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 × 3) = \$2934 > (204 × 2 + 36 × 2) = \$480 > (\$460 × 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.