

## Solutions: Problem Set #6

(1) The coefficient estimates for the regressions in this problem set are provided in the other file. These are the same as reported in the textbook.

(1a) In class, we derived the omitted variables bias formula:

$$\beta_2 \rightarrow \theta_2 + \theta_3 \frac{\text{Cov}(x, z)}{\text{Var}(x)}.$$

For this particular question, the omitted variable,  $z$  is the percentage of English learners (PctEl), and  $x$  refers to Str, or student-teacher ratio. Finally,  $\beta_2$  is the coefficient on Str from the simple regression that fails to include PctEl, and  $\theta_2$  is the coefficient on Str from the multiple regression model that includes Str.

So, we re-write the above as, approximately,

$$\theta_2 \approx \beta_2 - \theta_3 \frac{\text{Cov}(Str, PctEL)}{\text{Var}(Str)}.$$

In general, we think that in schools with a high percentage of English learners, test scores might be lower on average. This implies that  $\theta_3 < 0$ . In addition we might expect that  $\text{Cov}(Str, PctEl) > 0$ . That is, in schools with a high percentage of English learners, student-teacher ratios are also large. (See the discussion in section 5.1 of your book that lends empirical support to both of these claims). Putting these two pieces together, we find that

$$\theta_2 \approx \beta_2 + c,$$

for some constant  $c > 0$ . Since  $\beta_2$  was negative (-2.28), this derivation shows that  $\theta_2$ , the coefficient on Str from the multiple regression, will be less negative, or potentially, even positive. The results of our multiple regression analysis confirm this - when adding PctEl to the regression model, the impact of Str falls (i.e., moves closer to 0) by more than one-half.

(2a and b) The results are generally consistent with what we expect. Larger homes (higher square footage) have higher sales prices. Specifically, holding bdrms constant, a 100 square foot increase in the size of the home increases our dependent variable,

price, by 12.8. Since the dependent variable is measured in thousands of dollars, this translates into almost a \$13,000 increase in the sales price of the home. Similarly, *bdrms* has a positive effect on sales price - every added bedroom (holding square footage constant) increases the sales price by about \$15,200. Note, however, that *bdrms* is not statistically significant at either the 5 or 10 percent levels.

(2c) Adding both a bedroom and 150 square feet produces a change in expected price equal to

$$150(.128) + 15.20 = 34.4.$$

So, an additional bedroom which adds 150 square feet would increase the expected sales price of the home by \$34,400.

(2d). Note that the t-statistic for testing the null that the coefficient on *bdrms* equals zero is  $15.19/9.48 = 1.60$ . Following similar logic to the solution for the last problem set, we need to calculate

$$\Pr(z > 1.60) - 1 - \Pr(z \leq 1.60) = 1 - .9452 = .0548$$

and then multiply this by 2 to get the *p*-value. This is  $.0548(2) = .1096$ . Note that this is close to the *p*-value given by STATA, which is .113. What accounts for this small difference is that the correct distribution associated with our test statistic is a  $t_{88-3} = t_{85}$ . We do not have a table for the t-cdf, but instead use the Normal as an approximation.

(2e) The predicted sales price is

$$E(\text{Price} | \widehat{Sq\ ft} = 2438, \text{Bdrms} = 4) = -19.315 + .128(2438) + 15.19(4) = 353.5.$$

Thus, the predicted sales price for this house is \$ 353,500.

(2f) The estimated residual is

$$\hat{u}_1 = 300 - 353.5 = -53.5.$$

One might be tempted to say that the person underpaid for the house, but there are other factors not captured by the regression which may be at play. For example, the

house may be in a bad neighborhood with poor schools, or have a small lot size. These factors may help to explain the disparity between predicted and observed prices.

**((3))** This is not a correct interpretation of the confidence interval. The discussion seems to attribute randomness to the true population parameter by stating that it falls within  $[-1.95, -.026]$  95 percent of the time. In fact, the interval endpoints themselves are random while the parameter is fixed - it either falls in this interval or it does not. In repeated sampling (as discussed on the powerpoint slides with a generated data experiment), the confidence intervals generated will contain the true parameter 95 percent of the time. This is just one such realization of a confidence interval - it either contains the parameter, or it does not.