\Sigma Y_i X_i - \Sigma Y_i X_i)/[N\Sigma X_i^2 - (\Sigma X_i)^2]$. With some algebraic manipulation (in part using the fact that $\Sigma X_i = N\bar{X}$), this simplifies to Equation 2.4.

(d) To get Equation 2.5, solve the first normal equation for $\hat{\beta}_0$, using $\bar{X} = \Sigma X_i/N$.

2-13. (a) Yes. We'd expect bigger colleges to get more applicants, and we'd expect colleges that used the common application to attract more applicants. It might seem at first that the rank of a college ought to have a positive coefficient, but the variable is defined as 1 = best, so we'd expect a negative coefficient for RANK.

(b) The meaning of the coefficient of SIZE is that for every increase of one in the size of the student body, we'd expect a college to generate 2.15 more applications, holding RANK and COMMONAP constant. The meaning of the coefficient of RANK is that every one-rank improvement in a college's *U.S. News* ranking should generate 32.1 more applications, holding SIZE and COMMONAP constant. These results do not allow us to conclude that a college's ranking is 15 times more important than the size of that college because the units of the variables SIZE and RANK are quite different in magnitude. On a more philosophical level, it's risky to draw any general conclusions at all from one regression estimated on a sample of 49 colleges.

(c) The meaning of the coefficient of COMMONAP is that a college that switches to using the common application can expect to generate 1222 more applications, holding constant RANK and SIZE. However, this result does not prove that a given college would increase applications by 1222 by switching to the common application. Why not? First, we don't trust this result because there may well be an omitted relevant variable (or two) and because all but three of the colleges in the sample use the common application. Second, in general, econometric results are evidence that can be used to support an argument, but in and of themselves they don't come close to "proving" anything.

(e) If you drop COMMONAP from the equation, $\bar{R}^2$ falls from 0.681. This is evidence (but not proof) that COMMONAP belongs in the equation.

Chapter Three: Learning to Use Regression Analysis

3-3. (a) A male professor in this sample earns $817 more than a female professor, holding constant the other independent variables in the equation.

(b) Most students will expect a negative coefficient, so they will call this an unexpected sign. Most professors and administrators will expect a positive sign because of the growing competition among colleges for African-American professors, so they will call this an expected sign. A key point here is not to change expectations based solely on this result.

(c) R is not a dummy variable because it takes on more than two values. For each additional year in rank, the *i*th professor's salary will rise by $406, holding constant the other independent variables in the equation.

(d) Yes. The coefficient is large and, as we'll learn in Chapter 5, statistically significantly greater than zero. (In addition, it's quite robust.)

(e) There's no measure of the quality of the professor in the equation as it stands, so good suggestions might be the number of articles published by the *i*th professor or the average teaching evaluation (on a standard scale) of the *i*th professor.

3-4. (a) There are many possible omitted explanatory variables; for example, the number of parking spaces near the restaurant.

(b) The sample could be larger, for one thing.

3-5. (a) Positive.

(b) Obviously, the best equation includes the actual traffic data (which, it turns out, are available). Since the traffic dummy variable is correlated with the actual traffic variable and since the new equation has expected signs and reasonable coefficients, it seems slightly better than Equation 3.5.

(c) No! The theoretical underpinnings of the model are much more important. Of course, the higher $\bar{R}^2$ is certainly a plus.

3-6. (a) New P = Old P/1000, so $\hat{\beta}_P$ goes from 0.3547 to 354.7 and all other coefficients remain unchanged.

(b) $\hat{\beta}_P = 320$ and all other coefficients remain unchanged.

(c) No.

3-8. (a) A male student's GRE subject score in Economics is likely to be 39.7 points higher than a female's, holding constant their GPA and SATs.

(b) This result is evidence of, but not proof of, bias. If we were sure that we had the best possible specification (the topic of Chapter 6) and if this result turned out to be statistically significant (the topic of Chapter 5), and if we were able to reproduce this result in other samples, we'd be much closer to a "proof." Even then, there still would be a possibility that some factor other than bias was the cause of these results.

(c) Possible variables include the number of upper division economics courses taken, the number of mathematics classes taken, and dummy variables measuring whether the student had taken econometrics or international economics (two fields frequently covered in the test). It's vital that any suggested variable be cross-sectional by student, however.

(d) The equation would become

$$\widehat{GRE}_i = 212.1 - 39.7G_i + 78.9GPA_i + 0.203SATM_i + 0.110SATV_i.$$

3-9. (a) Negative; positive; none.

(b) Holding all other included explanatory variables constant, a car with an automatic transmission gets 2.76 miles less per gallon than a model with a manual transmission, and a car with a diesel engine gets 3.28 miles more per gallon than one without a diesel engine.

(c) Lovell added the EPA variable because he wanted to test the accuracy of EPA estimates. If these estimates were perfectly accurate, then the EPA variable would explain all the variation in miles per gallon.

3-10. (a) All positive except for the coefficient of F_i, which in today's male-dominated movie industry probably has a negative expected sign. The sign of $\hat{\beta}_B$ certainly is unexpected.

(b) Fred, because \$500,000 < (\$4,000,000 – \$3,027,000).

(c) Yes, since $200 \times 15.4 = \$3,080,000 > \$1,200,000$.

(d) Yes, since \$1,770,000 > \$1,000,000.

(e) Yes, the unexpected sign of the coefficient of $\hat{\beta}_B$.

3-11. (a) The best way to handle three discrete conditions is to specify two dummy variables. For example, one dummy variable could = 1 if the iPod is new (and 0 otherwise) and the other dummy variable could = 1 if the iPod is used but unblemished (and 0 otherwise). The omitted condition, that the iPod is used and scratched, would be represented by both dummy variables equaling zero.

(b) Positive; negative; positive.

(c) In theory, the narrower the time spread of the observations, the better the sample, but 3 weeks probably is a short enough time period to ensure that the observations are from the same population. If the 3 weeks included a major shock to the iPod market, however, then the friend would be right, and the sample should be split into "before the shock" and "after the shock" subsamples.

(d) Yes, they match with the answer to part b.

(e) $\overline{R}^2$ is missing!

(f) $\overline{R}^2$ is 0.431.

Chapter Four: The Classical Model

4-2. (a) An additional pound of fertilizer per acre will cause corn yield (bushels/acre) to increase by 0.10 bushel/acre, holding rainfall constant. An additional inch of rain will increase corn yield (bushels/acre) by 5.33 bushels/acre holding fertilizer/acre constant.

(b) No. (This is a typical student mistake.) First, since it's hard to imagine *zero* inches of rain falling in an entire year, this intercept has no real-world meaning. In addition, recall that the OLS estimate of the intercept includes the nonzero mean of the error term in order to validate Classical Assumption II (as explained in the text), so even if rainfall were zero, it wouldn't make sense to attempt to draw inferences from the estimate of the β_0 term unless it was known that the mean of the (unobservable) error term was zero.

(c) No; this could be an unbiased estimate. 0.10 is the estimated coefficient for this sample, but the mean of the coefficients obtained for the population could still equal the true β_F.

(d) Not necessarily; 5.33 could still be close to or even equal to the true value. An estimated coefficient produced by an estimator that is not BLUE could still be accurate. If the estimator is biased, its bias could be small and its variance smaller still.

4-3. Pair "c" clearly violates Assumption VI, and pair "a" probably violates it for most samples.

4-4. (a) Most experienced econometricians would prefer an unbiased nonminimum variance estimate.

(b) Yes; an unbiased estimate with an extremely large variance has a high probability of being far from the true value. In such a case, a slightly biased estimate with a very small variance would be better.

(c) The most frequently used possibility is to minimize the mean square error (MSE), which is the sum of the expected variance plus the square of any expected bias.

4-5. (a) Classical Assumption II.

(b) Classical Assumption VI.

(c) R: A one-unit increase in yesterday's R will result in a 0.1% increase in today's Dow Jones average, holding constant the other independent variables in the equation.

M: The Dow Jones will fall by 0.017% on Mondays, holding constant the other independent variables in the equation.

(d) Technically, C is not a dummy variable because it can take on three different values. Saunders assumed (at least implicitly) that all levels of cloud cover between 0% and 20% have the same impact on the Dow and also that all levels of cloud cover between 21% and 99% have the same impact on the Dow. In addition, by using the same variable to represent both sunny and cloudy days, he constrained the coefficient of sun and cloud to be equal.

(e) In our opinion, this particular equation does little to support Saunders' conclusion. The poor fit and the constrained specification combine to outweigh the significant coefficients of R_{t-1} and M.

4-7. (a) The estimated coefficient of C shows that (for this sample) a one percent increase in the nonwhite labor force in the *i*th city adds 0.002 percentage points to the overall labor force participation rate in that city, holding constant all the other independent variables in the equation. The estimated coefficient of the dummy variable, D, shows that if a city is in the South, the labor force participation rate will be 0.80 percentage points lower than in other cities, holding constant the other explanatory variables in the equation.

(b) Perfect collinearity is virtually impossible in a cross-section like this one because no variable is a perfect linear function of another; some are closely related, but none is a perfect linear function.

(c) This does not imply that one of the estimates is biased. The estimates were taken from two different samples and are quite likely to differ. In addition, the true value may have changed between decades.

(d) Disagree. Beginners often confuse the constant term with the *mean* of the dependent variable. While the estimated constant term shows the value of the dependent variable in the unlikely case that all of the explanatory variables equal zero, it also includes the mean of the observations of the error term as mentioned in Question 4-2 (b).

4-8. We know that $\Sigma e_i^2 = \Sigma(Y_i - \hat{Y}_i)^2 = \Sigma(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i)^2$. To find the minimum, differentiate Σe_i^2 with respect to $\hat{\beta}_0$ and $\hat{\beta}_1$ and set each derivative equal to zero (these are the "normal equations"):

$$\delta(\Sigma e_i^2)/\delta\hat{\beta}_0 = 2[\Sigma(Y_i - \hat{\beta}_0\hat{\beta}_1 X_i)] = 0$$

or

$$\Sigma Y_{i=} N(\hat{\beta}_0) + \hat{\beta}_1(\Sigma X_i)$$

$$\delta(\Sigma e_i^2)/\delta\hat{\beta}_1 = 2[\Sigma(Y_i - \hat{\beta}_0\hat{\beta}_1 X_i)X_i] = 0$$

or

$$\Sigma Y_i X_i = \hat{\beta}_0(X_i) + \hat{\beta}_1(\Sigma X_i^2)$$

Solve the two equations simultaneously and rearrange:

$$\hat{\beta}_1 = [N(\Sigma Y_i X_i) - \Sigma Y_i X_i]/N(\Sigma X_i^2) - (\Sigma X_i)^2]$$

where $x_i = (X_i - \overline{X})$ and $y_i = (Y_i - \overline{Y})$.

$$\hat{\beta}_0 = [\Sigma X_i^2 \Sigma Y_i - \Sigma X_i \Sigma X_i Y_i]/[N(\Sigma X_i^2) - (\Sigma X_i)^2] = \overline{Y} - \hat{\beta}_1\overline{X}$$

To prove linearity:

$$\begin{aligned}
\hat{\beta}_1 &= \Sigma x_i y_i/\Sigma x_i^2 = \Sigma x_i(Y_i - \overline{Y})/\Sigma X_i^2 \\
&= \Sigma x_i Y_i/\Sigma x_i^2 - \Sigma x_i(\overline{Y})/\Sigma X_i^2 \\
&= \Sigma x_i(Y_i)/\Sigma x_i^2 - \overline{Y}\Sigma x_i/\Sigma X_i^2 \\
&= \Sigma x_i(Y_i)/\Sigma x_i^2 \text{ since } \Sigma x_i = 0 \\
&= \Sigma k_i Y_i \text{ where } k_i = x_i/\Sigma x_i^2
\end{aligned}$$

$\hat{\beta}_1$ is a linear function of Y, since this is how a linear function is defined. It is also a linear function of the βs and ϵ, which is the basic interpretation of linearity. $\hat{\beta}_1 = \hat{\beta}_0\Sigma k_1 + \beta_1 k_i x_i + \Sigma k_i \epsilon_1$. $\hat{\beta}_0 = \overline{Y} - \hat{\beta}_1(\overline{X})$ where $\overline{Y} = \hat{\beta}_0 + \hat{\beta}_1(\overline{X})$, which is also a linear equation.

To prove unbiasedness:

$$\begin{aligned}
\hat{\beta}_1 &= \Sigma k_i Y_i = \Sigma k_i(\beta_0 + \beta_1 X_1 + \epsilon_i) \\
&= \Sigma k_i \beta_0 + \Sigma k_i \beta_1 X_i + \Sigma k_i \epsilon_i
\end{aligned}$$

Since $k_i = x_i/\Sigma X_i^2 = (X_i - \overline{X})/\Sigma X_i - \overline{X})^2$, then $\Sigma k_1 = 0, \Sigma X_i^2 = 1/\Sigma X_i^2, \Sigma k_i x_i = \Sigma k_i X_i = 1$

So, $\hat{\beta}_i = \beta_1 + \Sigma k_i \epsilon_1$, and given the assumptions of ϵ, $E(\hat{\beta}_1) = \beta_1 + \Sigma k_1 E(\epsilon_1) = \beta_1$, proving $\hat{\beta}_1$ is unbiased.

To prove minimum variance (of all linear unbiased estimators): $\hat{\beta}_1 = \Sigma k_i Y_i$. Since $k_i = x_i/\Sigma x_i^2 = (X_i - \bar{X})/\Sigma(X_i - \bar{X})/\Sigma(X_i - \bar{X})^2$, $\hat{\beta}_1$ is a weighted average of the Ys, and the k_i are the weights. To write an expression for any linear estimator, substitute w_i for k_i, which are also weights but not necessarily equal to k_i:

$$\beta_1^* = \Sigma w_i Y_i, \text{ so } E(\beta_1^*) = \Sigma x_i E(Y_i) = \Sigma w_i(\beta_0 + \beta_1 X_1)$$
$$= \beta_0 \Sigma w_i + \beta_i \Sigma w_i X_i$$

In order for β_1^* to be unbiased, $\Sigma w_1 = 0$ and $\Sigma w_i X_i = 1$. The variance of β_1^*:

$$VAR(\beta_1^*) = VAR\ \Sigma w_i Y_i = \Sigma w_i VAR\ Y_i = \sigma^2 \Sigma w_i^2$$
$$[VAR(Y_i) = VAR(\epsilon_i) = \sigma^2]$$
$$= \sigma^2 \Sigma (w_i - x_i/\Sigma X_i^2 x_i/\Sigma X_i^2)^2$$
$$= \sigma^2 \Sigma (w_i - x_i/\Sigma X_i^2)^2 + \sigma^2 \Sigma x (\Sigma x_i^2)^2$$
$$+ 2\sigma^2 \Sigma (w_i - x_i/\Sigma X_i^2)(x_i/\Sigma X_i^2)$$
$$= \sigma^2 \Sigma (w_i - x_i/\Sigma X_i^2)^2 + \sigma^2/(\Sigma X_i^2)$$

The last term in this equation is a constant, so the variance of β_1^* can be minimized only by manipulating the first term. The first term is minimized only by letting $w_i = x_i/\Sigma X_i^2$, then:

$$VAR(\beta_1^*) = \sigma^2 / \Sigma X_i^2 = VAR(\beta_1^*)$$

When the least-squares weights, k_i, equal w_i, the variance of the linear estimator β_1 is equal to the variance of the least-squares estimator, $\hat{\beta}_1$. When they are not equal, $VAR(\hat{\beta}_1^*) > VAR(\hat{\beta}_1)$ Q.E.D.

