Chapter 5
Regression with a Single Regressor: Hypothesis Tests and Confidence Intervals

■ Solutions to Exercises

1 (a) The 95% confidence interval for β_1 is $\{-5.82 \pm 1.96 \times 2.21\}$, that is $-10.152 \le \beta_1 \le -1.4884$.

(b) Calculate the *t*-statistic:

$$t^{act} = \frac{\hat{\beta}_1 - 0}{\text{SE}(\hat{\beta}_1)} = \frac{-5.82}{2.21} = -2.6335.$$

The *p*-value for the test $H_0 : \beta_1 = 0$ vs. $H_1 : \beta_1 \neq 0$ is

$$p\text{-value} = 2\Phi(-|t^{act}|) = 2\Phi(-2.6335) = 2 \times 0.0042 = 0.0084.$$

The *p*-value is less than 0.01, so we can reject the null hypothesis at the 5% significance level, and also at the 1% significance level.

(c) The *t*-statistic is

$$t^{act} = \frac{\hat{\beta}_1 - (-5.6)}{\text{SE}(\hat{\beta}_1)} = \frac{0.22}{2.21} = 0.10$$

The *p*-value for the test $H_0 : \beta_1 = -5.6$ vs. $H_1 : \beta_1 \neq -5.6$ is

$$p\text{-value} = 2\Phi(-|t^{act}|) = 2\Phi(-0.10) = 0.92$$

The *p*-value is larger than 0.10, so we cannot reject the null hypothesis at the 10%, 5% or 1% significance level. Because $\beta_1 = -5.6$ is not rejected at the 5% level, this value is contained in the 95% confidence interval.

(d) The 99% confidence interval for β_0 is $\{520.4 \pm 2.58 \times 20.4\}$, that is, $467.7 \le \beta_0 \le 573.0$.

2. (a) The estimated gender gap equals \$2.12/hour.

(b) The hypothesis testing for the gender gap is $H_0 : \beta_1 = 0$ vs. $H_1 : \beta_1 \neq 0$. With a *t*-statistic

$$t^{act} = \frac{\hat{\beta}_1 - 0}{SE(\hat{\beta}_1)} = \frac{2.12}{0.36} = 5.89,$$

the *p*-value for the test is

$$p\text{-value} = 2\Phi(-|t^{act}|) = 2\Phi(-5.89) = 2 \times 0.0000 = 0.000 \text{ (to four decimal places)}$$

The *p*-value is less than 0.01, so we can reject the null hypothesis that there is no gender gap at a 1% significance level.

(c) The 95% confidence interval for the gender gap β_1 is $\{2.12 \pm 1.96 \times 0.36\}$, that is, $1.41 \le \beta_1 \le 2.83$.

(d) The sample average wage of women is $\hat{\beta}_0 = \$12.52$/hour. The sample average wage of men is $\hat{\beta}_0 + \hat{\beta}_1 = \$12.52 + \$2.12 = \$14.64$/hour.

(e) The binary variable regression model relating wages to gender can be written as either

$$Wage = \beta_0 + \beta_1 Male + u_i,$$

or

$$Wage = \gamma_0 + \gamma_1 Female + v_i.$$

In the first regression equation, *Male* equals 1 for men and 0 for women; β_0 is the population mean of wages for women and $\beta_0 + \beta_1$ is the population mean of wages for men. In the second regression equation, *Female* equals 1 for women and 0 for men; γ_0 is the population mean of wages for men and $\gamma_0 + \gamma_1$ is the population mean of wages for women. We have the following relationship for the coefficients in the two regression equations:

$$\gamma_0 = \beta_0 + \beta_1,$$
$$\gamma_0 + \gamma_1 = \beta_0.$$

Given the coefficient estimates $\hat{\beta}_0$ and $\hat{\beta}_1$, we have

$$\hat{\gamma}_0 = \hat{\beta}_0 + \hat{\beta}_1 = 14.64,$$
$$\hat{\gamma}_1 = \hat{\beta}_0 - \hat{\gamma}_0 = -\hat{\beta}_1 = -2.12.$$

Due to the relationship among coefficient estimates, for each individual observation, the OLS residual is the same under the two regression equations: $\hat{u}_i = \hat{v}_i$. Thus the sum of squared residuals, $SSR = \sum_{i=1}^{n} \hat{u}_i^2$, is the same under the two regressions. This implies that both $SER = \left(\frac{SSR}{n-1}\right)^{\frac{1}{2}}$ and $R^2 = 1 - \frac{SSR}{TSS}$ are unchanged.

