Solution Assignment 2: EE425 1/2018

SOLUTIONS TO PROBLEMS

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.

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

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

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

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

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

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

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

2.7 (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.

2.8 (i) From equation (2.66),

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

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

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

After standard algebra, the numerator can be written as

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

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

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

Conditional on the x_i, we have

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

because E(u_i) = 0 for all i. Therefore, the bias in $\tilde{\beta}_1$ is given by the first term in this equation. This bias is obviously zero when $\beta_0 = 0$. It is also zero when $\sum_{i=1}^{n} x_i = 0$, which is the same as

$\bar{x}$ = 0. In the latter case, regression through the origin is identical to regression with an intercept.

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

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

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

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

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

2.9 (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:

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

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

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

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

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

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

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

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

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

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

which is what we wanted to show.

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

SOLUTIONS TO COMPUTER EXERCISES

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

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

(iii) The estimated equation is

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

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

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

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

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

$$wage = 116.99 + 8.30\, IQ$$

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

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

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

$$\log(wage) = 5.89 + .0088\, IQ$$

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

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

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

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

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

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

(iii) The regression results are

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

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

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

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

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

(iii) The estimated equation is

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

(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) ≈ 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.

$$n = 7430, R^2 = 0.5047.$$

(ii) No. The intercept does not have a meaningful interpretation because the reading test score is not close to zero for the population of interest.

(iii) No. R^2 = 0.5047 implies that about 50.47% variation in math score is explained by reading score.

(iv) If we run the regression of *read12* on *math12*, we obtained almost similar results as the regression of *math12* on *read12*. So it is better to hire more reading and math tutors.