4-9. (a) This possibly could violate Assumption III, but it's likely that the firm is so small that no simultaneity is involved. We'll cover simultaneous equations in Chapter 14.

(b) Holding constant the other independent variables, the store will sell 134.4 more frozen yogurts per fortnight if it places an ad. If we ignore long-run effects, this means that the owner should place the ad as long as the cost of the ad is less than the increase in profits brought about by selling 134.4 more frozen yogurts.

(c) The result doesn't disprove the owner's expectation. School is not in session during the prime yogurt-eating summer months, so the variable might be picking up the summer time increased demand for frozen yogurt from nonstudents.

(d) Answers will vary wildly, so perhaps it's best just to make sure that all suggested variables are time-series for 2-week periods. For students who have read Chapters 1–4 only, the best answer would be any variable that measures the existence of, prices of, or advertising of local competition. Students who have read Chapter 6 might reasonably be expected to try to find a variable whose expected omitted-variable bias on the coefficient of C is negative. Examples include the number of rainy days in the period or the number of college students returning home for vacation in the period.

4-10. (a) Yes; Yes. In particular, there's no measure of prices in the equation.
(b) Yes.
(c) Yes; very unlikely.
(d) No.
(e) No.
(f) No.
(g) The nightclub should hire a dancer, because the estimated coefficient is higher.

4-11. (a) The coefficient of DIVSEP implies that a divorced or separated individual will drink 2.85 more drinks than otherwise, holding constant the other independent variables in the equation. The coefficient of UNEMP implies that an unemployed individual will drink 14.20 more drinks than otherwise, holding constant the other independent variables in the equation. The signs of the estimated coefficients make sense, but we wouldn't have expected the coefficient of UNEMP to be five times the size of the coefficient of DIVSEP.
(b) The coefficient of ADVICE implies that an individual will drink 11.36 more drinks, holding constant the other independent variables in the equation, if a physician advises them to cut back on drinking alcohol. This coefficient certainly has an unexpected sign! Our guess is that DRINKS and ADVICE are simultaneously determined, since a physician is more likely to advise an individual to cut back on his or her drinking if that individual is drinking quite a bit. As a result, this equation almost surely violates Classical Assumption III.
(c) We'd expect each sample to produce different estimates of β_{ADVICE}. This entire group is called a sampling distribution of β-hats.
(d) The estimated β_{ADVICE} for this subsample is 8.62, which is a little lower than the coefficient for the entire sample. The other coefficients for this subsample differ even more from the coefficients for the entire sample, and the estimated coefficient of EDUC actually has an unexpected sign. These results are clear evidence of the advantages of large samples.

Chapter Five: Hypothesis Testing

5-3. (a) $H_0: \beta_1 \leq 0, H_A: \beta_1 > 0$
(b) $H_0: \beta_1 \geq 0, H_A: \beta_1 < 0; H_0: \beta_2 \leq 0, H_A: \beta_2 > 0;\ H_0: \beta_3 \leq 0, H_A: \beta_3 > 0$
(The hypothesis for β_3 assumes that it is never too hot to go jogging.)
(c) $H_0: \beta_1 \leq 0, H_A: \beta_1 > 0; H_0: \beta_2 \leq 0, H_A: \beta_2 > 0; H_0: \beta_3 \geq 0, H_A: \beta_3 < 0;$
(The hypothesis for β_3 assumes you're not breaking the speed limit.)
(d) $H_0: \beta_G = 0; H_A: \beta_G \neq 0$ (G for grunt.)

5-5. For β_N: Reject $H_0: \beta \geq 0$ if $|-4.42| > t_C$ and –4.42 is negative.
For β: Reject $H_0: \beta \leq 0$ if $|4.88| > t_C$ and 4.88 is positive.
For β_1: Reject $H_0: \beta \leq 0$ if $|2.37| > t_C$ and 2.37 is positive.
(a) $t_C = 1.943$; reject the null hypothesis for all three coefficients.
(b) $t_C = 1.311$; reject H_0 for all three coefficients.
(c) $t_C = 6.965$; cannot reject the null hypothesis for any of the three coefficients.

5-6. (a) $t_2 = (200 - 160)/25.0 = 1.6$; $t_C = 2.052$; therefore we cannot reject H_0. (Notice the violation of the principle that the null contains that which we do not expect.)

(b) $t_3 = 2.37$; $t_C = 2.756$; therefore we cannot reject the null hypothesis.

(c) $t_2 = 5.6$; $t_C = 2.447$; therefore we reject H_0 if it is formulated as in the exercise, but this poses a problem because the original hypothesized sign of the coefficient was negative. Thus the alternative hypothesis ought to have been stated: $H_A: \beta_2 < 0,$ and H_0 cannot be rejected.

5-7. This is a concern for part (a) but not for parts (b) and (c). In part (a), 160 probably is the coefficient we expect; after all, if our expectation was something else, why did we specify 160 in the null? In parts (b) and (c), however, it seems unlikely that we'd expect zero.

5-8. (a) For both $H_0: \beta \leq 0$ and $H_A: \beta > 0.$ For M, we can reject the null hypothesis because $t_M = 5.00$ and $|5.00| > 1.761,$ the 5% one-sided critical *t*-value for 14 degrees of freedom, and because 5.00 is positive. For Y, we cannot reject the null hypothesis because $t_Y = 1.25$ and $|1.25| < 1.761.$

(b) Here, $H_0: \beta_A = 0$ and $H_A: \beta_A \neq 0.$ We cannot reject the null hypothesis because $t_A = 0.80$ and $|0.80| < 2.861,$ the 1% two-sided critical *t*-value for 19 degrees of freedom.

(c) We think that B should have a negative effect on missed payments while C seems likely to have a positive effect (thought some students will argue that having more children indicates that the father likes children and will therefore miss fewer payments). Thus, for B, $H_0: \beta_B \geq 0$ and $H_A: \beta_B < 0.$ We cannot reject this null hypothesis because $t_B = 1.00$ and $|1.00| < 1.363,$ 1.363, the critical 10% one-sided *t*-value for 11 degrees of freedom. For C, $H_0: \beta_C \leq 0.$ Even though $t_C = -3.00$ and $|-3.00| > 1.363,$ we cannot reject the null hypothesis for C because t_C is negative, not positive.

5-9. (a) For all three, $H_0: \beta \leq 0,$ $H_A: \beta > 0,$ and the critical 5% one-sided *t*-value for 24 degrees of freedom is 1.711. For LOT, we can reject H_0 because $|+7.0| > 1.711$ and +7.0 is positive. For BED, we cannot reject H_0 because $|+1.0| < 1.711$ even though +1.0 is positive. For BEACH, we can reject H_0 because $|+10.0| > 1.711$ and +10.0 is positive.

(b) $H_0: \beta \geq 0,$ $H_A: \beta < 0,$ and the critical 10% one-sided *t*-value for 24 degrees of freedom is 1.318, so we reject H_0 because $|-2.0| > 1.318$ and −2.0 is negative.

(c) $H_0: \beta = 0,$ $H_A: \beta \neq 0,$ and the critical 5% two-sided *t*-value for 24 degrees of freedom is 2.064, so we cannot reject H_0 because $|-1.0| < 2.064.$ Note that we don't check the sign because the test is two-sided and both signs are in the alternative hypothesis.

(d) The main problems are that the coefficients of BED and FIRE are insignificantly different from zero.

(e) Given that we weren't sure what sign to expect for the coefficient of FIRE, the insignificant coefficient for BED is the most worrisome.

(f) Unless the students have read Chapter 6, this will be a difficult question for them to answer. It's possible that the dataset is unrepresentative, or that there's an omitted variable causing bias in the estimated coefficient of BED. Having said that, the most likely answer is that BED is an irrelevant variable if LOT also is in the equation. Beach houses on large lots tend to have more bedrooms than beach houses on small lots, so BED might be irrelevant if LOT is included.

5-10. (a) For both, $H_0: \beta \leq 0$ and $H_A: \beta > 0$. For WIN, we cannot reject H_0, even though the sign agrees with the sign implied by H_A, because $|+1.00| < 1.697$, the 5 percent one-sided critical *t*-value for 30 degrees of freedom. For FREE, we can reject H_0 at the 5 percent level of significance because $|2.00| > 1.697$ and because 2.00 has the sign implied by H_A.

(b) $H_0: \beta_{WEEK} \geq 0$ and $H_A: \beta_{WEEK} < 0$. We can reject H_0 at the 1 percent level because $|-4.00| > 2.457$, the 1 percent, one-sided critical *t*-value for 30 degrees of freedom and because –4.00 has the sign implied by H_A.

(c) $H_0: \beta_{DAY} = 0$ and $H_A: \beta_{DAY} \neq 0$. We cannot reject H_0 because $|-1.00| < 2.042$, the 5 percent two-sided critical *t*-value for 30 degrees of freedom.

(d) The coefficients of DAY and WIN are insignificantly different from zero. In addition, it's hard to rule out the possibility that a variable that belongs in the equation might have been omitted.

(e) A potential omitted variable is more worrisome than an insignificant coefficient.

(f) We'd suggest adding a variable that measures the weather (like inches of rainfall that day) to the equation. Even given San Diego's wonderful weather, there's a good chance that rainy or cold weather could cut down on attendance at an outdoor event.

5-11. (a) For the *t*-tests:

Coefficient:	β_P	β_M	β_S	β_T
Hypothesized sign:	+	+	–	–
t-value:	5.8	6.3	1.0	–3.3
$t_C = 1.671$ (5% one-sided with 60 d.f., as close to 73 as Table B–1 goes)	reject	reject	do not reject	reject

(b) No. We still agree with the authors' original expectations despite the contrary result.

(c) Keynes' point is well taken; empirical results will indeed allow an econometrician to discover a theoretical mistake now and then. Unfortunately, far too many beginning researchers use this loophole to change expectations to get "right" signs without enough thinking or analysis.

(d) Holding all other included explanatory variables constant, an increase in winning percentage of 150 points will increase revenues by $7,965,000 ($53.1 times 150 times 1000) and thus it would be profitable for this team to hire a $4,000,000 free agent who can raise its winning percentage to 500 from 350.

5-12. (a) NEW: $H_0: \beta \leq 0$, $H_A: \beta > 0$. Reject H_0 since $|-4.00| > 1.658$ and +5.34 has the sign of H_A.
SCRATCH: $H_0: \beta \geq 0$, $H_A: \beta < 0$. Reject H_0 since $|-4.00| > 1.658$ and –4.00 has the sign of H_A.

(b) BIDRS: $H_0: \beta \leq 0$, $H_A: \beta > 0$. Cannot reject H_0 since $|1.23| < 2.358$ even though +1.23 has the sign of H_A.

(c) Some experienced econometricians might drop BIDRS from the equation because of its low *t*-score, but we'd be inclined to keep the variable. The theory is strong, and the estimated coefficient is in the expected direction. As we'll see in Chapter 6, consistently dropping variables with low *t*-scores will result in coefficient bias.

(d) Most suggestions will be attributes of the iPod, but attributes of the auction of that iPod (like the length of time of the auction or whether there was a "buy it now" option available) also make sense.

5-13. (a) DIVSEP: $H_0: \beta \leq 0$, $H_A: \beta > 0$. . Cannot reject H_0 since $|1.11| < 1.658$ even though +1.11 has the sign of H_A.

UNEMP: $H_0: \beta \leq 0$, $H_A: \beta > 0$. Reject H_0 since $|2.75| > 1.658$ and +2.75 has the sign of H_A.

(b) EDUC: $H_0: \beta = 0$, $H_A: \beta \neq 0$. Cannot reject H_0 since $|-0.65| < 2.617$.

(c) ADVICE: $H_0: \beta \geq 0$, $H_A: \beta < 0$. Cannot reject H_0 since +5.37 doesn't have the sign of H_A, even though $|5.37| < 1.289$.

(d) No. We'd still expect ADVICE to have a negative impact on DRINKS in this structural equation. The problem is that the two variables almost surely are simultaneously determined, since a physician would be more likely to advise a patient to drink less if that patient was drinking quite a bit. This simultaneity violates Classical Assumption III. We'll learn how to estimate simultaneous equations in Chapter 14.

5-14. (a) All five tests are one-sided, so $t_C = 1.706$ throughout.

GDPN: $H_0: \beta \leq 0$, $H_A: \beta > 0$. Reject H_0 because $|+6.69| > 1.706$ and 6.69 is positive as in H_A.

CVN: $H_0: \beta \geq 0, H_A: \beta < 0$. Reject H_0 because $|-2.66| > 1.706$ and −2.66 is negative as in H_A.

PP: $H_0: \beta \leq 0, H_A: \beta > 0$. Do not reject H_0 because $|+1.19| < 1.706$.

DPC: $H_0: \beta \geq 0, H_A: \beta < 0$. Reject H_0 because $|-2.25| > 1.706$ and −2.25 is negative as in H_A.

IPC: $H_0: \beta \geq 0, H_A: \beta < 0$. Do not reject H_0 because $|-1.59| < 1.706$.

(b) Our confidence interval equation is $\hat{\beta} \pm t_C^* SE(\hat{\beta})$, and the 10% two-sided $t_C = 1.706$ (the same as a one-sided 5% t_C), so the confidence interval equals $\hat{\beta} \pm 1.706 \cdot SE(\hat{\beta})$, or:

GDPN: $1.07 < \hat{\beta} < 1.79$

CVN: $-0.98 < \hat{\beta} < -0.22$

PP: $-3.13 < \hat{\beta} < 17.75$

DPC: $-27.45 < \hat{\beta} < -3.81$

IPC: $-23.59 < \hat{\beta} < 0.83$

(c) Yes. The important signs were as expected and statistically significant, and the overall fit was good.

(d) The sizes of the coefficients would change, but not their signs or significance.

Chapter Six: Specification: Choosing the Independent Variables

6-3. (a)

Coefficient:	β_C	β_E	β_M
Hypothesized sign:	+	+	+
t-value:	4.0	4.0	–2.0
$t_C = 1.314$ (10% one-sided with 27 d.f.)	reject	reject	do not reject

The problem with the coefficient of M is that it is significant in the unexpected direction, one indicator of a possible omitted variable.

(b) The coefficient of M is unexpectedly negative, so we're looking for a variable the omission of which would cause negative bias in the estimate of β_M. We thus need a variable that is negatively correlated with meat consumption with a positive expected coefficient *or* a variable that is positively correlated with meat consumption with a negative expected coefficient. For the six variables listed, the expected bias is:

Possible Omitted Variable	Expected Sign of β	Correlation with M	Direction of Bias
B	+	+*	+
F	+	+	+
W	+*	+	+
R	–	–	+
H	–	+	–
O	–	–	+

*Indicates a weak expected sign or correlation.

(c) The best suggested variables are annual aggregate variables, the omission of which would cause negative bias. The expected bias equation is difficult to work with the first time around, so some students surely will suggest time-series variables the omission of which would cause positive expected bias. With luck, no students will suggest a disaggregate variable.

6-4. (a)

Coefficient:	β_E	β_I	β_T	β_V	β_R
Hypothesized sign:	–	+	–	–	–
Calculated *t*-score:	–3.0	1.0	–1.0	–3.0	3.0
$t_C = 1.682$, so:	sig.	insig.	insig. unexp.	sig.	sig. but sign

(b) Both income and tax rate are potential irrelevant variables not only because of the sizes of the *t*-scores but also because of theory. The significant unexpected sign for β_R is a clear indication that there is a potential omitted variable.

(c) It's prudent to attempt to solve an omitted variable problem before worrying about irrelevant variables because of the bias that omitted variables cause.

(d) The equation appears to show that television advertising is effective and radio advertising isn't, but you shouldn't jump to this conclusion. Improving the specification could change this result. In particular, although it's possible that radio advertising has little impact on smoking, it's very hard to believe that a radio antismoking campaign could cause a significant *increase* in cigarette consumption!