In summary, in regressing *Wages* on *Female*, we will get

$$\widehat{Wages} = 14.64 - 2.12 Female, \quad R^2 = 0.06, \quad \text{SER} = 4.2.$$

3. The 99% confidence interval is $1.5 \times \{3.94 \pm 2.58 \times 0.31)$ or 4.71 lbs ≤ WeightGain ≤ 7.11 lbs.

4. (a) $-3.13 + 1.47 \times 16 = \$20.39$ per hour

(b) The wage is expected to increase from \$14.51 to \$17.45 or by \$2.94 per hour.

(c) The increase in wages for college education is $\beta_1 \times 4$. Thus, the counselor's assertion is that $\beta_1 = 10/4 = 2.50$. The *t*-statistic for this null hypothesis is $t = \frac{1.47 - 2.50}{0.07} - 14.71$, which has a *p*-value of 0.00. Thus, the counselor's assertion can be rejected at the 1% significance level. A 95% confidence for $\beta_1 \times 4$ is $4 \times (1.47 \pm 1.97 \times 0.07)$ or \$5.33 ≤ Gain ≤ \$6.43.

5 (a) The estimated gain from being in a small class is 13.9 points. This is equal to approximately 1/5 of the standard deviation in test scores, a moderate increase.

(b) The *t*-statistic is $t^{act} = \frac{13.9}{2.5} = 5.56,$ which has a *p*-value of 0.00. Thus the null hypothesis is rejected at the 5% (and 1%) level.

(c) $13.9 \pm 2.58 \times 2.5 = 13.9 \pm 6.45.$

6. (a) The question asks whether the variability in test scores in large classes is the same as the variability in small classes. It is hard to say. On the one hand, teachers in small classes might able so spend more time bringing all of the students along, reducing the poor performance of particularly unprepared students. On the other hand, most of the variability in test scores might be beyond the control of the teacher.

(b) The formula in 5.3 is valid for heteroskesdasticity or homoskedasticity; thus inferences are valid in either case.

7. (a) The *t*-statistic is $\frac{3.2}{1.5} = 2.13$ with a *p*-value of 0.03; since the *p*-value is less than 0.05, the null hypothesis is rejected at the 5% level.

(b) $3.2 \pm 1.96 \times 1.5 = 3.2 \pm 2.94$

(c) Yes. If *Y* and *X* are independent, then $\beta_1 = 0$; but this null hypothesis was rejected at the 5% level in part (a).

(d) β_1 would be rejected at the 5% level in 5% of the samples; 95% of the confidence intervals would contain the value $\beta_1 = 0$.

8. (a) $43.2 \pm 2.05 \times 10.2$ or 43.2 ± 20.91, where 2.05 is the 5% two-sided critical value from the t_{28} distribution.

(b) The *t*-statistic is $t^{act} = \frac{61.5-55}{7.4} = 0.88,$ which is less (in absolute value) than the critical value of 20.5. Thus, the null hypothesis is not rejected at the 5% level.

(c) The one sided 5% critical value is 1.70; t^{act} is less than this critical value, so that the null hypothesis is not rejected at the 5% level.

9. (a) $\overline{\beta} = \frac{1}{\overline{X}}\frac{1}{n}(Y_1 + Y_2 + \cdots + Y_n)$ so that it is linear function of Y_1, Y_2, …, Y_n.

