Solutions: Problem Set #3

(4.1) Throughout this exercise, let $TS$ denote the test score and $CS$ denote class size.

(a) We seek $E(TS|CS = 22)$:

$$E(TS|CS = 22) = \hat{\beta}_1 + 22\hat{\beta}_2$$

$$= 520.4 - 22(5.82)$$

$$= 392.36$$

(b) Note that (in terms of the population regression function):

$$E(TS|CS = 23) = \beta_1 + (23)\beta_2$$

and

$$E(TS|CS = 19) = \beta_1 + (19)\beta_2.$$ 

Therefore, the difference, denoted $\Delta$, from going from a class size of 19 to a class size of 23 is

$$\Delta \equiv E(TS|CS = 19) - E(TS|CS = 23) = \beta_1 + (19)\beta_2 - [\beta_1 + (23)\beta_2] = -4\beta_2.$$ 

Replacing population parameters with their estimated values, we find

$$\hat{\Delta} = -4(-5.82) = 23.28.$$ 

So, test scores are predicted to be about 23.23 points larger when the class size is at its smaller value of 19. (Note: you could have defined $\Delta$ as the difference between a class size of 23 - a class size of 19, in which case your point estimate would be -23.28).

(e) We are given $CS = 21$.

Our OLS formula for the intercept implies:

$$\hat{\beta}_1 = \overline{y} - \hat{\beta}_2\overline{x},$$

or in this particular case:

$$\hat{\beta}_1 = \overline{TS} - \hat{\beta}_2\overline{CS} \Rightarrow \overline{TS} = \hat{\beta}_1 + \hat{\beta}_2\overline{CS}.$$
Based on what is given in this problem, we can evaluate the right hand side of this final expression to obtain:

\[ T \bar{S} = 520.4 - 5.82(21.4) = 395.85. \]

\((4.2)\) (a). Throughout this exercise, let \(M\) denote the Male dummy variable. (That is, \(M = 1\), for example, denotes that we are talking about (or conditioning on) Males, while \(M = 0\) denotes females).

Note that, based on the regression model:

\[ E(Wage|M = 1) = \beta_1 + \beta_2 \quad \text{and} \quad E(Wage|M = 0) = \beta_1. \]

Therefore, letting \(\Delta\) denote the wage gap (the difference in wages between men and women), we have

\[ \Delta = E(Wage|M = 1) - E(Wage|M = 0) = \beta_2. \]

Our estimate of this wage gap is then

\[ \hat{\Delta} = \hat{\beta_2} = 2.79. \]

In other words, based on the results of this regression analysis, men earn about 2.79 more dollars per hour than women.

(b) Our intuition is that, since \(\beta_1\) denotes the expected wage of women, and \(\beta_1 + \beta_2\) denotes the expected wage of men, then \(\hat{\beta_1}\), our estimator, will reflect the average of women’s wages in the sample, while \(\hat{\beta}_1 + \hat{\beta}_2\) will denote the average of men’s wages in the sample. If this is true, then the mean wage of women in the sample is \(\hat{\beta}_1 = 12.68\) and the mean wages of men is \(\hat{\beta}_1 + \hat{\beta}_2 = 12.68 + 2.79 = 15.47.\)

One can also be a bit more formal about this and prove this result. To do this, let us first go back to our general first order condition for the slope parameter (\(\beta_2\)) of the linear regression model:

\[ \sum_{i=1}^{n}(y_i - \hat{\beta}_1 - \hat{\beta}_2 x_i)x_i = 0. \]
In this case $x_i$ is our Male dummy variable. It is clear that all terms in the sum for which $x_i = 0$ (i.e., females) will contribute nothing to this summation, since everything inside the summation is multiplied by $x_i$. So, it reduces to:

$$
\sum_{i=1}^{n} (y_i - \hat{\beta}_1 - \hat{\beta}_2 x_i) x_i = \sum_{i: \text{Male}=1} (y_i - \hat{\beta}_1 - \hat{\beta}_2)
$$

$$
= \sum_{i: \text{Male}=1} y_i - \sum_{i: \text{Male}=1} (\hat{\beta}_1 + \hat{\beta}_2)
$$

$$
= \sum_{i: \text{Male}=1} y_i - n_m (\hat{\beta}_1 + \hat{\beta}_2)
$$

In the above, we noted that our sum is only for males, and let $n_m$ denote the number of males in the sample (which is this case is 250). Since this expression must be equal to zero, it follows that

$$(\hat{\beta}_1 + \hat{\beta}_2) = \frac{1}{n_m} \sum_{i: \text{Male}=1} y_i.$$}

The right-hand side is simply the average wage of men in the sample, so that $\hat{\beta}_1 + \hat{\beta}_2$ does indeed reduce sample average of men’s wages.