(e) *Theory*: Given the fairly price-inelastic demand for cigarettes, it's possible that T is irrelevant.
t-*score*: The estimated coefficient isn't significantly different from zero in the expected direction.
$\bar{R}^2$: $\bar{R}^2$ remains constant, which is exactly what will happen whenever a variable with a *t*-score with an absolute value of 1 is removed from (or added to) an equation.
Do other coefficients change?: None of the other estimated coefficients change significantly when T is dropped, indicating that dropping T caused no bias.
Conclusion: Based on these four criteria, it's reasonable to conclude that T is an irrelevant variable.

(f) You should not have been surprised. If a variable's coefficient has a *t*-score of exactly 1.00, then taking that variable out of an equation will not change $\bar{R}^2$.

6-5. (b) $\hat{Y} = 29.30 - 0.10\,PC + 0.036\,PB + 0.24\,YD - 0.027\,PRP$

	(0.03)	(0.018)	(0.025)	(0.036)
t =	– 2.95	1.98	9.78	–0.74

$N = 29 \qquad \bar{R}^2 = 0.9902$

Theory: Pork is one of many substitutes for chicken, so the theoretical case for PRP's inclusion in the equation is good but not overwhelming.
t-*score*: The estimated coefficient is insignificantly different from zero in the unexpected direction, providing evidence that PRP is irrelevant.
$\bar{R}^2$: $\bar{R}^2$ falls when PRP is added to the equation, providing evidence that it is irrelevant.
Do other coefficients change?: None of the other estimated coefficients change significantly when PRP is added, indicating that adding PRP has corrected no bias and providing evidence that it is irrelevant.
Conclusion: Based on the four criteria, it's reasonable to conclude that PRP is an irrelevant variable.

6-6. (a)

Coefficient	β_1	β_2	β_3	β_4
Hypothesized sign:	+	+	+	–
Calculated *t*-score:	5.0	1.0	10.0	3.0
$t_c = 2.485$ (1% level), so:	sig.	insig.	sig.	unexpected sign

(b) The significant unexpected sign of β_4 is evidence of a possible omitted variable that is exerting positive bias. The omitted variable must either be correlated positively with X_4 *and* have a positive expected coefficient or else be correlated negatively with X_4 *and* have a negative expected coefficient.

(c) A second run might add an independent variable that is theoretically sound and that could have caused positive bias in β_4. For example, the number of "attractions" like theaters or malls in the area would have a positive expected coefficient and be positively correlated with the number of nearby competing stores.

6-7. Some students will come to the conclusion that sequential specification searches are perfectly reasonable in business applications, and they need to be reminded that the regular use of such searches will produce consistently biased coefficient estimates.

6-8. Expected bias in $\hat{\beta} = (\beta_{\text{omitted}}) \cdot \text{f}(r_{\text{omitted, included}})$

(a) Expected bias = (−) · (+) = (−) = negative bias.

(b) (+) · (+) = (+) = positive bias; this bias will be potentially large since age and experience are highly correlated.

(c) (+) · (+) = (+) = positive bias.

(d) (−) · (0) = 0 = no bias; it may seem as though it rains more on the weekends, but there is no theoretical relationship between the two.

6-9. (a) In a supply equation, the coefficient of price will have a positive expected sign because the higher the price, holding all else constant, the more suppliers would be willing to produce.

(b) The price of inputs (such as labor, seeds, transporation, machinery, and fertilizer), the price of a "production substitute" (a crop or product that could be produced instead of the crop or product being modeled), and exogenous factors (like local growing conditions, local strikes that don't have an impact on the price, etc.) are just a few examples of important variables in a supply-side equation.

(c) Lag those independent variables that influence the production decision on a delayed basis. In particular, lag them by the length of time it takes for that particular event to have an impact on production. For example, if growers must make production decisions a year before the crop is harvested, then price should be lagged one year, etc. If a product can be stored at a fairly low cost, then such a lag might not be appropriate because producers could choose to wait until prices rose before going to market.

6-10. (a) Consumers and producers can react differently to changes in the same variable. A rise in price causes consumers to demand a lower quantity and producers to supply a greater quantity.

(b) Include variables affecting demand ("demand-side variables") only in demand equations and variables affecting supply ("supply-side variables") only in supply equations.

(c) Review the literature, decide whether the equation you wish to estimate is a supply or a demand equation, and when specifying the model, think carefully about whether an independent variable is appropriate for a demand or supply equation.

6-11. (a)

Coefficient	β_{PARENT}	β_{HSRANK}
Hypothesized sign:	−	+
Calculated *t*-score:	−11.26	4.22
$t_C = 1.679$ (5% level), so:	reject H_0	reject H_0

(b) There are no obvious signs of an omitted or irrelevant variable, but it seems probable that more than two variables determine financial aid grants in most colleges, so an omitted variable is very likely from a theoretical point of view.

(d) The estimated coefficient of MALE implies that a male financial aid applicant will receive $1570 less in grant aid than a female applicant, holding constant PARENT and HSRANK.

(e) *Theory*: When asked, most colleges will state that they award financial aid without regard to gender, but liberal arts colleges attract more females than males, so it's possible that a particular college might try to tilt its financial aid toward males. Given given this possibility and even given the charge of bias, however, the theory behind MALE is fairly weak.

t-*score*: The absolute value of the *t*-score is greater than the new critical *t*-value of 1.680, but the sign of the *t*-score is opposite that implied by H_A, so we cannot reject the null hypothesis.

$\bar{R}^2$: $\bar{R}^2$ increases when MALE is added, providing evidence that the variable belongs in the equation.

bias: Neither estimated slope coefficient changes by anything close to a standard error when MALE is added to the equation, providing evidence that omitting MALE from the equation does not cause any bias.

Three of the four specification criteria favor Equation 6.22, so we prefer Equation 6.22 to Equation 6.23. However, the significant unexpected sign in Equation 6.23 cannot be ignored. It indicates that there very likely is an omitted variable in Equation 6.22. Since we were concerned about the possibility of an omitted variable on theoretical grounds already, this empirical evidence is very convincing. In essence, *neither* equation is the best equation! Most beginning econometricians will not be very happy with this answer, but it's an important learning opportunity.

6-12. (a) No bias (+ · 0) unless weather patterns indicate a correlation between rainfall and temperature. If it tends to rain more when it's cold, then there would be a small negative bias (+ · −).

(b) Positive bias (+ · +).

(c) Positive bias (+ · +).

(d) Negative bias (+ · −) given a likely negative correlation between hours studied for the test and hours slept.

6-13. (a) X_1 = either dummy variable

X_2 = either dummy variable

X_3 = Parents' educational background

X_4 = Iowa Test score

(b) We have two variables for which we expect positive coefficients (Iowa score and Parents' education) and two positive estimated coefficients ($\hat{\beta}_3$ and $\hat{\beta}_4$), so we'd certainly expect X_3 and X_4 to be those two variables. In choosing between them, it's fair to expect a larger and more significant coefficient for Iowa than for Parents. Next, we have two variables for which we expect a zero coefficient (the dummies) and two estimated coefficients ($\hat{\beta}_1$ and $\hat{\beta}_2$) that are not significantly different from zero, so we'd certainly expect X_1 and X_2 to be the dummies. There is no evidence to allow us to distinguish which dummy is X_1 and which is X_2. (Students who justify this answer by expecting negative signs for coefficients of the two dummies are ignoring the presence of the Iowa test score variable in the equation that holds constant the test-taking skills of the student.)

(c)

Coefficient:	β_D	β_D	β_{PE}	β_{IT}
Hypothesized sign:	0	0	+	+
t-value:	−1.0	−0.25	+2.0	+12.0
$t_C = 2.093$ (5% two-sided with 19 d.f.)	do not reject	do not reject		
$t_C = 1.729$ (5% one-sided with 19 d.f.)			reject	reject

(d) As you can see, we used a one-sided test for those coefficients for which we had a specific prior expectation but a two-sided test around zero for those coefficients for which we did not.

6-14. (a) *Theory:* If PERCENT is the best proxy available for the quality and reliability of the seller, then it has a strong theoretical basis until a better variable can be found.

t-*score:* The coefficient is in the expected direction, but it's insignificant at the 5% level.

$\bar{R}^2$: $\bar{R}^2$ is not given, but it turns out that the addition of any variable with a *t*-score greater than one in absolute value will increase $\bar{R}^2$.

Bias: None of the coefficients change significantly.

Thus the four criteria are inconclusive. Because PERCENT appears to be the best available measure of seller quality, and because the sign of the coefficient is in the expected direction, we'd tend to keep PERCENT.

(b) In theory, PERCENT seems like the best we can do, but it might be an unreliable measure if there are very few transactions.

(c) When you drop PERCENT from the equation, $\bar{R}^2$ falls from 0.434 to 0.431.

6-15. (a) (i) The coefficient of CV is −0.19 with a SE $(\hat{\beta})$ of 0.23 and a *t*-score of −0.86. The $\bar{R}^2$ is 0.773, and the rest of the equation is extremely similar to Equation 5.14 except that the coefficient of CVN falls to −0.48 with a *t*-score of −1.86.

(ii) The coefficient of N is 0.00054 with a SE $(\hat{\beta})$ of 0.063 and a *t*-score of 0.0086. The $\bar{R}^2$ is 0.766, and the rest of the equation is identical (for all intents and purposes) to Equation 5.10.

(b) *Theory*: P is a price ratio, and while it's possible that a price ratio would be a function of the size of a market or a country, it's not at all obvious that either variable would add anything since CVN is already in the equation.

t-*score*: Both *t*-scores are insignificant.

$\bar{R}^2$: $\bar{R}^2$ falls when either variable is added.

Bias: None of the coefficients change at all when N is added, so it clearly is irrelevant. The omission of CV does change the coefficient of CVN somewhat, making it likely that CV is redundant since CVN is in the equation.

(c) Since CVN = f[CV/N], it would make little theoretical sense to include all three variables in an equation, even though technically you don't violate Classical Assumption VI by doing so.

(d) It's good econometric practice to report all estimated equations in a research report, especially those that were undertaken for specification choice or sensitivity analysis.

6-16. (a)

Coefficient:	β_{PR}	β_{PRCOMP}	β_{ADS}	β_{YD}
Hypothesized sign:	−	+	+	+
Calculated *t*-score:	1.0	3.0	3.0	3.0
$t_C = 1.711$, so:	insig./unexpected sign	sig.	sig.	sig.

(b) PR is hardly irrelevant, but there could be an omitted variable.

(c) The obvious addition is advertising for the competing instant oatmeal.

(d) Amish Oats competes with regular oatmeal, other cereals, and other breakfast foods, not just with the competing instant oatmeal.

Chapter Seven: Specification: Choosing a Functional Form

7-3. (a) Linear in the coefficients but not the variables.

(b) Linear in the coefficients but not the variables.

(c) Linear in the coefficients but not the variables.

(d) Nonlinear in both.

(e) Nonlinear in both.

7-4. (a)

Coefficient	β_1	β_2
Hypothesized sign:	+	+
Calculated *t*-score:	4.0	2.20
$t_C = 1.708$ at the 5% level, so:	sig.	sig.

(b) It is the sum of the constant effect of omitted independent variables and the nonzero mean of the sample error term observations; it does not mean that salaries (logged) could be negative.

(c) For this semilog function, the slopes are $\beta_1 SAL_i$ and $\beta_2 SAL_i$, which both increase as the Xs rise. This implies that a one-unit change in ED_i will cause a β_1 *percent* change in SAL_i, which makes sense for salaries.

(d) $\bar{R}^2$s cannot be compared because the dependent variables differ.

7-5. (a) To avoid confusion with β, let's use α_S as the coefficients.

Coefficient	α_{BETA}	α_{EARN}	α_{DIV}
Hypothesized sign:	−	+	+
Calculated *t*-score:	−1.99	1.45	3.33
$t_C = 1.671$ (5% level), so:	sig.	insig.	sig.

(b) It's unusual to have a lagged variable in a cross-sectional model, but in this equation all the variables are for 1996–2000 except for BETA, which is for 1958–1994 and therefore is indeed lagged. Fair assumed that the risk characteristics of companies don't change rapidly over time and stated that "five observations per company is not enough to get trustworthy estimates" (p. 17).

(c) We don't believe that any of Fair's variables are potentially irrelevant, because the theory behind each variable is exceptionally strong. Some students will think that EARN might be irrelevant because its coefficient has a low *t*-score, but we disagree with this concern because earnings growth is one of the most important determinants of stock prices. A student who drops EARN should conclude, based on the four specification criteria, that the variable belongs in the equation, because three of the four criteria support keeping EARN in the equation, and the *t*-score is close to being significant in the expected direction.

(d) The functional form is a semilog left, which is indeed appropriate both on a theoretical basis and also because two of the independent variables are expressed as percentages.

(e) This optional question is intentionally difficult. EARN and DIV both include negative values, so many students will give up. Since the negative values are extremely small, one possible way to estimate the equation is to set all the negative values equal to + 0.01, obtaining:

$$\begin{aligned} &\text{LNPE} = 3.23 - 0.18\ \text{LN BETA} + 0.071\ \text{LN EARN} + 0.098\ \text{LN DIV} \\ &\qquad\qquad (0.11) \qquad\quad (0.035) \qquad\quad (0.028) \\ &\text{t} = \qquad -1.69 \qquad\quad 2.02 \qquad\quad 3.49 \\ &\text{N} = 65 \qquad \bar{R}^2 = 0.23 \end{aligned}$$

However, these results, while completely reasonable, shed very little light on whether to use a double-log functional form, because we urge researchers to focus on theory, and not fit, to choose their functional forms. We think that Fair's choice of a semilog left is supported by the literature and by the fact that two of the independent variables are expressed in percentage growth terms. His optional question is intentionally difficult. EARN and DIV both include negative values, so many students will give up. Since the negative values are extremely small, it's reasonable to set all the negative values equal to + 0.01 (or to add to each variable the smallest amount necessary to make the most negative observation of that variable flip positive). However, even if you do this, the results shed very little light on whether to use a double-log functional form, because we urge researchers to focus on theory, and not fit, to choose their functional forms. We think that Fair's choice of a semilog left is supported by the literature and by the fact that two of the independent variables are expressed in percentage growth terms.

7-6. (a) The Midwest (the fourth region of the country).

(b) Including the omitted condition as a variable will cause the dummies to sum to a constant (1.0). This constant will be perfectly collinear with the constant term.

(c) Positive.

(d) Most correct = III; least correct = I.

(e) Any number of worker attributes make sense; for example the gender, age, or experience of the *i*th worker.

7-7. (a) Since the equations are double-log, the elasticities are the coefficients themselves:

Industry	**Labor**	**Capital**
Cotton	0.92	0.12
Sugar	0.59	0.33

(b) The sum indicates whether or not returns to scale are constant, increasing, or decreasing. In this example, Cotton is experiencing increasing returns to scale while Sugar is experiencing decreasing returns to scale.

(c) This question contains a hidden difficulty in that the sample size is purposely not given. "D" students will give up, while "C" students will use an infinite sample size. "B" students will state the lowest sample size at which each of the coefficients would be significantly different from zero (listed below), and "A" students will look up the article in *Econometrica* and discover that there were 125 cotton producers and 26 sugar producers, leading to the t_cs and hypothesis results listed below.

Coefficient:	β_{1C}	β_{2C}	β_{1S}	β_{2S}
Hypothesized sign:	+	+	+	+
t-value:	30.667	3.000	4.214	1.914
Lowest d.f. at which signif. (5%)	1	2	2	7
5% t_c given actual d.f.	1.645	1.645	1.714	1.714

(So all four coefficients are significantly different from zero in the expected direction.)

