CHAPTER 2
The Simple Regression Model

SOLUTIONS TO PROBLEMS

2.1 (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.

2.3 (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 (2.19), 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

$$GPA = .5681 + .1022\, ACT$$

$$n = 8.$$

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, *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*	*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 $ACT = 20$, $GPA = .5681 + .1022(20) \approx 2.61$.

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

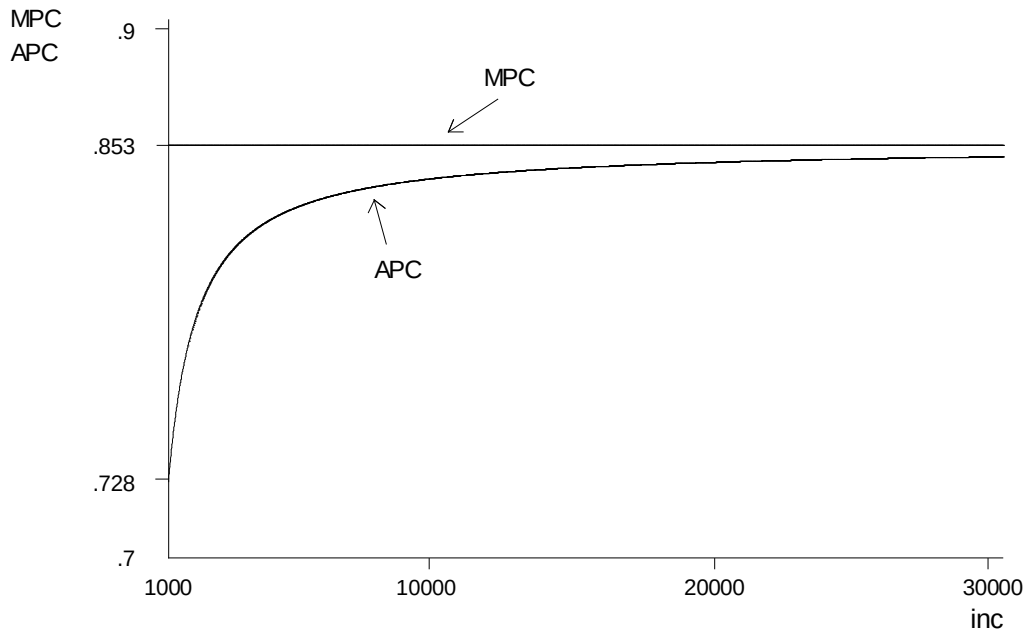
$$R^2 = 1 - \text{SSR/SST} \approx 1 - (.4347/1.0288) \approx .577.$$

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

2.5 (i) The intercept implies that when *inc* = 0, *cons* is predicted to be negative \$124.84. This, of course, cannot be true, and reflects the 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: $cons = -124.84 + .853(30{,}000) = 25{,}465.16$ dollars.

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



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

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

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

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

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

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

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

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

(iii) We can simply apply part (ii) because $\log(c_1 y_i) = \log(c_1) + \log(y_i)$. In other words, replace c_1 with log(c_1), replace 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 $\tilde{\beta}_0 = \hat{\beta}_0 - \log(c_2)\hat{\beta}_1$.

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

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

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

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

2.13 (i) Since x_i is a binary variable, it is equal to either 0 or 1. Thus, the number of observations with $x_i = 0$ will be $n_0 = \sum(1 - x_i)$ since the value in the summation is equal to 1 whenever $x_i = 0$ and equal to 0 whenever $x_i = 1$. Similarly, $n_1 = \sum x_i$ will give us the number of observations with $x_i = 1$ since we are only going to be counting instances in which x_i is not equal to 0.

We know that $\bar{x} = \frac{1}{n}\sum x_i$. We also have shown that $\sum x_i = n_1$. Thus, $\bar{x} = \frac{n_1}{n}$.
We can write $1 - \bar{x} = 1 - \frac{n_1}{n} = \frac{n - n_1}{n} = \frac{n_0}{n}$ since $n_0 + n_1 = n$.
$\bar{x} = \frac{n_1}{n}$ tells us the fraction of observations for which $x_i = 1$.

(ii) We know that $\bar{y}_0 = n_0^{-1}\sum_{i=1}^{n_0} y_i$, where we are just looking at values of y_i for which $x_i = 0$. One way we can formalize this is by recognizing that $(1 - x_i)y_i = y_i$ if $x_i = 0$ and 0 otherwise. Thus, $\bar{y}_0 = n_0^{-1}\sum_{i=1}^{n_0} y_i = n_0^{-1}\sum_{i=1}^{n}(1 - x_i)y_i$.
Similarly, we can show that $\bar{y}_1 = n_1^{-1}\sum_{i=1}^{n_1} y_i = n_1^{-1}\sum_{i=1}^{n} x_i y_i$, since $x_i y_i = y_i$ if $x_i = 1$ and 0 otherwise.

