Introductory Econometrics

2005-10-10

Exercise 1 solutions

CHAPTER 2 SOLUTIONS

SOLUTIONS TO THE PROBLEM EXERCISES

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.3 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 (2.9), 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 $ACT = 20$, $\hat{GPA} = .5681 + .1022(20) \approx 2.61$.

(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.

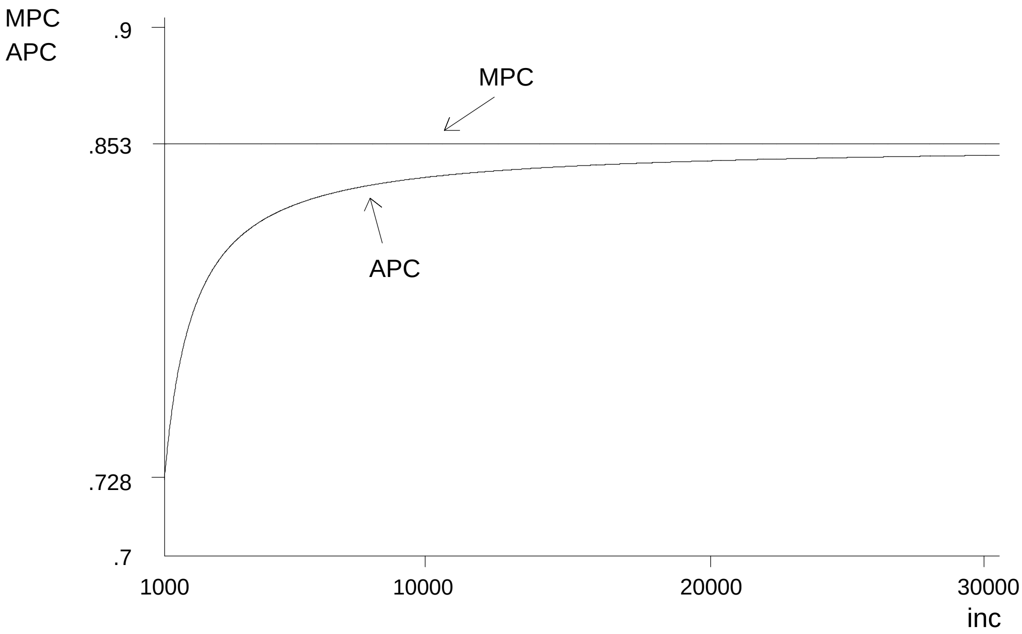
(iii) If we want a predicted *bwght* of 125, then *cigs* = (125 – 119.77)/(–.524) $\approx$ –10.18, or about –10 cigarettes! This is nonsense, of course, and it shows what happens when we are trying to predict something as complicated as birth weight with only a single explanatory variable. The largest 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%.

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 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).


MPC
APC
.9
.853
.728
.7
MPC
APC
1000
10000
20000
30000
inc

SOLUTIONS TO COMPUTER EXERCISES

2.10 (i) The average *prate* is about 87.36 and the average *mrate* is about .732.

(ii) The estimated equation is

$$\widehat{prate} = 83.05 + 5.86\, mrate$$

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

(iii) The intercept implies that, even if *mrate* = 0, the predicted participation rate is 83.05 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 $\widehat{prate}$ = 83.05 + 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 *mrate* ≥ 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.

2.11 (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

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

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

We obtain the approximate percentage change in *salary* given Δ*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%.

2.12 (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 $\Delta \widehat{sleep}$ = –.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.

2.13 (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:

$$\widehat{wage} = 116.99 + 8.30\, IQ$$

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

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:

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

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

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

2.14 (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

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

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

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

CHAPTER 3 SOLUTIONS

SOLUTIONS TO PROBLEMS

3.1 (i) *hsperc* is defined so that the smaller it is, the lower the student's standing in high school. Everything else equal, the worse the student's standing in high school, the lower is his/her expected college GPA.

(ii) Just plug these values into the equation:

$$\widehat{colgpa} = 1.392 - .0135(20) + .00148(1050) = 2.676.$$

(iii) The difference between A and B is simply 140 times the coefficient on *sat*, because *hsperc* is the same for both students. So A is predicted to have a score .00148(140) $\approx$ .207 higher.