7-8. Let PCIi = per capita income in the *i*th period, GRi = rate of growth in the *i*th period, and ϵ_i = a classical error term.

(a) $GR_i = \propto_0 + \propto_1 PCI_i + \propto_2 PCI_i^2 + \epsilon_i$ where we'd expect $\propto_1 > 0$ and $\propto_2 < 0$.

(b) A semilog function alone cannot change from positive to negative slope, so it is not appropriate.

(c) $GR_i = \beta_0 + \beta_1 PCI_i + \beta_2 D_i + \beta_3 D_i PCI_i + \epsilon_i$, where $D_i = 0$ if $PCI_i \leq \$2{,}000$ and $D_i = 1$ if $PCI_i > \$2{,}000$. ($2,000 is an estimate of the turning point.).

7-9. (a) $H_0: \beta \leq 0; H_A: \beta > 0;$ for both.

(b) L: t = 2.02; K: t = 5.86, since $t_c = 1.717$, we can reject H_0 for both.

(c) The relative prices of the two inputs need to be known before this question can be answered.

7-10. (a) The expected signs are β_1, + or ?; β_2, + β_3, + ; β_4, +.

(b) AD_i/SA_i: The inverse form implies that the larger sales are, the smaller will be the impact of advertising on profits. CAP_i, ES_i, DC_i: The semilog right functional form implies that as each of these variables increases (holding all others in the equation constant), PR increases at a decreasing rate.

(c) β_2, β_3, and β_4, all have positive expected signs, so (+) · (+) = (+) = positive expected bias on β_1 if one of the other Xs were omitted.

7-11. (a) Polynomial (second-degree, with a negative expected coefficient for age and a positive expected coefficient for age squared).

(b) Double-log. (We would not quibble with those who chose a linear form to avoid the constant elasticity properties of a double-log.)

(c) Semilog (lnX).

(d) Linear. (All intercept dummies have a linear functional relationship with the dependent variable by definition.)

(e) Inverse. (Most students will remember from the text that a U-shaped polynomial typically is used to model a cost curve and will want to apply it here. The problem is that the telephone industry appears to be an industry in which costs continually decrease as size increases, making an inverse our choice.)

7-12. (a) The estimated coefficients all are in the expected direction, and those for A and S are significant. $\bar{R}^2$ seems fairly low, even for a cross-sectional data set of this nature.

(b) It implies that wages rise and then fall with respect to age but does not imply perfect collinearity.

(c) With a semilog left functional form (lnY), a slope coefficient represents the percentage change in the dependent variable caused by a one-unit increase in the independent variable (holding constant all the other independent variables). Since pay raises are often discussed in percentage terms, such a functional form frequently is used to model wage rates and salaries.

(d) It's a good habit to ignore $\hat{\beta}_0$ (except to make sure that one exists) even if it looks too large or too small.

(e) The poor fit and the insignificant estimated coefficient of union membership are all reasons for being extremely cautious about using this regression to draw any conclusions about union membership.

7-13. (a)

Coefficient:	β_{LQ}	β_A	β_V
Hypothesized sign:	+	−	−
t-value:	4.0	−2.0	0.13
$t_c = 1.725$ (5% one-sided with 20 d.f.)	reject	reject	do not reject

(b) Q constant; A and V non-constant.

(c) No. The coefficient of V is quite insignificant, and the equation (simplified from an unpublished article) is flawed to boot. Note, however, that the violence may be causing the absentee rate to rise, so that the significant coefficient for A does indicate some support for the charge.

(d) In our opinion, this is a classic case of "spurious correlation" because actual total output appears on both sides of the equation, causing almost all of the fit by definition. If we could make one change, we'd drop LQ from the equation, but we worry that little will be left when we do.

7-14. (a)

Coefficient	β_B	β_S	β_D
Hypothesized sign:	+	+	−
Calculated *t*-score:	−0.08	1.85	−1.29
$t_c = 1.682$, so:	insig.	sig.	insig.

The insignificance of $\hat{\beta}_B$ could be caused by an omitted variable, but it's likely that the interaction variable has "soaked up" the entire effect of beer consumption. Although we cannot reject the null hypothesis for $\hat{\beta}_D$, we see no reason to consider D to be an irrelevant variable because of its sound theory and reasonable statistics.

(b) The interaction variable is a measure of whether the impact of beer drinking on traffic fatalities rises as the altitude of the city rises. For each unit increase in the multiple of B and A, F rises by 0.011, holding constant all the other independent variables in the equation. Thus the size of the coefficient has no real intuitive meaning in and of itself.

(c) $H_0: \beta_{BA} \leq 0$

$H_A: \beta_{BA} > 0$

Reject H_0 because $|+4.05| > t_C = 1.682$ and 4.05 is positive and thus matches the sign implied by H_A.

(d) Although there is no ironclad rule (as there is with slope dummies) most econometricians include both interaction-term components as independent variables. The major reason for this practice is to avoid the possibility that an interaction term's coefficient might be significant only because it is picking up the effect of the omitted interaction-term component (bias).

(e) The exception to this general practice occurs when there is no reason to expect the interaction-term component to have any theoretical validity on its own. We prefer Equation 7.25 to 7.26 because we don't believe that altitude typically would be included as an independent variable in a highway fatality equation. Of our other three specification criteria, only the increase in $\bar{R}^2$ supports considering A to be a relevant variable. However, even moderate theoretical support for the inclusion of A on its own would result in our preferring Equation 7.26.

7-15. (a) Let's look at β_D. The easy answer is that an increase of 100,000 residential customers will cause an increase of \$5.00 in advertising and promotional expense per 1000 residential kilowatt hours, holding constant G, D, G*D, and S*D. However, this technically correct answer ignores the existence of the interaction variables. We'd rather say that for duopolies, an increase of 100,000 residential customers will cause an increase of \$5.00 in advertising and promotional expense per 1000 residential kilowatt hours, holding constant G (because the other terms fall out of the equation), and for monopolies an increase of 100,000 residential customers will cause a decrease of \$15.00 in advertising and promotional expense per 1000 residential kilowatt hours, holding constant G. The answers for the other coefficients are similarly annoying.

(b)

Coefficient:	β_S	β_G	β_D	β_{SD}	β_{GD}
Hypothesized sign:	+	+	+	+?	+
t-value:	4.5	0.4	2.9	–5.0	2.3
$t_C = 1.645$	reject	do not	reject	do not	reject
(5% one-sided	reject			reject	
with infinite d.f.)					

(c) As Primeaux puts it (on page 622 of his article), "A duopoly firm of small size spends more than a monopoly firm of the same size. However, as scale increases, eventually, the duopoly firm spends less."

(d) Again, from page 622, "There is no difference between monopoly and duopoly firms at zero rates of growth in sales. However, as growth takes place, the duopoly firms engage in more sales promotion activity."

7-16. (a)

Coefficient	β_{TOP}	β_{WEIGHT}	β_{HP}
Hypothesized sign:	–	+	–
Calculated *t*-score:	–6.49	2.23	–7.74
$t_C = 2.479$, so:	sig.	insig.	sig.

(b) At first glance, all three problems seem possible.

(c) Since TIME and HP are inversely related by theory, an inverse functional form should be used.

(d) Positive, because as HP gets bigger, 1/HP gets smaller and TIME gets smaller. NOTE: In the first printing of the text, the sign in front of 1/HP is incorrectly stated as negative. The *t*-score below it, which is positive, is correct.

(e) All of our specification criteria except $\bar{R}^2$ favor Equation 7.28, but the theory behind the inverse functional form is so clear-cut that we would stick with the inverse even if the other criteria favored Equation 7.27.

(f) $\bar{R}^2$ can indeed be used to compare the fits of the equations because the dependent variable has not been transformed.

7-16. "Part g version"

$$SPEED_i = 137.7 + 0.49\ TOP_i - 0.018\ WEIGHT_i + 0.93\ HP_i$$
$$\qquad\qquad (5.90) \qquad (0.006) \qquad\quad (0.12)$$
$$\bar{R}^2 = .889 \quad N = 30$$

(a)

Coefficient	β_{TOP}	β_{WEIGHT}	β_{HP}
Hypothesized sign:	+	−	+
Calculated *t*-score:	−0.18	−2.81	13.19
$t_C = 2.479$, so:	insig.	sig.	sig.

(b) At first glance, all three problems seem possible.

(c) An increase in HP would indeed increase SPEED at a diminishing rate, holding WEIGHT constant, so an inverse functional form, while not explicitly called for by the underlying physics, is plausible.

(d) Negative, because as HP gets bigger, 1/HP gets smaller and SPEED gets bigger.

(e) For SPEED, switching to 1/HP produces:

$$SPEED_i = 298.5 - 13.25\ TOP_i - 0.025 WEIGHT_i - 23770(1/HP_i)$$
$$\qquad\qquad (10.58) \qquad (0.010) \qquad (3118)$$
$$t = \qquad -1.25 \qquad -2.46 \qquad -7.62$$
$$\bar{R}^2 = 0.738 \quad N = 30$$

While the statistical results support the linear functional form, the choice should be made on the basis of theory. As mentioned above, 1/HP is supported by the underlying physics in a TIME equation but not in a SPEED equation. As a result, the choice isn't obvious. We tend to prefer the linear functional form in this case, but other researchers might disagree.

(f) $\bar{R}^2$ can indeed be used to compare the fits of the equations because the dependent variable has not been transformed.

Chapter Eight: Multicollinearity

8-3. Perfect multicollinearity; each can be stated as an exact function of the other two. To solve the perfect multicollinearity problem, one of the three explanatory variables must be dropped.

8-4. Dominant variables are likely in a and d. In a., the number of games won should equal the number of games played (which is a constant) minus the number of games lost. In d., the number of autos produced should equal four times the number of tires bought (if no spare is sold with the cars or five if a spare is included).

8-5. a, c

8-6. (a)

Coefficient	β_F	β_S	β_A
Hypothesized sign:	+	+	+
Calculated *t*-score:	2.90	−1.07	5.07
$t_C = 2.447$, so:	sig.	insig.	sig
		unexpected sign	

(b) All three are possibilities.

(c) Multicollinearity is a stronger possibility.

(d) Yes; the distribution of the β_S is wider with multicollinearity.

8-7. (a) No; no explanatory variable is a perfect function of another.

(b) Yes; income in any quarter will be strongly correlated with income in previous quarters.

(c) If all the variables were specified in terms of first differences, it's likely that much of the multicollinearity would be avoided.

8-8. (a) Don't change your regression just because a fellow student says you are going to have a problem; in particular, even if you do have multicollinearity, you may well end up doing nothing about it.

(b) There is a reasonable $\bar{R}^2$ (0.36) with all low *t*-scores (the highest is 0.84). Furthermore, the simple correlation coefficient between HR and RBI is 0.93. Also, the VIFs for HR and RBI are >5.

(c) Since a sample of eight is extremely small, the first solution to try is to increase the sample size. In this particular case, a larger sample doesn't rid the equation of damaging multicollinearity, so we'd favor dropping one of the redundant variables. There also are a number of potential omitted variables.

8-9. (a)

Coefficient:	β_{PC}	β_{PQ}	β_Y	β_C	β_N	
Hypothesized sign:		+	−	+	+	+
t-value:	0.801	−1.199	0.514	−1.491	1.937	

$t_C = 1.725$ at the 5% level, so only $\hat{\beta}_N$ is significantly different from zero in the expected direction.

(b) The obviously low *t*-scores could be caused by irrelevant variables, by omitted variables biasing the estimated coefficients toward zero, or by severe imperfect multicollinearity.

(c) The high simple correlation coefficient between Y and C indicates that the two are virtually identical (redundant), which makes sense theoretically. The simple correlation coefficient between the two price variables is not as high, but mild multicollinearity exists nonetheless.

(d) Y_t and C_t both serve as measures of the aggregate buying power of the economy, so they are redundant, and one should be dropped. It doesn't matter statistically which one is dropped, but Y_t seems analytically more valid than C_t, so we'd drop C. Dropping one of the price variables would be a mistake, since they have opposite expected signs. While forming a relative price variable is an option, the low level of multicollinearity, the reasonable coefficients, and the possibility that C_t is also multicollinear with prices (so dropping it will improve things) all argue for making just one change.

8-10. (a)

Coefficient	β_C	β_P	β_E
Hypothesized sign:	+	+	+
Calculated *t*-value:	31.15	–0.07	–0.85
t_C = 1.684 at the 5% level, so:	sig.	insig.	sig
		unexpected sign	

(b) There is definite multicollinearity in the equation.

(c) The payroll for defense workers and the number of civilians employed in defense industries are redundant; they measure the same thing. As a result, one or the other should be dropped.

8-11. (a)

Coefficient:	β_M	β_B	β_A	β_S
Hypothesized sign:	+	+	+	+
t-value:	5.0	1.0	–1.0	2.5
t_C = 1.645 (5% one-sided with infinite d.f.)	reject	do not reject	do not reject	reject

(b) The insignificant *t*-scores of the coefficients of A and B could have been caused by omitted variables, irrelevance, or multicollinearity (a good choice, since that's the topic of this chapter). In particular, since most students graduate at about the same age, the collinearity between A and B must be fairly spectacular (Stanford gave us no clues).

(c) It's probably a good idea, since the improvement in GPA caused by extra maturity may eventually be offset by a worsening in GPA due to separation from an academic environment.

(d) We believe in making just one change at a time to best be able to analyze the impact of the change on the estimated regression. Thus our first choice would be to drop either A or B (we'd prefer to drop A, but on theoretical grounds, not as a result of the unexpected sign). Switching to a polynomial *before* dropping one of the redundant variables will only make things worse, in our opinion.

8-12. (a) 2.35, 2.50, 1.18.

(b) 8.09, 1.29, 9.07

(c) Since X_1 and X_2 are the only independent variables in the equation, VIF(X_1) must equal VIF(X_2) and hence VIF(X_1) = 3.8.

(d) In a two-variable equation, $r^2 = R^2$. Thus $R^2 = (0.80)^2$, and VIF(X_1) = VIF(X_2) = $1/(1 - 0.64) = 2.78$.

8-13. (a)

Coefficient:	β_{PF}	β_{PB}	β_{lnYD}	β_P	
Hypothesized sign:		–	+	+	–?
t-value:	–0.98	0.24	0.31	–0.48	
t_C = 1.725 % one-sided with 20 d.f.)	We cannot reject any of the null hypotheses!				

(b) Omitted variables, irrelevant variables, and multicollinearity all are possibilities

(c) No. The dependent variable is the average amount of fish consumed per capita by everyone in the United States so the number of Catholics has no theoretical relevance. To make things worse, it's extremely likely to be redundant with the log of disposable income.

(d) P may well be an irrelevant variable, but dropping it from the equation is a bad suggestion, since the purpose of the research was to determine whether the Pope's decision had an impact on fish consumption. If you drop P, you won't be able to answer your research question.

(e) There's no doubt that the two price variables are highly correlated, but they're far from redundant, because they measure quite different things.

(f) PF: 36.44; PB: 17.27; LNYD: 8.98; P: 2.35.

(g) Positive.

(h) We prefer Equation 8.24. The four specification criteria aren't necessarily applicable when you replace two variables with a third, but in this case the criteria shed at least some light on the specification choice:

Theory: The theory behind the two models is similar. Many economists would prefer a relative price ratio (as compared to two individual price variables) when consumers are choosing between the two products, as is the case in Equation 8.24.

t-test: The coefficient of RP is significantly different from zero in the expected direction while the coefficients for PB and PF are not, supporting Equation 8.24.