(iii) We can express $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i = \frac{1}{n}\left(\sum_{i=1}^{n_0} y_{0,i} + \sum_{i=1}^{n_1} y_{1,i}\right) = \frac{n_0}{n}\bar{y}_0 + \frac{n_1}{n}\bar{y}_1$
We have shown that $1 - \bar{x} = \frac{n_0}{n}$ and $\bar{x} = \frac{n_1}{n}$. Thus, $\bar{y} = (1 - \bar{x})\bar{y}_0 + \bar{x}\bar{y}_1$

(iv) If x_i is binary, then $x_i^2 = x_i$. Thus, $n^{-1}\sum x_i^2 - \bar{x}^2 = n^{-1}\sum x_i - \bar{x}^2 = \bar{x} - \bar{x}^2 = \bar{x}(1-\bar{x})$

(v) First note that $n^{-1}\sum_{i=1}^{n} x_i y_i = n^{-1}\sum_{i=1}^{n_1} y_i = \frac{n_1}{n}\bar{y}_1$ since we will only be counting instances in which $x_i = 1$. Using the result from part i, we show that $n^{-1}\sum_{i=1}^{n} x_i y_i = \bar{x}\bar{y}_1$
Therefore $n^{-1}\sum_{i=1}^{n} x_i y_i - \bar{x}\bar{y} = \bar{x}\bar{y}_1 - \bar{x}\bar{y} = \bar{x}\bar{y}_1 - \bar{x}[(1-\bar{x})\bar{y}_0 + \bar{x}\bar{y}_1]$, using the result from part iii.
With some manipulation, we can thus show that $n^{-1}\sum_{i=1}^{n} x_i y_i - \bar{x}\bar{y} = \bar{x}(1-\bar{x})(\bar{y}_1 - \bar{y}_0)$

(vi) Equation 2.74 states that $\hat{\beta}_1 = \bar{y}_1 - \bar{y}_0$. This is derived in the same way as the usual OLS estimator: minimizing the sum of the squared residuals to get $\hat{\beta}_1 = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2}$
Expanding the numerator and denominator: $\hat{\beta}_1 = \frac{\sum x_i y_i - n\bar{x}\bar{y}}{\sum x_i^2 - n\bar{x}^2}$
Using the results from parts v and vi: $\hat{\beta}_1 = \frac{\bar{x}(1-\bar{x})(\bar{y}_1 - \bar{y}_0)}{\bar{x}(1-\bar{x})} = \bar{y}_1 - \bar{y}_0$

(vii) We know that $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x}$. Using the results from parts iii and vi, we have:
$\hat{\beta}_0 = (1-\bar{x})\bar{y}_0 + \bar{x}\bar{y}_1 - (\bar{y}_1 - \bar{y}_0)\bar{x} = \bar{y}_0 - \bar{x}\bar{y}_0 + \bar{x}\bar{y}_1 - \bar{x}\bar{y}_1 + \bar{x}\bar{y}_0 = \bar{y}_0$

2.15 (i) The population average treatment effect $\tau_{ate} = E[y(1) - y(0)]$. If we are able to observe $y_i(0)$ and $y_i(1)$ for each observation i (e.g. we can observe the same observation in both states of the world), then $E\left[\frac{1}{n}\sum y_i(1)\right] = \frac{1}{n}\sum E[y_i(1)] = \frac{1}{n} * nE[y(1)] = E[y(1)]$. Similarly, $\frac{1}{n}\sum y_i(0)$ is an unbiased estimator for $E[y(0)]$. Therefore $E\left\{\frac{1}{n}\sum[y_i(1) - y_i(0)]\right\} = E[y(1) - y(0)]$

(ii) We can write $y_i = (1-x_i)y_i(0) + x_i y_i(1)$. Then, $\bar{y}_0 = \frac{1}{n}\sum_{i=1}^{n}[(1-x_i)y_i(0) + x_i y_i(1)] = \frac{1}{n_0}\sum_{i=1}^{n_0} y_i(0)$ since $x_i = 0$ for all observations used to calculate $\bar{y}_0$. By a similar logic, $\bar{y}_1 = \frac{1}{n_1}\sum_{i=1}^{n_1} y_i(1)$ since $x_1 = 1$ for all observations used to calculate $\bar{y}_1$. Thus, both $\bar{y}_0$ and $\bar{y}_1$ are computed using a subset of the sample. By contrast $\bar{y}(0)$ and $\bar{y}(1)$ would be computed using the entire sample, using the outcomes of each observation in both states of the world.

2.17 (i) $Var(u_i|x_i) = Var[(1-x_i)u_i(0) + x_i u_i(1)] = (1-x_i)^2\sigma_0^2 + x_i^2\sigma_1^2$. Recall that we are conditioning on x_i, so we can treat it as a deterministic variable. Finally, note that since x_i is binary, we can write $Var(u_i|x_i) = (1-x_i)\sigma_0^2 + x_i\sigma_1^2$. The conditional variance is a weighted average of the variances from the two different states of the world.