(iv) With *hsperc* fixed, $\Delta\widehat{colgpa}$ = .00148Δ*sat*. Now, we want to find Δ*sat* such that $\Delta\widehat{colgpa}$ = .5, so .5 = .00148(Δ*sat*) or Δ*sat* = .5/(.00148) $\approx$ 338. Perhaps not surprisingly, a large ceteris paribus difference in SAT score – almost two and one-half standard deviations – is needed to obtain a predicted difference in college GPA or a half a point.

3.2 (i) Yes. Because of budget constraints, it makes sense that, the more siblings there are in a family, the less education any one child in the family has. To find the increase in the number of siblings that reduces predicted education by one year, we solve 1 = .094(Δ*sibs*), so Δ*sibs* = 1/.094 $\approx$ 10.6.

(ii) Holding *sibs* and *feduc* fixed, one more year of mother's education implies .131 years more of predicted education. So if a mother has four more years of education, her son is predicted to have about a half a year (.524) more years of education.

(iii) Since the number of siblings is the same, but *meduc* and *feduc* are both different, the coefficients on *meduc* and *feduc* both need to be accounted for. The predicted difference in education between B and A is .131(4) + .210(4) = 1.364.

3.4 (i) A larger rank for a law school means that the school has less prestige; this lowers starting salaries. For example, a rank of 100 means there are 99 schools thought to be better.

(ii) $\beta_1 > 0$, $\beta_2 > 0$. Both *LSAT* and *GPA* are measures of the quality of the entering class. No matter where better students attend law school, we expect them to earn more, on average. β_3, $\beta_4 > 0$. The number of volumes in the law library and the tuition cost are both measures of the school quality. (Cost is less obvious than library volumes, but should reflect quality of the faculty, physical plant, and so on.)

(iii) This is just the coefficient on *GPA*, multiplied by 100: 24.8%.

(iv) This is an elasticity: a one percent increase in library volumes implies a .095% increase in predicted median starting salary, other things equal.

(v) It is definitely better to attend a law school with a lower rank. If law school A has a ranking 20 less than law school B, the predicted difference in starting salary is 100(.0033)(20) = 6.6% higher for law school A.

3.7 Only (ii), omitting an important variable, can cause bias, and this is true only when the omitted variable is correlated with the included explanatory variables. The homoskedasticity assumption. MLR.5, played no role in showing that the OLS estimators are unbiased. (Homoskedasticity was used to obtain the standard variance formulas for the $\hat{\beta}_j$.) Further, the degree of collinearity between the explanatory variables in the sample, even if it is reflected in a correlation as high as .95, does not affect the Gauss-Markov assumptions. Only if there is a *perfect* linear relationship among two or more explanatory variables is MLR.4 violated.

SOLUTIONS TO COMPUTER EXERCISES

3.13 (i) Probably $\beta_2 > 0$, as more income typically means better nutrition for the mother and better prenatal care.

(ii) On the one hand, an increase in income generally increases the consumption of a good, and *cigs* and *faminc* could be positively correlated. On the other, family incomes are also higher for families with more education, and more education and cigarette smoking tend to be negatively correlated. The sample correlation between *cigs* and *faminc* is about −.173, indicating a negative correlation.

(iii) The regressions without and with *faminc* are

$$\widehat{bwght} = 119.77 - .514\, cigs$$

$$n = 1{,}388, \; R^2 = .023$$

and

$$\widehat{bwght} = 116.97 - .463\, cigs + .093\; faminc$$

$$n = 1{,}388, \; R^2 = .030.$$

The effect of cigarette smoking is slightly smaller when *faminc* is added to the regression, but the difference is not great. This is due to the fact that *cigs* and *faminc* are not very correlated, and the coefficient on *faminc* is practically small. (The variable *faminc* is measured in thousands, so $10,000 more in 1988 income increases predicted birth weight by only .93 ounces.)

3.14 (i) The estimated equation is

$$\widehat{price} = -19.32 + .128\, sqrft + 15.20\, bdrms$$

$$n = 88, \; R^2 = .632$$

(ii) Holding square footage constant, $\Delta\widehat{price} = 15.20\; \Delta bdrms$, and so $\widehat{price}$ increases by 15.20, which means $15,200.