$\bar{R}^2$: $\bar{R}^2$ falls when PB and PF are replaced by RP. This result supports Equation 8.23.

Bias: The coefficient of lnYD changes by far more than a standard error, probably because PF was acting as a proxy for disposable income in Equation 8.23.

(i) The coefficient of P is quite insignificant in all specifications, providing robust support for the conclusion that the Pope's decision did not cut down on the consumption of fish. This is in contrast with the findings of the original empirical work on the issue, Frederick Bell's "The Pope and the Price of Fish," *American Economic Review*, Dec. 1968, pp. 1346–1350. Bell studied a different sample (the demand for fish in New England in the first 9 months after the Pope's decision), so it's possible that both results are valid.

8-14. (a)

Coefficient	β_Y	β_Y^2	β_H	β_A
Hypothesized sign:	+	−	+	+
Calculated *t*-score:	3.00	−0.80	6.50	−1.00
$t_C = 1.282$, so:	sig.	insig.	sig.	insig. unexpected sign

(b) The functional form appears reasonable. The coefficient of Y can be greater than 1.0 since Y^2 is in the equation with a negative coefficient.

(c) A and H seem potentially redundant.

(d) The high VIFs strengthen the answer.

(e) Either drop A or, if the purpose behind A was to measure the differential eating habits of children, change the two variables to A and (H – A).

8-15. The ever-innovative Rao and Miller used this example (developed by Professor Maurice G. Kendall) to show that the inspection of simple correlation coefficients is not an adequate test for multicollinearity.

(a) Since R and L are obviously correlated but R and (L – R) are not, many beginning students will want to drop either R or L from Model A, but this would leave out the difference between leg lengths that is the inherent causal variable.

(b) As Rao and Miller point out (on page 48 of their text), the implicit estimates of the coefficients are identical "because the conditions imposed on the residuals for estimation in either case are implicitly the same." To calculate the coefficients of one model from the other, multiply out the β_2 term of Model B, reconfigure to correspond to Model A, and solve for the coefficients of Model B: $\beta_0 = \alpha_0$, $\beta_1 = (\alpha_1 - \alpha_2)$, and $\beta_2 = \alpha_2$.

(c) Since the coefficient estimates are *identical* for every sample, their distributions must also be identical, meaning that the two models are identically vulnerable to multicollearity.

(d) If you drop L from Model A, then the linkage between the Models cited in the answers above is lost.

Hints for Section 8.7.2: The SAT Interactive Regression Learning Exercise:

1. Severe multicollinearity between APMATH and APENG is the only possible problem in this regression. You should switch to the AP linear combination immediately.

2. An omitted variable is a distinct possibility, but be sure to choose the one to add on the basis of theory.

3. Either an omitted or irrelevant variable is a possibility. In this case, theory seems more important than any mild statistical insignificance.

4. On balance, this is a reasonable regression. We see no reason to worry about theoretically sound variables that have slightly insignificant coefficients with expected signs. We're concerned that the coefficient of GEND seems larger in absolute size than those reported in the literature, but none of the specification alternatives seems remotely likely to remedy this problem.

5. An omitted variable is a possibility, but there are no signs of bias and this is a fairly reasonable equation already.

6. We'd prefer not to add PREP (since many students take prep courses because they did poorly on their first shots at the SAT) or RACE (because of its redundancy with ESL and the lack of real diversity at Arcadia High). If you make a specification change, be sure to evaluate the change with our four specification criteria.

7. Either an omitted or irrelevant variable is a possibility, although GEND seems theoretically and statistically strong.

8. The unexpected sign makes us concerned with the possibility that an omitted variable is causing bias or that PREP is irrelevant. If PREP is relevant, what omission could have caused this result? How strong is the theory behind PREP?

9. This is a case of imperfect multicollinearity. Even though the VIFs are only between 3.8 and 4.0, the definitions of ESL and RACE (and the high simple correlation coefficient between them) make them seem like redundant variables. Remember to use theory (and not statistical fit) to decide which one to drop.

10. An omitted variable or irrelevant variable is a possibility, but there are no signs of bias and this is a fairly reasonable equation already.

11. Despite the switch to the AP linear combination, we still have an unexpected sign, so we're still concerned with the possibility that an omitted variable is causing bias or that PREP is irrelevant. If PREP is relevant, what omission could have caused this result? How strong is the theory behind PREP?

12. All of the choices would improve this equation except switching to the AP linear combination. If you make a specification change, be sure to evaluate the change with our four specification criteria.

13. To get to this result, you had to have made at least three suspect specification decisions, and you're running the risk of bias due to a sequential specification search. Our advice is to stop, take a break, review Chapters 6–8, and then try this interactive exercise again.

14. We'd prefer not to add PREP (since many students take prep courses because they did poorly on their first shots at the SAT) or ESL (because of its redundancy with RACE and the lack of real diversity at Arcadia High). If you make a specification change, be sure to evaluate the change with our four specification criteria.

15. Unless you drop one of the redundant variables, you're going to continue to have severe multicollinearity.

16. From theory and from the results, it seems as if the decision to switch to the AP linear combination was a waste of a regression run. Even if there were severe collinearity between APMATH and APENG (which there isn't), the original coefficients are significant enough in the expected direction to suggest taking no action to offset any multicollinearity.

17. On reflection, PREP probably should not have been chosen in the first place. Many students take prep courses only because they did poorly on their first shots at the SAT or because they anticipate doing poorly. Thus even if the PREP courses improve SAT scores, which they probably do, the students who think they need to take them were otherwise going to score worse than their colleagues (holding the other variables in the equation constant). The two effects seem likely to offset each other, making PREP an irrelevant variable. If you make a specification change, be sure to evaluate the change with our four specification criteria.

18. Either adding GEND or dropping PREP would be a good choice, and it's hard to choose between the two. If you make a specification change, be sure to evaluate the change with our four specification criteria.

19. On balance, this is a reasonable regression. We'd prefer not to add PREP (since many students take prep courses because they did poorly on their first shots at the SAT), but the theoretical case for ESL (or RACE) seems strong. We're concerned that the coefficient of GEND seems larger in absolute size than those reported in the literature, but none of the specification alternatives seems remotely likely to remedy this problem. If you make a specification change, be sure to evaluate the change with our four specification criteria.

Chapter Nine: Serial Correlation

9-3. The coefficient estimates for all three orders are the same:
$\widehat{HS}_t = -28187 + 16.86P_t$. The Durbin-Watson d results differ, however:

(a) DW = 3.08
(b) DW = 3.08
(c) DW = 0.64

Note that any order change will be likely to change the DW except for the reverse order (for which DW will always be exactly the same).

9-4. (a) Reject H_0 of no positive serial correlation ($d < d_L = 1.03$).
(b) Cannot reject H_0 of no positive serial correlation ($d > d_U = 1.25$).
(c) Inconclusive ($d_L = 0.98 < d < 1.73 = d_U$).
(d) Inconclusive ($4 - d_U = 4 - 1.63 = 2.37 < d < 4 - d_L = 4 - 1.13 = 2.87$).
(e) Cannot reject H_0 of no positive serial correlation ($d > d_U = 1.57$).
(f) Reject H_0 of no serial correlation ($d < d_L = 1.04$).
(g) Inconclusive ($d_L = 0.90 < d < 1.99 = d_U$).

9-6. The inconclusive region has expanded because of the small sample size and the large number of explanatory variables. As a result, even if the DW = 2 you cannot conclude that there is no positive serial correlation.

9-7. (a) $d_L = 1.44$; DW = 0.81 < 1.44, so we'd reject the null hypothesis of no positive serial correlation.
(b) This is not necessarily a sign of pure serial correlation. It's reasonable to think that residuals from the same country would have more in common than would residuals from other countries (that is, the model could be consistently underestimating for France and overestimating for Canada, producing six positive residuals followed by six negative residuals). As a result, the DW for such pooled datasets will at times give indications of serial correlation when it does not indeed exist. The appropriate measure is the Durbin-Watson *d* for each country taken individually, since the order of the countries will influence the overall DW statistic, and that order is arbitrary.
(c) If the serial correlation is impure, then a variable needs to be added to the equation to help distinguish better between the countries. If the serial correlation is judged to be pure, however, then generalized least squares might be applied one country at a time. It is possible to specify different first-order serial correlation coefficients for each country and then estimate one pooled regression equation.

9-8. (a) Except for the first and last observations in the sample, the DW test's ability to detect first-order serial correlation is unchanged.
(b) GLS can be applied mechanically to correct for serial correlation, but this procedure generally does not make sense; this time's error term is now hypothesized to be a function of *next* time's error term.

(c) First-order serial correlation in data that have been entered in reverse chronological order means that this time's observation of the error term is a function of next time's, which would be very unusual. This might occur if decision makers are able to accurately predict and adjust to future random events before they occur (which would be the case in a world of rational expectations and perfect future information).

9-9. (a) An outlier in the residuals can occur even if no outlier exists in the dataset if all the Xs are very low (or very high) simultaneously, producing an unusually low or high $\hat{Y}$. In such a situation, $\hat{Y}$ would be dramatically lower (or higher) than Y.

When an extreme outlier exists in the residuals, the Durbin-Watson test will not necessarily produce an accurate measure of the existence of serial correlation because the outlier will give the appearance of severe negative serial correlation. That is, there will be a large $(e_t - e_{t-1})$ of one sign followed by a large $(e_t - e_{t-1})$ of the opposite sign, so the two large squared terms will move the DW dramatically toward four. In such a circumstance, some researchers will drop the outlier from the DW calculation (but not from the dataset). A one-time dummy equal to one in the observation with the outlier residual will solve this problem by in essence setting the residual equal to zero; this is almost (but not quite) the same as dropping the observation. Neither solution is particularly attractive.

9-10. (a) $\beta_0^* = \beta_0(1-\hat{\rho})$, so to get β_0, divide β_0^* by $(1 - \hat{\rho})$.

(b) To account for the fact that the equation was estimated with GLS.

(c) $\hat{\beta}_0 = 23.5/(1 - 0.80) = 117.5$.

(d) The equations are inherently different, and different equations can have drastically different constant terms, because $\hat{\beta}_0$ acts as a "garbage collector" for the equation it is in. As a result, we should not analyze the estimated values of the constant term.

9-11. (a) As we've mentioned, we prefer a one-sided Durbin-Watson d test, so with K = 3 and N = 40, the 5% critical values are $d_L = 1.34$ and $d_U = 1.66$. Since DW = 0.85 is less than D_L, we can reject the null hypothesis of no positive serial correlation.

(b)

Coefficient:	β_L	β_P	β_W
Hypothesized sign:	+	+	+
t-value:	0.04	2.6	3.0
$t_C \cong 2.423$ (1% one-sided with 40 − closest to 36 in Table B-1 − d.f.)	do not reject	reject	reject

(c) The estimated coefficient of P looks reasonable in terms of size and significance, but the one for L looks pathetically small. We would never expect such similar variables to have such dramatically different coefficients. Many students will want to drop L, pointing out that the Lakers "almost always play well," so fans may not pay much attention to exactly how the Lakers are doing at any given point. We'd guess that a long losing streak would show the true relevance of this variable, however.

(d) Pure serial correlation is certainly a possibility, but the fact that some fans "are most interested in games played late in the season" implies that an omitted variable with a temporal pattern exists. We'd want to include such a variable before concluding that pure serial correlation exists.

(e) We prefer dropping the first observation to including zeroes for L and P, but an even better alternative might be to use last season's winning percentages as proxies for this season's for opening day (or even a few games thereafter). While opening day might have always sold out in the past, there is no guarantee that it always will be sold out in the future.

9-12. (a)

Coefficient	β_1	β_2	β_3	β_4	β_5
Hypothesized sign:	+	+	+	+	+
Calculated *t*-score:	3.93	2.41	5.88	1.64	1.11
$t_C = 1.668$, so:	sig.	sig.	sig.	insig.	insig.

(b) $d = 1.27$, which is less than the 5% $d_L = 1.51$, so we can reject the null hypothesis of no positive serial correlation.

(c) We're concerned about the possibilities of an omitted variable or positive serial correlation, but we're are not concerned about an irrelevant variable or multicollinearity. Both PROM and TEAM are theoretically strong variables with estimated coefficients that are in the expected direction and have reasonably large *t*-scores, so we would not consider dropping either variable from the equation even though their coefficients are not significantly different from zero in the expected direction.

(d) ATT = 34712 + 4576 MANNY + 3330 PM + 6644 WKND + 3314 PROM + 5545 TEAM

	(1216)	(934)	(1085)	(926)	(5866)
t = 3.76	3.57	6.12	3.58	0.95	
N = 80	$\hat{P} = 0.276$	$\bar{R}^2 = 0.638$			

(e) ATT = 32349 + 3763 MANNY + 2057 PM + 5621 WKND + 3917 PROM + 12121TEAM

(975)	(1065)	(1039)	(1013)	(5670)
t = 3.86	1.93	5.41	3.87	2.14

+ 3182 RIVAL
(1035)
t = 3.08

N = 81 DW = 1.52 $\bar{R}^2 = 0.592$

(f) Our guess is that the serial correlation in the original equation was impure, so we prefer the equation with RIVAL. This conclusion is supported by the inconclusive DW in the RIVAL equation and also by the fact that teams in baseball tend to play three (or so) games in a row against the same opponent. Given this pattern, an omitted variable that was related to the opponent would generate a residual grouping that resembled serial correlation.

(g) On balance, the estimates of β_{MANNY} are indeed robust, so we'd estimate that Manny Ramirez brought in more than 4,000 extra fans per game. Over 25 home games, that's an extra 100,000 fans, and if each fan brings in a net of $40, that's an additional $4 million dollars of profit.

9-13. (a) This is a cross-sectional dataset and we normally wouldn't expect autocorrelation, but we'll test anyway since that's what the question calls for. DL for a 5% one-sided, K = 3, test is approximately 1.61, substantially higher than the DW of 0.50. (Sample sizes in Table B-4 only go up to 100, but the critical values at those sample sizes turn out to be reasonable estimates of those at 200.) As a result, we can reject the null hypothesis of no positive serial correlation, which in this case seems like evidence of impure serial correlation caused by an omitted variable or an incorrect functional form.

(b)

Coefficient:	β_G	β_D	β_F
Hypothesized sign:	+	+	–?
t-value:	3.5	7.0	–2.5
$t_C = 1.645$ (5% one-sided with infinite d.f.)	reject	reject	reject

We certainly have impure serial correlation. In addition, some students will conclude that F has a coefficient that is significant in the unexpected direction. (As it turns out, the negative coefficient could have been anticipated because the dependent variable is in percentage terms but F is in aggregate terms. We'd guess that the more food a pig eats, the bigger it is, meaning that its chances of growing at a high *rate* are low, thus the negative sign.)

(c) The coefficient of D is significant in the expected direction, but given the problems with this equation, we'd be hesitant to conclude much of anything just yet.

(d) In this case, the accidental ordering was a lucky stroke (not a mistake), because it allowed us to realize that younger pigs will gain weight at a higher rate than their older counterparts. If the data are ordered by age, positive residuals will be clustered at one end of the dataset, while negative ones will be clustered at the other end, giving the appearance of serial correlation.

9-14. (a) Equation 9.26:

Coefficient	β_1	β_2	β_3
Hypothesized sign:	+	+	+
Calculated *t*-score:	0.76	14.98	1.80
$t_C = 1.721$, so:	insig.	sig.	sig.

Equation 9.27:

Coefficient	β_1	β_2
Hypothesized sign:	+	+
Calculated *t*-score:	1.44	28.09
$t_c = 1.717$, so:	insig.	sig.