To show that $\hat{\beta}_1$ is the sample average of women’s wages, we must go back to our first order condition for the intercept in the linear regression model:

$$
\sum_{i=1}^{n} (y_i - \hat{\beta}_1 + \hat{\beta}_2 x_i) = 0.
$$

We can break this sum into two parts: one part that arises for the subsample of women and one part that arises from the subsample of men:

$$
\sum_{i=1}^{n} (y_i - \hat{\beta}_1 + \hat{\beta}_2 x_i) = \sum_{i: \text{Male}=1} (y_i - \hat{\beta}_1 + \hat{\beta}_2) + \sum_{i: \text{Male}=0} (y_i - \hat{\beta}_1)
$$

$$
= [ \sum_{i: \text{Male}=1} y_i - \sum_{i: \text{Male}=1} (\hat{\beta}_1 + \hat{\beta}_2) ] + \sum_{i: \text{Male}=0} (y_i - \hat{\beta}_1)
$$

$$
= [ \sum_{i: \text{Male}=1} y_i - n_m (\hat{\beta}_1 + \hat{\beta}_2) ] + \sum_{i: \text{Male}=0} (y_i - \hat{\beta}_1)
$$

$$
= [ \sum_{i: \text{Male}=1} y_i - \sum_{i: \text{Male}=1} y_i ] + \sum_{i: \text{Male}=0} (y_i - \hat{\beta}_1)
$$

$$
= \sum_{i: \text{Male}=0} (y_i - \hat{\beta}_1)
$$

The only tricky part of this derivation is the second to last line, where we have subbed in our pervious result which showed that $\hat{\beta}_1 + \hat{\beta}_2$ picks up the sample average of men’s
wages. As a result, the first part of our expression drops out. Since the first order condition requires that this expression equals zero, it follows that

$$\sum_{i: Male=0} (y_i - \hat{\beta}_1) = 0$$

whence

$$\hat{\beta}_1 = (1/n_W) \sum_{i: Male=0} y_i,$$

(with $n_W$ denoting the number of women in the sample). This derivation shows that $\hat{\beta}_1$ is the average of women’s wages from the sample.

(e) In our original specification, we had

$$Wage = \beta_1 + \beta_2 M + u.$$

In our new specification, we write

$$Wage = \theta_1 + \theta_2 F + u,$$

where $F$ is a dummy variable taking on the value 1 if the observation corresponds to a female, and 0 otherwise. Clearly, $F$ and $M$ are connected by the simple relation: $F = 1 - M$. Subbing this into our second model, we obtain

$$Wage = \theta_1 + \theta_2 (1 - M) + u$$

$$= (\theta_1 + \theta_2) - \theta_2 M + u$$

This regression equation is of the same form as our original representation of the model, where we regressed wages on a constant and a male dummy. Therefore, it follows that the coefficients on the $M$ dummy variable must be the same, i.e., $\theta_2 = -\beta_2$. In addition, the intercept coefficients must be the same, which imposes that $\theta_1 + \theta_2 = \beta_1$, or in other words, $\theta_1 = \beta_1 - \theta_2 = \beta_1 + \beta_2$ Putting these together in terms of the coefficient estimates themselves, we have $\hat{\theta}_1 = 12.68 + 2.79 = 15.47$ and $\hat{\theta}_2 = -2.79$. (Note that the estimated conditional expectations are now the same regardless of the representation of the model). Finally, the $R^2$ value will not change - the model is the same, the predicted values are unchanged and the $R^2$ remains constant at .06.
We are asked to show that the intercept estimator is unbiased, given that the slope estimator is unbiased. Our intercept estimator is defined as

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}.$$ 

Since

$$\bar{y} = \beta_0 + \beta_1 \bar{x} + \bar{u},$$

it follows that

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = \beta_0 + \beta_1 \bar{x} + \bar{u} - \hat{\beta}_1 \bar{x} + \bar{u} = \beta_0 + (\beta_1 - \hat{\beta}_1) \bar{x} + \bar{u}$$

To establish unbiasedness, we must take the expectation of this expression:

$$E(\hat{\beta}_0) = E[\beta_0 + (\beta_1 - \hat{\beta}_1) \bar{x} + \bar{u}] = E(\beta_0) + E[(\beta_1 - \hat{\beta}_1) \bar{x}] + E(\bar{u}) = \beta_0 + E[(\beta_1 - \hat{\beta}_1) \bar{x}] + E(\bar{u}) = \beta_0 + E[(1/n) \sum_i u_i] = \beta_0 + (1/n) \sum_i E(u_i) = \beta_0$$

The third step follows since we regard \(x\) as a constant (a better way to do this is to apply the law of iterated expectations, but we do not take up that level of generality here). The fourth step applies the fact that \(\hat{\beta}_1\) is unbiased, while the fifth notes that \(E(u_i) = 0\) by the assumptions of our regression model.

(4.5) (a)

$$E(\text{Weight} \mid \text{Height} = 70) = -99.41 + 3.94(70) = 176.39$$
$$E(\text{Weight} \mid \text{Height} = 65) = -99.41 + 3.94(65) = 156.69$$
$$E(\text{Weight} \mid \text{Height} = 74) = -99.41 + 3.94(74) = 192.15$$
(b) By a similar logic to 4.1b and 4.2a, the increase will simply be estimated by $1.5(3.94) = 5.91$. In other words, growing 1.5 inches will increase your weight by an average of approximately 6 pounds.