

## CHAPTER 2

### SOLUTIONS TO PROBLEMS

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

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

$n = 8.$

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

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

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

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

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

(iv) The sum of squared residuals,  $\sum_{i=1}^n \hat{u}_i^2$ , is about .4347 (rounded to four decimal places),

and the total sum of squares,  $\sum_{i=1}^n (y_i - \bar{y})^2$ , is about 1.0288. So the  $R$ -squared from the regression is

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

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

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

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

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

From equation (2.7), we should try to get as much variation in  $hours_i$  as is feasible.

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

(iii) If preparation courses are effective,  $\beta_1$  should be positive: other factors equal, an increase in  $hours$  should increase  $sat$ .

(iv) The intercept,  $\beta_0$ , has a useful interpretation in this example: because  $E(u) = 0$ ,  $\beta_0$  is the average SAT score for students in the population with  $hours = 0$ .

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

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

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

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

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

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

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

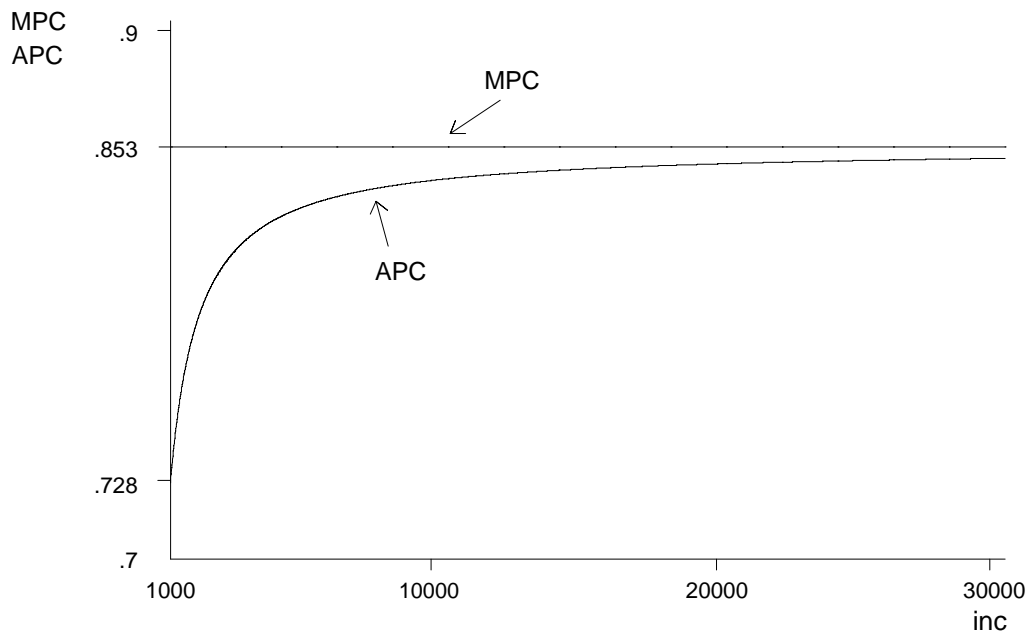
(iii) We can simply apply part (ii) because  $\log(c_1 y_i) = \log(c_1) + \log(y_i)$ . In other words, replace  $c_1$  with  $\log(c_1)$ ,  $y_i$  with  $\log(y_i)$ , and set  $c_2 = 0$ .

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

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

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

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



## SOLUTIONS TO COMPUTER EXERCISES

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

(ii) The estimated equation is

$$\widehat{prate} = 83.05 + 5.86 \text{ } mrte$$

$$n = 1,534, R^2 = .075.$$

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

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

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

**C2.3** (i) The estimated equation is

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

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

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

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

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

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

where  $\beta_1$  is the elasticity of *rd* with respect to *sales*.

(ii) The estimated equation is

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

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

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

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

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

(iii) The estimated equation is

$$\widehat{gift} = 2.01 + 2.65 \text{ mailsyear}$$

$$n = 4,268, R^2 = .0138$$

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

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