(*Note:* The authors explain a positive sign for $\hat{\beta}_{SP}$ by stating that the Soviet leadership became "more competitive" after 1977, leading the USSR to increase defense spending as SP increased.)

(b) All three statistical specification criteria imply that SP is a relevant variable: $\bar{R}^2$ increases when SP is added, SP's coefficient is significantly different from zero, and the estimated coefficient of 1nSP to be positive, most readers would expect that the sign would be negative (an idea supported by the fact that the authors obtained a negative sign for $\hat{\beta}_{SP}$ for the subset of the sample from 1960 to 1976) and that Equation 9.26 therefore has a significant unexpected sign caused by an omitted variable. No matter which sign you expect, however, SP cannot be considered irrelevant.

(c) For both equations, DW is far below the critical value for a 5% one-sided test, so we can reject the null hypothesis of no positive serial correlation. (For Equation 9.26, 0.49 < 1.12, and for Equation 9.27, 0.43 < 1.21.) This result makes us worry that $\hat{\beta}_{SP}$'s *t*-score might be inflated, making it more likely that SP is an irrelevant variable.

(d) Such a small improvement in the DW statistic is no evidence whatsoever that the serial correlation is impure.

(e) Just as we suspected, running GLS makes $\hat{\beta}_{SP}$ insignificant, making it even more likely that SP is an irrelevant variable.

9-15. (a) With a 1%, one-sided test and N = 19, K = 1, the critical values are $d_L = 0.93$ and $d_U = 1.13$. Since $d = 0.48 < 0.93$, we can reject the null hypothesis of no positive serial correlation. (Impure serial correlation caused by an incorrect functional form tends to be positive.)

(b) See the answer to Exercise 1-6(c).

(c) $1.22 > 1.13$, so we can't reject the null hypothesis.

(d) 9.30, but, as we'll learn in Chapter 13, neither equation is perfect because the $\hat{P}_S$ are not limited by zero and one, even though in theory they should be.

Chapter Ten: Heteroskedasticity

10-4. (a) At the 1% level, $t_C = 2.787$; reject the null hypothesis of homoskedasticity.

(b) At the 1% level, $t_C = 4.032$; $t = 1.30$; cannot reject null hypothesis of homoskedasticity.

(c) It depends on the underlying theory that led you to choose Z as a good proportionality factor. If you believe that the absolute value of Z is what makes the variance of ϵ large, then there is no difference between −200 and +200. On the other hand, if you believe that the *relative* value of Z is important, then you are forced to add a constant (greater than 200) to each Z (which changes the nature of Z) and run the Park test.

10-5. (a) $\widehat{CO}_i = 1273.2 + 0.72I_i \qquad \bar{R}^2 = 0.97$

(0.044)

$t = 16.21$

where: CO = average consumption

I = average income.

(b) $\widehat{\ln(e_i^2)} = 29.54 - 2.34(\ln I_i) \quad \bar{R}^2 = 0.39$

(0.94)

$t = -2.49$

$t_C = 2.365$ at the 5% level (two-tailed) so we can reject the null hypothesis of homoskedasticity.

(c) The White test confirms the Park test result of heteroskedasticity.

(d) Most econometricians would switch to HC standard errors if the Park test or White test indicated heteroskedasticity. In this case, however, there's another reason for considering HC standard errors. The ranges of the income brackets are not constant in Ando and Modigliani's dataset, so the variables are means of ranges of differing widths. Thus it would seem reasonable to think that different range widths might produce different variances for the error term, making heteroskedasticity even more likely.

10-6. (a) LIKELY: the number of professors, the number of undergraduates.

(b) LIKELY: aggregate gross investment, population.

(c) LIKELY: U.S. disposable income, population, and, less likely but still possible, U.S. per capita disposable income.

10-7. $R^2 = 0.122$, N = 33, so $NR^2 = 4.026 < 21.7$ = the critical 1% Chi-square value with 9 d.f.; so we cannot reject the null hypothesis of homoskedasticity. Thus both tests show evidence of heteroskedasticity.

10-8. $NR^2 = 33.226 > 15.09$ = critical chi-square value, so reject H_0 of homoskedasticity. Thus, both tests agree.

10-9. (a) Multicollinearity and heteroskedasticity (but not positive serial correlation) appear to exist. We'd tackle the multicollinearity first. Since the heteroskedasticity could be impure, you should get the best specification you can before worrying about correcting for heteroskedasticity.

(b) For all intents and purposes, the two equations are identical. Given that, and given the reasonably strong *t*-score of STU, we'd stick with Equation 10.22. Note that the ratio of the FAC/STU coefficients is far more than 10/1 in Equation 10.22. This means that Equation 10.22 overemphasizes the importance of faculty compared to Equation 10.23. (On second thought, what's wrong with overemphasizing the importance of faculty?)

(c) Both methods show evidence of heteroskedasticity. For instance, if TOT = Z, the Park test t = 4.85 > 2.67, the critical *t*-value for a two-sided, 1% test with 57 degrees of freedom (interpolating).

(d) There are many possible answers to this question, including HC standard errors, but the interesting possibility might be to reformulate the equation, using SAT and STU/FAC (the student/faculty ratio) as proxies for quality:

$$\widehat{VOL/TOT_i} = 0.067 + 0.00011SAT_i - 0.0045STU_i/FAC_i$$
$$(0.00007) \qquad (0.0013)$$
$$t = 1.59 \qquad -3.44$$
$$\bar{R}^2 = N\ .19 = DW \quad 60 = 2.11$$

10-10. (a)

Coefficient	β_1	β_2	β_3	β_4
Hypothesized sign:	+	+	+	+
Calculated *t*-value:	7.62	2.19	3.21	7.62
$t_C = 1.645$ (5% level), so:	sig.	sig.	sig.	sig.

(b) Some authors suggest the use of a double-log equation to avoid heteroskedasticity because the double-log functional form compresses the scales on which the variables are measured, reducing a 10-fold difference between two values to a 2-fold difference.

(c) A reformulation of the equation in terms of output per acre (well, stremmata) would likely produce homoskedastic error terms.

10-11. (a)

Coefficient:	β_W	β_U	β_{lnP}
Hypothesized sign:	+	−	−?
t-value:	10.0	−1.0	−3.0
$t_C = 1.66$ (5% one-sided with 90 d.f.—interpolating)	reject	do not reject	reject

(b) We disagree, mainly because econometrics cannot "prove" that discrimination is the cause of any differences in employment. Less importantly, we disagree that the nondiscriminatory expected value of the coefficient of W is 1.00; for starters, a constant term and other variables are in the equation.

(c) Heteroskedasticity seems reasonably unlikely, despite the cross-sectional nature of the dataset, because the dependent variable is stated in per capita terms.

(d) The two-sided 1% critical *t*-value is approximately 2.64 (interpolating), so we cannot reject the null hypothesis of homoskedasticity.

(e) The theory behind P or lnP seems quite weak (despite its significance). Our preference would be to change P to a non-aggregate measure, for instance the percentage of the population that is black in the *i*th city, or some measure of the unemployment benefits available in the *i*th city.

10-12 (a) Stock and Watson accurately describe the standard practice of many experienced econometricians.

(b) Stock and Watson are entirely correct, but it's rare to find datasets with absolutely no heteroskedasticity. As a result, HC standard errors can be different from OLS standard errors in equations where the Park and White tests do not indicate heteroskedasticity.

(c) We think that it's crucial that beginning econometricians understand what heteroskedasticity is and how to combat it, so we continue to see benefits in covering heteroskedasticity in an elementary text.

10-13. (a)

Coefficient:	β_P	β_I	β_Q	β_A	β_S	β_T
Hypothesized sign:	−	+	+	+	−	+?
t-value:	–0.97	6.43	3.62	1.93	1.6	–2.85
$t_C = 1.684$ (5% one-sided with 40 d.f., closest to 43)	do not reject	reject	reject	reject	do not reject	do not reject

The last two variables cause some difficulties for most students when hypothesizing signs. Our opinion is that having more suburban newspapers should hurt metropolitan newspaper circulation but that the number of television stations is a measure more of the size of a city than of the competition a newspaper faces. By the way, we see Q as a proxy for quality and A as an endogenous variable (note that the authors did indeed estimate the equation with 2SLS, a technique beyond the scope of this chapter).

(b) Heteroskedasticity seems extremely likely, since larger cities will have larger newspaper circulation, leading to larger error term variances. Using a two-sided 1% test, we can reject the null hypothesis of homoskedasticity since 3.13 > 2.66, the critical *t*-value with 60 degrees of freedom (closest to 48).

(c) Heteroskedasticity, multicollinearity, and omitted variables all seem likely.

(d) While it's tempting to divide the equation through by population (or reformulate the equation by making the dependent variable per capita circulation), this would dramatically lessen the equation's usefulness. Instead, we would attempt to improve the specification. Reasonable answers would include attempting to reduce some of the multicollinearity (redundancy) among the independent variables, trying to find a better measure of quality than the number of editorial personnel or substituting the number of major metropolitan newspaper competitors for S and T.

10-14. (a) Heteroskedasticity is still a theoretical possibility. Young pigs are much more likely to grow at a high percentage rate than are old ones, so the variance of the error terms for young pigs might be greater than that of the error terms for old pigs.

(b) Yes, $|-6.31|$ is greater than the two-tailed 1% t_c of 2.576.

(c) An analysis of the sign of the coefficient can be useful in deciding how to correct any heteroskedasticity. In this case, the variance of the error term decreases as the proportionality factor increases, so dividing the equation again by weight wouldn't accomplish much.

(d) One possibility would be to regroup the sample into three subsamples by age and rerun the equation. This is an unusual solution but since the sample is so large, it's a feasible method of obtaining more homogeneous groups of pigs.

10-15. (a) To test for serial correlation, first run:

$$\hat{S}_t = 0.73 + 0.22 I_t + 0.46 \ln(1 + V_t)$$
$$(0.05) \quad (0.16)$$
$$t = 4.50 \quad 2.85$$

N = 58 (monthly) $\bar{R}^2$ DW 0.556 = 1.54

Since DW = 1.54, the Durbin-Watson test is inconclusive at the 5% one-sided level. Lott and Ray, the source of these data, reach the same inconclusion but with slightly different numbers. This means that there is a chance that we transcribed the dataset incorrectly. If so, comparability with Lott and Ray is reduced, but the quality of the exercise for students is undiminished.

(b) As mentioned in the text, we do not recommend running GLS if the DW is in the inconclusive range. Our major reason is that a poor estimate of rho can introduce bias into an equation while pure serial correlation will not. This is especially true when the *t*-scores are not being used to decide whether to drop a variable, as is the case in this example.

(c) A mere doubling in the size of the dependent variable should not, in and of itself, cause you to be concerned about heteroskedasticity in a time-series equation. If the dependent variable had gone up ten times, then heteroskedasticity (or nonstationarity, depending on the situation) would be a real concern.

(d) A Park test with T as a proportionality factor produces a *t*-score of 0.45, providing no evidence whatsoever of heteroskedasticity. A White test produces an R^2 of 0.20, for an NR^2 of 11.6, which indicates heteroskedasticity at the 5% level but not at the 1% level we often use for heteroskedasticity tests.

(e) A Park test with T as a proportionality factor produces a *t*-score of 0.99, once again providing no evidence of heteroskedasticity.

(f) Our first instinct would be to use HC standard errors, but we'd do so only after investigating the possibility of nonstationarity, to be discussed in Chapter 12. Nonstationarity is a completely reasonable concern in a time-series study of the Brazilian black market for dollars.

Chapter Eleven: Running Your Own Regression Project

Hints for The Housing Price Interactive Exercise

The biggest problem most students have with this interactive exercise is that they run far too many different specifications "just to see" what the results look like. In our opinion, all but one or two of the specification decisions involved in this exercise should be made before the first regression is estimated, so one measure of the quality of your work is the number of different equations you estimated. Typically, the fewer the better.

As to which specification to run, most of the decisions involved are matters of personal choice and experience. Our favorite model on theoretical grounds is:

$$\underset{+}{P} = f(\underset{+}{S}, \underset{-}{N}, \underset{-}{A}, \underset{+}{A^2}, \underset{+}{Y}, \underset{+}{CA})$$

We think that BE and BA are redundant with S. In addition, we can justify both positive and negative coefficients for SP, giving it an ambiguous expected sign, so we'd avoid including it. We would not quibble with someone who preferred a linear functional form for A to our quadratic. In addition, we recognize that CA is quite insignificant for this sample, but we'd retain it, at least in part because it gets quite hot in Monrovia in the summer.

As to interactive variables, the only one we can justify is between S and N. Note, however, that the proper variable is not $S \cdot N$ but instead is $S \cdot (5 - N)$, or something similar, to account for the different expected signs. This variable turns out to improve the fit while being quite collinear (redundant) with N and S.

In none of our specifications did we find evidence of serial correlation or heteroskedasticity, although the latter is certainly a possibility in such cross-sectional data.

Chapter Twelve: Time-Series Models

12-3. (a) $\hat{Y}_t \approx 13.0 + 12.0X_t + 0.48X_{t-1} + 0.02X_{t-2}$ (smoothly decreasing impact)

(b) $\hat{Y}_t \approx 13.0 + 12.0X_t + 0.96X_{t-1} + 0.08X_{t-2} + 0.01X_{t-3}$ (smoothly decreasing impact)

(c) $\hat{Y}_t \approx 13.0 + 12.0X_t + 24.0X_{t-1} + 48.0X_{t-2} + \cdots$ (explosively positive impact)

(d) $\hat{Y}_t \approx 13.0 + 12.0X_t + 4.8X_{t-1} + 1.92X_{t-2} - \cdots$ (damped oscillating impact)

(e) $0 < \lambda < 1$

12-4. (a)

Coefficient	β_{Pt}	β_{Pt-1}	β_U
Hypothesized sign:	+	+	−
Calculated *t*-value:	4.55	0.06	3.89
$t_c = 1.341$, so:	sig.	insig.	sig.

(b) The hypothesis being tested here is that the impact of a change in price on wages is distributed over time rather than instantaneous. Such a distributed lag (in this case ad hoc) could occur because of long-term contracts, slowly adapting expectations, and so forth. P_{t-1} is extremely insignificant in explaining W, but it's not obvious that it should be dropped from the equation. Collinearity might be the culprit, or the lag involved may be more or less than a year. In the latter case, it would not be a good idea to test many different lags on the same dataset, but if another dataset could be developed, such tests (scans) would probably be useful.

(c) The equation would no longer be an ad hoc distributed lag.

12-5. (a) $\widehat{SALES} = -243 + 5.2AD_t + 1.9AD_{t-1} + 3.1AD_{t-2} + 1.0AD_{t-3} + 3.3AD_{t-4}$

(b) $\widehat{SALES} = -38.86 + 2.98AD_t + 0.79SALES_{t-1}$

The lag structure in the ad hoc distributed lag equation makes no economic sense, because the estimated coefficients don't follow the smoothly declining pattern that economic theory would suggest and that results from using a dynamic model.

12-6. $LM = NR^2 = 24 * 0.005622 = 0.135 < 3.84 = 5\%$ critical chi-square value with one degree of freedom, so we cannot reject the null hypothesis of no serial correlation.

12-7. (a) Second-order serial correlation means that the current observation of the error term is a function of the observations of the error term from the previous two time periods.

(b) $e_t = a_0 + a_1X_t + a_2Y_{t-1} + a_3e_{t-1} + a_4e_{t-2} + u_t$
There would be 2 degrees of freedom in the test because there are two restrictions in the null hypothesis ($a_3 = a_4 = 0$).

