Answers to Small Project 2

1.

(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

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

$$n = 680, \quad R^2 = 0.291.$$

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, or with values even close to zero.)

(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* are 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 who had more promise (which could mean more innate ability) think they can get by 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 predictions like this is always possible when using regression methods for dependent variables with natural upper or lower bounds. In practice, we would predict a 100% attendance rate for this student. (In fact, this student had an actual attendance rate of 87.5%.)

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

2.

(i) The results of the regression are

$$\widehat{math10} = -20.36 + 6.23\log(expend) - 0.305\ lnchprg$$

$$n = 408, \quad R^2 = 0.180.$$

The signs of the estimated slopes imply that more spending increases the pass rate (holding *lnchprg* fixed) and a higher poverty rate (proxied well by *lnchprg*) decreases the pass rate (holding spending fixed). These are what we expect.

(ii) As usual, the estimated intercept is the predicted value of the dependent variable when all regressors are set to zero. Setting *lnchprg* = 0 makes sense, as there are schools with low poverty rates. Setting log(*expend*) = 0 does not make sense, because it is the same as setting *expend* = 1, and spending is measured in dollars per student. Presumably this is well outside any sensible range. Not surprisingly, the prediction of a -20 pass rate is nonsensical.

(iii) The simple regression results are

$$\widehat{math10} = -69.34 + 11.16\log(expend)$$

$$n = 408, \quad R^2 = 0.030,$$

and the estimated spending effect is larger than it was in part (i) – almost double.

(iv) The sample correlation between *lexpend* and *lnchprg* is about -0.19, which means that, on average, high schools with poorer students spent less per student. This makes sense, especially in 1993 in Michigan, where school funding was essentially determined by local property tax collections.

(v) We can use equation (3.23) in the textbook (p.75). Because Corr(x_1,x_2) < 0, which means $\tilde{\delta}_1 < 0$, and $\hat{\beta}_2 < 0$, the simple regression estimate, $\tilde{\beta}_1$, is larger than the multiple regression estimate, $\hat{\beta}_1$. Intuitively, failing to account for the poverty rate leads to an overestimate of the effect of spending.