(ii) There are three ways in which $Var(u_i|x_i)$ could be constant:
1. $x_i = 0$ for all i
2. $x_i = 1$ for all i
3. $\sigma_0^2 = \sigma_1^2$

The first two scenarios imply that we have no observations in either the treatment or control groups and thus the experiment is not of any use. The last case suggests that outcomes for non-treated state have exactly the same variability as outcomes for the treated state. While this is certainly a possibility, it would imply that the only effect of the treatment is to change the mean outcome, not any other aspects of its distribution.

SOLUTIONS TO COMPUTER EXERCISES

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

(ii) The estimated equation is

$$\widehat{prate} = 83.08 + 5.86\, mrate$$

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

(iii) The intercept implies that, even if *mrate* = 0, the predicted participation rate is 83.08 percent. The coefficient on *mrate* implies that a one-dollar increase in the match rate – a fairly large increase – is estimated to increase *prate* by 5.86 percentage points. This assumes, of course, that this change *prate* is possible (if, say, *prate* is already at 98, this interpretation makes no sense).

(iv) If we plug *mrate* = 3.5 into the equation, we get $\widehat{prate} = 83.08 + 5.86(3.5) = 103.59$. This is impossible, as we can have at most a 100 percent participation rate. This illustrates that, especially when dependent variables are bounded, a simple regression model can give strange predictions for extreme values of the independent variable. (In the sample of 1,534 firms, only 34 have *mrate* $\geq$ 3.5.)

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

C2.3 (i) The estimated equation is

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

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

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

(ii) If someone works two more hours per week, then $\Delta totwrk = 120$ (because *totwrk* is measured in minutes), and so $\Delta sleep = -.151(120) = -18.12$ minutes. This is only a few minutes a night. If someone were to work one more hour on each of five working days, $\Delta sleep = -.151(300) = -45.3$ 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 β_1 is the elasticity of *rd* with respect to *sales*.

(ii) The estimated equation is

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

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

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

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

$$n = 4{,}268, \quad R^2 = .0138.$$

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

(v) Because the smallest *mailsyear* in the sample is .25, the smallest predicted value of *gifts* is $2.01 + 2.65(.25) \approx 2.67$. Even if we look at the overall population, where some people have

received no mailings, the smallest predicted value is about two. So, with this estimated equation, we never predict zero charitable gifts.

C2.9 (i) In 1996, 1,051 counties had zero murders. Out of 2,197 counties, 31 counties had at least one execution and the largest number of executions is 3.

(ii) The estimated equation is

$$murders = 5.46 + 58.56\, execs$$
$$n = 2197, R^2 = 0.0439.$$

(iii) The slope coefficient on *execs* implies that if the number of executions increases by one, the estimated number of murders increases largely by about 59. No, the estimated equation does not suggest a deterrent effect of capital punishment.

(iv) The smallest number of murders can be predicted by the equation is 5.46, that is about 5 murders. The residual for a county with zero executions and zero murders is -5.46.

(v) This simple linear regression equation predicts that if the number of executions increases by one, the estimated number of murders increases largely by about 59, which means capital punishment does not have a deterrent effect on murders — capital punishment is not discouraging people from doing murders. The sign and magnitude of the estimate +58.56 makes us suspect that the error term *u* and the independent variable *execs* are correlated. Therefore, the regression model is not well suited for prediction.

C2.11 (i) There are 141 students in the sample. The average college GPA is a 3.057, while the maximum GPA is a 4.0 (1 student with this GPA).

(ii) 56 students (39.7% of the sample) owned their own PC.

(iii) $\widehat{colGPA} = \underset{(0.040)}{2.989} + \underset{(0.063)}{0.170} * PC$. We estimates that students who do not own a PC have an average college GPA of 2.989. Those who own a PC have college GPA's that are on average 0.17 points higher than those who do not. Both coefficient estimates are statistically significant at the 1% level. In terms of magnitude, a 0.17 point increase in GPA represents about a 5.7% increase from the mean college GPA for those without a computer.

(iv) The $R^2 = 0.05$. We are only explaining about 5% of the variation in college GPAs with computer ownership. Thus, there are a lot of other factors that influence college GPA.

(v) If we truly believed that PC ownership was randomly assigned, then we could infer a causal effect on college GPA. However, it is highly unlikely that we have random assignment in this experiment. There are a wide range of factors that could both influence college GPA and PC ownership that we are overlooking. For example, consider parent's income. Students from wealthier families are more likely to own PCs. These same students are also less likely to have to work during college, freeing up more time for their studies and thus earning higher GPAs. As such, we cannot disentangle the effects of PC ownership from parental income. There are any number of other omitted variables that could cause us to violate the random assignment assumption.