(c) $\hat{e}_t = -11.8 - 0.22\ A_t + 0.04S_{t-1} - 0.06e_{t-1} - 0.25e_{t-2}$

$R^2 = N \quad 0.066 = 23$ (1978–2000)

$LM = NR^2 = 23 * 0.066 = 1.52 < 5.99$, the 5% critical Chi-square value with 2 d.f., so we cannot reject the null hypothesis of no second-order serial correlation.

12-8. An F-test with I Granger causing Y, F = [(307532 − 263343)/4]/[263343/19] = 0.80. Since this observed F-value is less than the critical F-value of 2.90 (5% level with 4 degrees in the numerator and 19 degrees in the denominator), we cannot reject the null hypothesis that the coefficients of the lagged Is are jointly zero.

An F-test with Y Granger causing I, F = [(175642 − 136493)/4]/[136493/19] = 1.36. Since the observed F-value is greater than the critical *F*-value of 2.90, we cannot reject the null hypothesis that the coefficients of the lagged Ys are jointly equal to zero. Since neither null hypothesis is rejected, neither GDP nor investment appears to Granger-cause the other.

12-9. We suggest that the farmers rethink either the form of their equation or their expectations. Their current equation is a dynamic model, so it posits that corn growth is a distributed lag function of rainfall, a not unreasonable idea. However, lambda is restricted to between zero and one, so the likelihood of observing a negative lambda is small, and in theory a negative lambda would be very difficult to explain for the impact of rainfall on corn.

12-10. (a) Y: $t = 6.54$ and $t_C = 3.12$, so we cannot reject the null hypothesis of nonstationarity at the 5% level (the sign of t does not agree with H_A).

(b) PC: $t = -0.36$ and $t_c = 3.12$, so we cannot reject the null hypothesis of nonstationarity at the 5% level.

(c) PB: $t = 0.02$ and $t_C = 3.12$, so we cannot reject the null hypothesis of nonstationarity at the 5% level.

(d) YD: $t = 12.55$ and $t_C = 3.12$, so we cannot reject the null hypothesis of nonstationarity at the 5% level (the sign of t does not agree with H_A).
Thus all the variables in the chicken demand equation are nonstationary.

12-11. (a) $t = 11.17$ and $t_C = 3.12$, so we cannot reject the null hypothesis of nonstationarity for Y.

(b) $t = -0.78$ and $t_C = 3.12$, so we cannot reject the null hypothesis of nonstationarity for r.

(c) $t = 16.04$ and $t_C = 3.12$, so we cannot reject the null hypothesis of nonstationarity for CO.

(d) $t = 2.45$ and $t_C = 3.12$, so we cannot reject the null hypothesis of nonstationarity for I.
All four variables appear to be nonstationary, at least at the 2.5% level. This is a bit surprising, because interest rate variables often are stationary. It's not a coincidence that the interest rate is the only variable to have a negative *t*-score.

12-12. (a) Such a split result is not at all unusual in correctly done applications of Granger causality when there is no real underlying causal relationship or when the causal relationship switches under various circumstances.

(b) Based on this research, it's impossible to draw any general conclusions about the causal relationship between economic growth and democracy for two good reasons. First, the results are inconclusive.

Second, reaching a conclusion about causality involves more than just the results of a Granger causality test, so even if the results for all 32 countries had provided evidence of Granger causality in the same direction, we would not feel justified in drawing a conclusion about the relationship between economic growth and democracy.

(c) An interesting next step in this research project would be to see if national characteristics shed any light on the results of the Granger causality test. To do this, we'd research the literature to find the attributes of a country that might impact the direction of the Granger causality, and then we'd estimate a model where the dependent variable would be the direction of the Granger causality and the independent variables would be these national attributes. Since the dependent variable in this case would be a dummy variable, we'd probably estimate the equation with a logit (or multinomial logit), and such dummy dependent variable techniques will be covered in the next chapter.

Chapter Thirteen: Dummy Dependent Variable Techniques

13-3. (a) This equation is a linear probability model, and as such it can encounter at least two problems in addition to any "normal" specification problems it might have: $\bar{R}^2$ is not an accurate measure of the overall fit, and $\hat{Y}_i$ is not bounded by 0 and 1.

(b) Some might question the importance of PV in the equation, and others might suggest adding variables to measure recent changes in profits or book value, but our inclination would be to switch to a logit before analyzing the specification much further.

(c) The best way to win this argument is to note that the linear probability model produces nonsensical forecasts (greater than 1.0 and less than 0.0).

13-4. Start with $\ln[D/(1 - D)] = Z$ and take the anti-log, obtaining $D/(1 - D) = e^Z$. Then cross-multiply and multiply out, which gives $D = e^Z - De^Z$. Then solve for $D = e^Z/(1 + e^Z)$. Finally, multiply the right-hand side by e^{-Z}/e^{-Z}, obtaining $D = 1/(1 + e^{-Z})$.

13-5. (a) $\hat{D}_i > 1$ if $X_i > 7$ and $\hat{D}_i < 0$ if $X_i < -3$.

(b) $\hat{D}_i > 1$ if $X_i < 10$ and $\hat{D}_i < 0$ if $X_i > 15$.

(c) $\hat{D}_i > 1$ if $X_i > 6.67$ and $\hat{D}_i < 0$ if $X_i < 3.33$.

(d–f) It won't take long for students to confirm that with a logit, $\hat{D}_i$ is never greater than 1 or less than 0.

13-6. (a)

Coefficient	β_{UNIT}	β_{ALCO}	β_{YEAR}	β_{GREEK}
Hypothesized sign:	+	–	–	–
Calculated *t*-score:	0.84	–1.55	–8.25	–1.38
$t_C = 1.289$, so:	insig.	insig.	sig.	sig.

(b) Defining YEAR this way constrains the coefficients of three classes to be related to each other when there is no reason to expect that to be the case. For example, the definition forces a junior to be exactly 1.33 times more likely to live off campus than a sophomore when there is no reason to expect this relationship. In fact, we'd expect seniors to be by far more likely to live off campus than juniors or sophomores, and this definition wouldn't allow that to happen.

A much better approach would have been to define two dummy variables, one equal to 1 for seniors (0 otherwise) and one equal to 1 for juniors (0 otherwise), which would make being a sophomore the omitted condition. We'd expect a positive coefficient for each variable, with the coefficient of senior being substantially larger than the coefficient of junior.

(c) The estimate of β_{ALCO} tells us that for each additional night (per week) that a student consumes alcohol, the log of the odds that that student will live on campus will decrease by 0.13, holding constant the other independent variables in the equation. If we divide 0.13 by 4, this turns out to be equivalent to saying that that for each additional night (per week) that a student consumes alcohol, probability of that student living on campus will decrease by 3.25 percentage points, holding constant the other independent variables in the equation. This is a little lower than we might have expected, but it certainly is plausible.

(d) We'd first fix the definition of YEAR as suggested in part (b). After that, we'd add one of a number of potentially relevant variables, for instance the gender of the *i*th student or whether the *i*th student's home was within 10 miles of campus.

13-7. (a) If $D_i = 2$, then the logit computer program will almost surely balk at taking the log of a negative number (–2). As mentioned in the text, however, logit programs iterate using Equation 13.7 (or a version thereof), so it's possible that a software package *could* produce some sort of estimates. (We've never seen one, however.)

(b) With a linear probability model, the answer is unambiguous. The estimated coefficients of WN and ME will double in size. If we divide each coefficient by 2, the theoretical meanings will be unchanged from the original model.

13-8. (a) There are only two women in the sample over 65. Because this causes a near-singular matrix, many Logit programs will not be able to estimate this equation or will produce estimates quite different from ours, which was estimated with EViews.

(b) We prefer Equation 13.13 because AD gives every appearance of being an irrelevant variable, at least as measured by the four criteria developed in Chapter 6.

13-9. In all three models, we find evidence that A is an irrelevant variable. The coefficient of A is insignificantly different from zero in all three models, and $\bar{R}_p^2$ falls when A is added to all three models. (Note that in some datafiles, D = "J".)

(a) $\hat{D}_i = -0.22 - 0.38\, M_i - 0.001\, A_i + 0.09\, S_i$

$(0.16) \quad (0.007) \quad (0.04)$

$t = -2.43 \quad -0.14 \quad 2.42$

$\bar{R}^2 = 0.29 \quad N = 30 \quad \bar{R}_p^2 = 0.806$

(b) $\widehat{\ln[D_i/(1-D_i)]} = -5.27 - 2.61\ M_i - 0.01\ A_i + 0.67\ S_i$

$$\begin{array}{rccc} & (1.20) & (0.04) & (0.32) \\ t = & -2.17 & -0.25 & 2.10 \\ \bar{R}_p^2 = & 0.76 & & \end{array}$$

(c) $\hat{Z}_i = \widehat{F^{-1}(P_i)} = -3.05 - 1.45\ M_i - 0.006\ A_i + 0.39\ S_i$

$$\begin{array}{rccc} & (0.63) & (0.02) & (0.18) \\ t = & -2.31 & -0.26 & 2.20 \end{array}$$

$\bar{R}_p^2 = 0.76$ iterations = 5 LR = 13.85

13-10. (a) All signs meet expectations except that of wait time, which we would expect to have a negative impact because a longer wait time should deter riders from taking public transportation.

(b) The fact that the estimated coefficient of walk time is larger in absolute value than that of travel time supports this hypothesis, but the large positive coefficient for wait time does not.

(c) Yes, if train commuters know train schedules and actually adjust their station arrival to minimize wait time, then setting the wait time for trains high allowed wait time to become a proxy for being the preferred mode of travel in Boston.

13-11. (a) Three.

(b) The three dependent variables are $\ln(P_u/P_c)$, $\ln(P_j/P_c)$, and $\ln(P_a/P_c)$, which are the log of the ratio of the probability of attending the choice in question divided by the probability of attending your college.

13-12. (a) The trick here is getting the expected sign right, because it won't be obvious to everyone that DISTANCE can be negative (if the patient lives farther from Cedars Sinai than he does from UCLA). Once you take this into account, it's clear that the larger DISTANCE is, the less likely the *i*th patient is to choose Cedars Sinai, so the expected sign of the coefficient is negative, and we can reject the null ($t_C = 1.645$).

(b) For every extra mile that it takes a patient to get to Cedars Sinai as compared to UCLA, the probability of that patient choosing Cedars Sinai falls by 9.5%, holding constant INCOME and OLD. [9.5% is the coefficient of distance (0.38) divided by 4.]

(c) Our guess is that most patients care about the relative distance to the two hospitals, not the absolute values of the individual distances to the hospitals.

(d) The coefficient of DISTANCE in the linear probability model is −0.072, and in the probit it is −0.226. We avoid estimating linear probability models when the dependent variable is a dummy variable because of the unlimited range of the dependent variable, so we have a strong preference for the probit in this and most other examples.

(e) The hypothesis behind this interaction term is that an elderly patient might be more likely than a younger patient to try to minimize the distance traveled to a hospital because of the limited mobility of elderly patients. Thus:

$$H_0 : \beta \geq 0$$
$$H_A : \beta < 0.$$

Sure enough, the estimated coefficient of the interaction term is negative and produces a *t*-score of −3.03, which is greater than the critical value of 1.645 and is in the expected direction, so we can reject H_0. We prefer the slope dummy logit, and all four of our specification criteria support that preference.

Chapter Fourteen: Simultaneous Equations

14-3. (a) If ε_2 decreases, Y_2 decreases and then Y_1 decreases.

(b) If ε_D increases, Q_D increases, then Q_S increases (equilibrium condition) and P_t increases. (Remember that the variables are simultaneously determined, so it doesn't matter which one is on the left-hand side.)

(c) If ε_1 increases, CO increases, and then Y increases and YD increases.

14-4. (a) the first two equations are simultaneous, but the third equation is a recursive equation that feeds into the first two, so Y_3 is not simultaneously determined.
Endogenous variables = Y_{1t}, Y_{2t}
Predetermined variables: Y_{3t}, X_{1t}, X_{1t-1}, X_{2t-1}, X_{3t}, X_{4t}, X_{4t-1}

(b) All three equations are simultaneous. Note that Y is predetermined.
Endogenous variables = Z_t, X_t, H_t
Predetermined variables: Y_t, P_{t-1}, B_t, CS_t, D_t

(c) The equations are recursive; solve for Y_2 first and use it to get Y_1.

14-5. All these cases can be shown to involve a positive correlation between the εs and the Ys.

14-6. (a) There are three predetermined variables in the system, and both equations have three slope coefficients, so both equations are exactly identified. (If the model specified that the price of beef was determined jointly with the price and quantity of chicken, then it would not be predetermined, and the equations would be underidentified.)

(b) There are two predetermined variables in the system, and both equations have two slope coefficients, so both equations are exactly identified.

(c) There are seven predetermined variables in the system, and there are three slope coefficients in both equations, so the first two equations are overidentified. Note that we don't worry about the identification properties of the third equation because it isn't part of the simultaneous system.

(d) There are five predetermined variables in the system, and there are three, two, and four slope coefficients in the first, second, and third equations, respectively, so all three equations are overidentified.

14-7. (a) A: Predetermined = $2 < 3$ = # of slope coefficients, so underidentified.
B: Predetermined = $2 = 2$ = # of slope coefficients, so exactly identified.

(b) Note that X_2 is endogenous to this system, so:
Y_1: Predetermined = $3 < 4$ = # of slope coefficients, so underidentified.
Y_2: Predetermined = $3 > 1$ = # of slope coefficients, so overidentified.
X_2: Predetermined = $3 = 3$ = # of slope coefficients, so exactly identified.

(c) Note that you can consider the change in Y to be endogenous to the system with a non-stochastic equation in which it equals $Y_t - Y_{t-1}$. Given this, there are six predetermined variables, Y_{t-1}, E_t, D_t, M_t, R_{t-1}, and G_t, so the identification properties of the four stochastic equations can be determined by using the order condition (which is necessary but not sufficient):
CO_t: Predetermined = $6 > 1$ = # of slope coefficients, so overidentified.
I_t: Predetermined = $6 > 4$ = # of slope coefficients, so overidentified.
R_t: Predetermined = $6 > 3$ = # of slope coefficients, so overidentified.
Y_t: Predetermined = $6 > 3$ = # of slope coefficients, so overidentified.

14-8. Stage one: Apply OLS to the second of the reduced-form equations:

$$Q_{St} = Q_{Dt} = \pi_0 + \pi_1 X_{1t} + \pi_2 X_{2t} + \pi_3 X_{3t} + V_{1t}$$
$$P_t = \pi_4 + \pi_5 X_{1t} + \pi_6 X_{2t} + \pi_7 X_{3t} + v_{2t}$$

Stage two: Substitute the reduced-form estimates of the endogenous variables for the endogenous variables that appear on the right side of the structural equations. This would give:

$$Q_{Dt} = \alpha_0 + \alpha_1 \hat{P}_t + \alpha_2 X_{1t} + \alpha_3 X_{2t} + u_{Dt}$$
$$Q_{St} = \beta_0 + \beta_1 \hat{P}_t + \beta_2 X_{3t} + u_{St}$$

To complete stage two, estimate these revised structural equations with OLS.

14-9. (a) $\hat{I}_t = -267 + 0.19 Y_t - 9.26 r_{t-1}$

(0.01) (11.19)

$t = 15.87 \quad -0.83$

$\bar{R}^2 = 0.956 \quad N = 32 \quad DW = 0.47$

(b) $\hat{Y}_t = -258.6 + 0.78 G_t - 0.37 NX_t + 1.52 T_t + 0.67 CO_{t-1} + 37.6 r_{t-1}$

$\bar{R}^2 = 0.999 \quad N = 32$

(c) $\hat{I}_t = -299 + 0.19 \hat{Y}_t - 9.10 r_{t-1}$

(Standard errors obtained from this estimation are biased and should be disregarded.)