(b) $E(Y_i|X_1, \ldots, X_n) = \beta_1 X_i$, thus

$$E(\overline{\beta}|X_1, \ldots, X_n) = E\frac{1}{\overline{X}}\frac{1}{n}(Y_1 + Y_2 + \cdots + Y_n)|X_1, \ldots, X_n)$$
$$= \frac{1}{\overline{X}}\frac{1}{n}\beta_1(X_1 + \cdots + X_n) = \beta_1$$

10. Let n_0 denote the number of observation with $X = 0$ and n_1 denote the number of observations with $X = 1$; note that $\sum_{i=1}^{n} X_i = n_1;$ $\overline{X} = n_1|n;$ $\frac{1}{n_1}\sum_{i=1}^{n} X_i Y_i = \overline{Y}_1;$
$\sum_{i=1}^{n}(X_i - \overline{X})^2 = \sum_{i=1}^{n} X_i^2 - n\overline{X}^2 = n_1 - \frac{n_1^2}{n} = n_1\left(1 - \frac{n_1}{n}\right) = \frac{n_1 n_0}{n};$ $n_1\overline{Y}_1 + n_0\overline{Y}_0 = \sum_{i=1}^{n} Y_i,$ so that
$\overline{Y} = \frac{n_1}{n}\overline{Y}_1 + \frac{n_0}{n}\overline{Y}_0$

From the least squares formula

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(X_i - \bar{X})^2} = \frac{\sum_{i=1}^{n} X_i(Y_i - \bar{Y})}{\sum_{i=1}^{n}(X_i - \bar{X})^2} = \frac{\sum_{i=1}^{n} X_i Y_i - \bar{Y}n_1}{n_1 n_0 | n}$$

$$= \frac{n}{n_0}(\bar{Y}_1 - \bar{Y}) = \frac{n}{n_0}\left(\bar{Y} - \frac{n_1}{n}\bar{Y}_1 - \frac{n_0}{n}\bar{Y}_0\right) = \bar{Y}_1 - \bar{Y}_0$$

and $\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X} = \left(\frac{n_0}{n}\bar{Y}_0 + \frac{n_1}{n}\bar{Y}_1\right) - (\bar{Y}_1 - \bar{Y}_0)\frac{n_1}{n} = \frac{n_1 + n_0}{n}\bar{Y}_0 = \bar{Y}_0$

11. Using the results from 5.10, $\hat{\beta}_0 = \bar{Y}_m$ and $\hat{\beta}_1 = \bar{Y}_w - \bar{Y}_m$. From Chapter 3, $\text{SE}(\bar{Y}_m) = \frac{s_m}{\sqrt{n_m}}$ and $\text{SE}(\bar{Y}_w - \bar{Y}_m) = \sqrt{\frac{s_m^2}{n_m} + \frac{s_w^2}{n_w}}$. Plugging in the numbers $\hat{\beta}_0 = 523.1$ and $\text{SE}(\hat{\beta}_0) = 6.22$; $\hat{\beta}_1 = -38.0$ and $\text{SE}(\hat{\beta}_1) = 7.65$.

12. Equation (4.22) gives

$$\sigma^2_{\hat{\beta}_0} = \frac{\text{var}(H_i u_i)}{n\left[E\left(H_i^2\right)\right]^2}, \quad \text{where } H_i = 1 - \frac{\mu_x}{E\left(X_i^2\right)}X_i.$$

Using the facts that $E(u_i|X_i) = 0$ and $\text{var}(u_i|X_i) = \sigma_u^2$ (homoskedasticity), we have

$$E(H_i u_i) = E\left(u_i - \frac{\mu_x}{E\left(X_i^2\right)}X_i u_i\right) = E(u_i) - \frac{\mu_x}{E\left(X_i^2\right)}E[X_i E(u_i|X_i)]$$

$$= 0 - \frac{\mu_x}{E\left(X_i^2\right)} \times 0 = 0,$$

and

$$E[(H_i u_i)^2] = E\left\{\left(u_i - \frac{\mu_x}{E\left(X_i^2\right)}X_i u_i\right)^2\right\}$$

$$= E\left\{u_i^2 - 2\frac{\mu_x}{E\left(X_i^2\right)}X_i u_i^2 + \left[\frac{\mu_x}{E\left(X_i^2\right)}\right]^2 X_i^2 u_i^2\right\}$$

$$= E\left(u_i^2\right) - 2\frac{\mu_x}{E\left(X_i^2\right)}E\left[X_i E\left(u_i^2|X_i\right)\right] + \left[\frac{\mu_x}{E\left(X_i^2\right)}\right]^2 E\left[X_i^2 E\left(u_i^2|X_i\right)\right]$$