(iii) Now $\Delta\widehat{price} = .128\,\Delta sqrft + 15.20\,\Delta bdrms = .128(140) + 15.20 = 33.12$, or $33,120. Because the size of the house is increasing, this is a much larger effect than in (ii).

(iv) About 63.2%.

(v) The predicted price is –19.32 + .128(2,438) + 15.20(4) = 353.544, or $353,544.

(vi) From part (v), the estimated value of the home based only on square footage and number of bedrooms is $353,544. The actual selling price was $300,000, which suggests the buyer underpaid by some margin. But, of course, there are many other features of a house (some that we cannot even measure) that affect price, and we have not controlled for these.

3.15 (i) The constant elasticity equation is

$$\widehat{\log(salary)} = 4.62 + .162\,\log(sales) + .107\,\log(mktval)$$

$$n = 177,\;\; R^2 = .299.$$

(ii) We cannot include profits in logarithmic form because profits are negative for nine of the companies in the sample. When we add it in levels form we get

$$\widehat{\log(salary)} = 4.69 + .161\,\log(sales) + .098\,\log(mktval) + .000036\,profits$$

$$n = 177, R^2 = .299.$$

The coefficient on *profits* is very small. Here, *profits* are measured in millions, so if profits increase by $1 billion, which means $\Delta profits$ = 1,000 – a huge change – predicted salary increases by about only 3.6%. However, remember that we are holding sales and market value fixed.

Together, these variables (and we could drop *profits* without losing anything) explain almost 30% of the sample variation in log(*salary*). This is certainly not "most" of the variation.

(iii) Adding *ceoten* to the equation gives

$$\widehat{\log(salary)} = 4.56 + .162\,\log(sales) + .102\,\log(mktval) + .000029\,profits + .012 ceoten$$

$$n = 177,\;\; R^2 = .318.$$

This means that one more year as *CEO* increases predicted salary by about 1.2%.

(iv) The sample correlation between log(*mktval*) and *profits* is about .78, which is fairly high. As we know, this causes no bias in the OLS estimators, although it can cause their variances to be large. Given the fairly substantial correlation between market value and firm profits, it is not too surprising that the latter adds nothing to explaining CEO salaries. Also, *profits* is a short term measure of how the firm is doing while *mktval* is based on past, current, and expected future profitability.

3.16 (i) The minimum, maximum, and average values for these three variables are given in the table below:

Variable	Average	Minimum	Maximum
atndrte	81.71	6.25	100
priGPA	2.59	.86	3.93
ACT	22.51	13	32

(ii) The estimated equation is

$$\widehat{atndrte} = 75.70 + 17.26\ priGPA - 1.72\ ACT$$

$$n = 680, \quad R^2 = .291.$$

The intercept means that, for a student whose prior GPA is zero and ACT score is zero, the predicted attendance rate is 75.7%. But this is clearly not an interesting segment of the population. (In fact, there are no students in the college population with *priGPA* = 0 and *ACT* = 0.)

(iii) The coefficient on *priGPA* means that, if a student's prior GPA is one point higher (say, from 2.0 to 3.0), the attendance rate is about 17.3 percentage points higher. This holds *ACT* fixed. The negative coefficient on *ACT* is, perhaps initially a bit surprising. Five more points on the *ACT* is predicted to lower attendance by 8.6 percentage points at a given level of *priGPA*. As *priGPA* measures performance in college (and, at least partially, could reflect, past attendance rates), while *ACT* is a measure of potential in college, it appears that students that had more promise (which could mean more innate ability) think they can get by with missing lectures.

(iv) We have $\widehat{atndrte}$ = 75.70 + 17.267(3.65) – 1.72(20) $\approx$ 104.3. Of course, a student cannot have higher than a 100% attendance rate. Getting predications like this is always possible when using regression methods with natural upper or lower bounds on the dependent variable. In practice, we would predict a 100% attendance rate for this student. (In fact, this student had an attendance rate of only 87.5%.)

(v) The difference in predicted attendance rates for A and B is 17.26(3.1 – 2.1) – (21 – 26) = 25.86.