(d) $\hat{I}_t = -261.5 + 0.19 \hat{Y}_t - 9.55 r_{t-1}$

(0.01) (11.2)

$t = 15.8 \quad -0.85$

$\bar{R}^2 = 0.956 \quad N = 32 \quad DW = 0.47$

14-10. (a) You don't know that OLS and 2SLS will be the same until the system is estimated with both.

(b) Not necessarily. It indicates only that the fit of the reduced form equation from stage one is excellent and that $\hat{Y}$ and Y are virtually identical. Since bias is only a general tendency, it does not show up in every single estimate. Indeed, it is possible to have estimated coefficients in the opposite direction. That is, even though positive bias exists with OLS, an estimated coefficient less than the true coefficient can be produced.

14-11. Most reasonable models of the labor market are simultaneous and therefore potentially involve simultaneity bias and should be estimated with 2SLS.

14-12. (a) The serial correlation is so severe that it can be detected by the Durbin-Watson *d* test even though that statistic is biased toward 2. $DW = 0.83 < 1.31 = d_L$ for N = 32, K = 2 at a 5% level of significance.

(b) Since the OLS and 2SLS estimates of this equation are similar, and since the serial correlation is quite severe, we'd choose to correct for serial correlation if we could correct for only one problem.

(c) One possibility is to estimate a reduced form for YD_t that includes $C0_{t-2}$ and YD_{t-1} on the right-hand side, and then substituting $\widehat{YD_t}$ into a GLS equation. This approach might be called "2SLS/GLS" since the 2SLS portion of the procedure is carried out before the GLS portion. A second possibility is to include an AR(1) term in a 2SLS model. A second possibility is to use Newey-West standard errors if your computer program's 2SLS program provides that option.

14-13. (a) OLS estimation will still encounter simultaneity bias because price and quantity are simultaneously determined. Not all endogenous variables will appear on the left-hand side of a structural equation.

(b) The direction of the bias depends on the correlation between the error term and the right–hand-side endogenous variable. If the correlation between the error term and price is positive, as it most likely is, then the simultaneity bias will also be positive.

(c) Three: stage one: $P = f(YD, W)$
stage two: $Q_D = f(\hat{P}, YD)$ and $Q_S = f(\hat{P}, W)$

(d) OLS: $\hat{Q}_D = 57.3 - 0.86P + 1.03YD$

$\hat{Q}_S = 167.5 + 3.95P - 1.42W$

2SLS: $\hat{Q}_D = 95.1 - 6.11\hat{P} + 2.71YD$

$\hat{Q}_S = 480.2 + 13.5\hat{P} - 5.50W$

14-14. (a) QU: –, –, –, +, +, +
UR: +, +, +, +, +

(b) Yes, since UR and QU are jointly determined in this system.

(c) This tells us that the UR equation is exactly identified but tells us nothing about the identification properties of the QU equation.

(d) The lack of significance makes us wonder if UR and QU are indeed simultaneously determined. We should be hesitant to jump to this conclusion, however, because: (1), the theory indicates simultaneity; (2), multicollinearity or other specification problems may be causing the insignificance; and (3), the pooled cross section/time-series dataset makes it difficult to draw inferences.

(e) Given the above reservations, we should be cautious. However, the results tend to confirm the theory that states interested in lowering their unemployment rates and lowering their budget deficits might consider lowering their unemployment benefits.

14-15. (a) All three variables are nonstationary. In Exercise 12-11, we showed that both CO_t and Y_t are nonstationary. If CO_t is nonstationary, then so too must be CO_{t-1}. YD_t and Y_t are highly correlated, so it's reasonable to think that if one is nonstationary then so too is the other. As a test of this, a Dickey-Fuller test on YD_t produces a *t*-score of 12.86, further evidence that YD_t is nonstationary.

(b) If we run a Dickey-Fuller test on the residuals, we get a *t*-score of –3.25, which is greater in absolute value that the t_c of 3.12 and which has the sign of H_A, so we can reject the null hypothesis of nonstationarity and conclude that the residuals are stationary. This implies that Equation 14.30 is reasonably cointegrated.

(c) A dynamic model distributed lag equation is more likely to be cointegrated because the lagged values of the dependent variable that appear on the right-hand side of the equation should have the same degree of nonstationarity as the dependent variable.

(d) We agree with the majority of applied econometricians who think that the concept of cointegration is unrelated to the estimation technique. As a result, we do not hesitate to recommend the use of the Dickey-Fuller test when testing 2SLS residuals for cointegration. There are those who disagree, however, by pointing out that nonstationarity in a truly simultaneous system implies that a test for cointegration should go beyond testing the residuals of just one of the equations in the system.

Chapter Fifteen: Forecasting

15-4. (a) 160.82 ± 17.53

(b) 800 ± 344.73

15-5. (a) P isn't a dummy variable. Instead, a variable whose sole function is to be multiplied by other variables so that the sign of the resulting interaction variable changes depending on the incumbent's party.

(b) The interaction variables were required because the dependent variable measures the percentage of votes won by the Democrats, but the independent variables measure items that support (or damage) public support for the incumbent party. For example, if a Democrat is in office in a time of high growth, that growth should increase the share of votes won by Democrats, so a positive sign makes sense. However, if a Republican is in office in a time of high growth, the growth should decrease the share of votes won by the Democrats, so a negative sign makes sense. Multiplying GROWTH by +1 if the incumbent is a Democrat and −1 if the incumbent is a Republican thus makes sense, and that's what multiplying by P accomplishes.

(c) $\widehat{\text{VOTE}} = 47.30 + 0.068\ \text{DUR} * \text{P} + 0.119\ \text{DOW} * \text{P} + 0.779\ \text{GROWTH} * \text{P} + \downarrow$

(0.837) (0.086) (0.244)

t = 0.08 2.09 3.20

$+ 0.014\ \text{ARMY} * \text{P} - 0.070\ \text{INFLATION} * \text{P} - 0.041\ \text{SPEND} * \text{P}$

(0.045) (0.400) (0.045)

t = 0.31 0.17 0.90

N = 21 $\bar{R}^2 = 0.59$ DW = 2.12

(d) We expect positive signs for the coefficients of the first three interaction variables and negative signs for the coefficients of the second three. Thus all the signs are as expected except for the coefficient of ARMYP. We can reject the null only for the coefficients of DOWP and GROWTHP.

(e) Plugging the actual values for 2000 into the equation, we got a forecast of 49.058, which is 2.4% less than the actual 50.265. For 2004, we got a forecast of 46.640, which is 4% less than the actual 48.586.

(f) To do this, we should estimate Equation 15.22 with data through 2004, producing:

$\widehat{\text{VOTE}} = 47.51 + 0.148\ \text{DUR} * \text{P} + 0.182\ \text{DOW} * \text{P} + 0.760\ \text{GROWTH} * \text{P} + \downarrow$

(0.770) (0.078) (0.244)

t = 0.19 2.32 3.48

$+ 0.014\ \text{ARMY} * \text{P} - 0.112\ \text{INFLATION} * \text{P} - 0.046\ \text{SPEND} * \text{P}$

(0.042) (0.367) (0.041)

t = 0.33 −0.31 −1.10

N = 23 $\bar{R}^2 = 0.60$ DW = 2.11

Plugging the actual values for 2008 into this equation, we get a forecast of 40.67, surprisingly below the share that Barack Obama actually earned.

15-6. (a) $Y_t^* = 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1$

$Y_t^{**} = 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0$ (d = 1)

(b) $Y_t^* = 0, 1, 1, 1, 1, 2, 2, 2, 3, 4, 5$

$Y_t^{**} = 1, 0, 0, 0, 1, 0, 0, 1, 1, 1$ (d = 2)

(c) $Y_t^* = 1, 3, -3, 1, -2, 1, 2, -4, 3, 0, 2$

$Y_t^{**} = 2, -6, 4, -3, 3, 1, -6, 7, -3, 2$ (d = 0)

15-7. If the answers to Exercise 15-6 were calculated correctly, then calculating "backwards" will indeed reproduce the original series.

15-8.

		Model A	**Model T**
(a)	1997	30.50	29.50
	1998	30.25	30.25
	1999	30.13	29.87
(b)	1998	31.50	28.50
	1999	30.75	30.75

(c) Model A should exhibit smoother behavior because of the negative coefficient in model T.

15-9. (a) $e_{99} = Y_{99} - \hat{Y}_{99} = 27 - 27.5 = -0.5$

(b) $Y_{100} = 0 + 1(27) - 0.5(-0.5) = 27.25$

$Y_{101} = 0 + 1(27.25) = 27.25$

$Y_{102} = 0 + 1(27.25) = 27.25$

15-10. (a) For period one, this would be an unconditional distributed lag forecast:

$$\hat{S}_{t+1} = \hat{\beta}_0 + \hat{\beta}_1 \hat{Y}_{t+1} + \hat{\beta}_2 \hat{S}_t$$

For period two, this would become a conditional distributed lag forecast:

$$\hat{S}_{t+2} = \hat{\beta}_0 + \hat{\beta}_1 \hat{Y}_{t+2} + \hat{\beta}_2 \hat{S}_{t+1}$$

(b) For both periods, this would be a conditional distributed lag forecast:

$$\hat{S}_{t+1} = \hat{\beta}_0 + \hat{\beta}_1 \hat{Y}_{t+1} + \hat{\beta}_2 \hat{S}_t$$
$$\hat{S}_{t+2} = \hat{\beta}_0 + \hat{\beta}_1 \hat{Y}_{t+2} + \hat{\beta}_2 \hat{S}_{t+1}$$

(c) Here, we'd build a (simultaneous) simulation model using the equations in part (b) in addition to something like: $Y_t = \propto_0 + \propto_1 S_t + \propto_2 T_t + \epsilon_t$

Chapter Sixteen: Experimental and Panel Data

16-3. In theory, one could design a random assignment experiment for which no additional explanatory variables would be necessary, but it's virtually impossible to imagine a natural experiment not needing such variables. There are sure to be some differences between the "control" and the "treatment" groups, and we need to account for those differences.

16-4. (a) This is a natural experiment dataset that also happens to be a panel dataset because it contains observations on the same variable from the same cross-sectional sample from two different time periods.

(b) The appropriate technique is the difference-in-differences estimator, resulting in:

$$\widehat{\Delta \text{OUTCOME}} = -2.43 - 0.73\ \text{TREATMENT}$$
$$(0.57)$$
$$t = -1.29$$
$$N = 45 \quad \bar{R}^2 = 0.015$$

(c) The estimated coefficient is almost significant in the expected direction, but the fit is terrible. This will surprise many students. However, most experienced researchers won't be surprised, because of the design of the research. In particular, it seems extremely optimistic to expect to explain cigarette consumption by state using a dummy for whether the cigarette tax rate increased as the only independent variable. Variables other than tax rates certainly play a role, as does the fact that some states increased cigarette taxes by substantially more than did others, and yet that information is lost if you limit yourself to a dummy variable, since it tells you only whether taxes increased, not the amount by which they increased.

16-5. (a) a_i represents the unobserved impact of the time-invariant omitted variables.

(b) V and ε have two subscripts because they can have different values not only for each of the i cross-sectional entities but also for each of the t time-series entities. a_i, in contrast, is invariant over time, so it can have different values only for each of the cross-sectional entities.

(c) We need to remove a_i to avoid omitted variable bias. If the impact of time-invariant omitted variables is in the error term, then we're very likely to be violating Classical Assumption III.

16-6. (a) The estimated slope is positive, which certainly runs counter to our expectations:

$$\hat{Q} = -1.41 + 0.0457\ P$$
$$(0.014)$$
$$t = 3.28$$
$$N = 4 \quad \bar{R}^2 = 0.76$$

(b) While the fit and the size of the estimated coefficients differ from those in part (a), the sign of the estimated slope coefficient continues to be unexpected.

$$\hat{Q} = -0.22 + 0.0237P$$
$$(0.014)$$
$$t = 1.64$$
$$N = 4 \quad \bar{R}^2 = 0.36$$

(c) As expected, the sign reverses.

(d) As expected, the fixed effects model is superior.

16-7. (a) $\widehat{\Delta Q} = 0.039 - 0.025\Delta P$

(0.002)

$t = -12.33$

$N = 8 \quad \bar{R}^2 = 0.53$

(b) They produce identical answers.

(c) The error term in the differencing model certainly appears to be defined in such a way as to be serially correlated.

Chapter Seventeen: Statistical Principles

17-3. The mean is 16.89 and the standard deviation is 6.43. Thus the 1999 P/E ratio was more than two standard deviations above the mean.

17-4. Because the numbers on each side are equally likely, we can reason directly that a six-sided die has an expected value of 3.5 and a four-sided die has an expected value of 2.5. Because the possibilities are more spread out on the six-sided die, it has the larger standard deviation.

17-5. Standardized scores: 1.9, 0.0, and –0.8, raw score: 90.

17-6. The z values and normal probabilities are:

$$P[x > 270] = P\left[\frac{x - \mu}{\sigma} > \frac{270 - 266}{16}\right] = P[z > 0.25] = 0.4013$$

$$P[x > 310] = P\left[\frac{x - \mu}{\sigma} > \frac{310 - 266}{16}\right] = P[z > 2.75] = 0.003$$

17-7. Mean = 500,000.
Standard deviation = 77,460.

17-8. The high school seniors who take the SAT are not a random sample because this test is taken by students who intend to go to college; these are generally students with above-average scholastic aptitude. The relationship between the fraction of a state's seniors taking the SAT and the state's average SAT score is negative. If a small fraction of the state's seniors takes the SAT, it will mostly consist of the state's best students. As the fraction of a state's students taking the SAT increases, the group of students taking the SAT is increasingly composed of weaker students, who bring down the state's average SAT.

17-9. People who have visited France for pleasure more than once during the past two years are more likely to have had good experiences than are people who visited France just once and then never returned and/or people making their first visit to France.

17-10. The mean is 299,756.2174 and the standard deviation is 107.1146. Table B-4 in Appendix B shows that with 23 − 1 = 22 degrees of freedom, the appropriate *t*-value for a 99% confidence interval is 2.819. A 99% confidence interval does include the value 299,710.5 that is now accepted as the speed of light:

$$\bar{x} \pm t^*\left(\frac{s}{\sqrt{N}}\right) = 299{,}756.2174 \pm 2.819\left(\frac{107.1146}{\sqrt{23}}\right)$$
$$= 299{,}756.2 \pm 63.0$$

17-11. The standard deviation is the square root of the variance, or 0.686, and the 95% two-sided t_c with 34 degrees of freedom is approximately 2.03, so a 95% confidence interval is 6.19 + 2.03 $(0.686/\sqrt{35})$ or 6.19 + 0.24. There is a 95% probability that a confidence interval constructed in this fashion will include the true value of the mean prediction of the population, so:

(a) No. This says nothing about how accurate or inaccurate the forecasters are.

(b) No. If anything, we might estimate that approximately 95% of the forecasts are in an interval equal to our estimate of the mean plus or minus 2 standard deviations of the individual observations:

$$6.19 + 1.96\ (0.686) \quad \text{or} \quad 6.19 + 1.34.$$

17-12. If x is N[215, 10] then for a random sample of size N = 20:

$$P[\bar{x} \le 257] = P\left[\frac{\bar{x} - \mu}{\sigma/\sqrt{N}} \ge \frac{257 - 215}{10/\sqrt{20}}\right] = P[z \ge 18.8] \approx 0$$

Dr. Frank's patients may choose to be medical patients because they have heart problems. Any trait they happen to share will then seemingly explain the heart disease; however, the standard statistical tests are not valid if these are not a random sample from the population of all people with earlobe creases.