$$= \sigma_u^2 - 2\frac{\mu_x}{E\left(X_i^2\right)}\mu_x\sigma_u^2 + \left[\frac{\mu_x}{E\left(X_i^2\right)}\right]^2 E\left(X_i^2\right)\sigma_u^2$$

$$= \left(1 - \frac{\mu_x^2}{E\left(X_i^2\right)}\right)\sigma_u^2.$$

Because $E(H_i u_i) = 0$, $\text{var}(H_i u_i) = E[(H_i u_i)^2]$, so

$$\text{var}(H_i u_i) = E[(H_i u_i)^2] = \left(1 - \frac{\mu_x^2}{E\left(X_i^2\right)}\right)\sigma_u^2.$$

We can also get

$$E\left(H_i^2\right) = E\left\{\left(1 - \frac{\mu_x}{E\left(X_i^2\right)}X_i\right)^2\right\} = E\left\{1 - 2\frac{\mu_x}{E\left(X_i^2\right)}X_i + \left[\frac{\mu_x}{E\left(X_i^2\right)}\right]^2 X_i^2\right\}$$

$$= 1 - 2\frac{\mu_x^2}{E\left(X_i^2\right)} + \left[\frac{\mu_x}{E\left(X_i^2\right)}\right]^2 E\left(X_i^2\right) = 1 - \frac{\mu_x^2}{E\left(X_i^2\right)}.$$

Thus

$$\sigma_{\hat{\beta}_0}^2 = \frac{\text{var}(H_i u_i)}{\left[nE\left(H_i^2\right)^2\right]} = \frac{\left(1 - \frac{\mu_x^2}{E\left(X_i^2\right)}\right)\sigma_u^2}{n\left(1 - \frac{\mu_x^2}{E\left(X_i^2\right)}\right)^2} = \frac{\sigma_u^2}{n\left(1 - \frac{\mu_x^2}{E\left(X_i^2\right)}\right)}$$

$$= \frac{E(X_i^2)\sigma_u^2}{n[E\left(X_i^2 - \mu_x^2\right)]} = \frac{E\left(X_i^2\right)\sigma_u^2}{n\sigma_X^2}.$$

13. (a) Yes
    (b) Yes
    (c) They would be unchanged
    (d) (a) is unchanged; (b) is no longer true as the errors are not conditionally homosckesdastic.

14. (a) From Exercise (4.11), $\hat{\beta} = \sum a_i Y_i$ where $a_i = \frac{X_i}{\sum_{j=1}^n X_j^2}$. Since the weights depend only on X_i but not on Y_i, $\hat{\beta}$ is a linear function of *Y*.

    (b) $$E(\hat{\beta}|X_1, \ldots, X_n) = \beta + \frac{\sum_{i=1}^n X_i E(u_i|X_1, \ldots, X_n)}{\sum_{i=1}^n X_j^2} = \beta \text{ since } E(u_i|X_1, \ldots, X_n) = 0$$

    (c) $$Var(\hat{\beta}|X_1, \ldots, X_n) = \frac{\sum_{i=1}^n X_i^2 Var(u_i|X_1, \ldots, X_n)}{\left[\sum_{i=1}^n X_j^2\right]^2} = \frac{\sigma^2}{\sum_{i=1}^n X_j^2}$$

    (d) This follows the proof in the appendix.

15. Because the samples are independent, $\hat{\beta}_{m,1}$ and $\hat{\beta}_{w,1}$ are independent. Thus $\text{var}(\hat{\beta}_{m,1} - \hat{\beta}_{w,1}) = \text{var}(\hat{\beta}_{m,1}) + \text{var}(\hat{\beta}_{w,1})$. $\text{Var}(\hat{\beta}_{m,1})$ is consistently estimated as $[SE(\hat{\beta}_{m,1})]^2$ and $\text{Var}(\hat{\beta}_{w,1})$ is consistently estimated as $[SE(\hat{\beta}_{w,1})]^2$, so that $\text{var}(\hat{\beta}_{m,1} - \hat{\beta}_{w,1})$ is consistently estimated by $[SE(\hat{\beta}_{m,1})]^2 + [SE(\hat{\beta}_{w,1})]^2$, and the result follows by noting the SE is the square root of the estimated variance.