Problem Set 4

Junwoo Yang

April 18, 2021

1 Exercises

Exercise 4.2. A random sample of 100 20-year-old men is selected from a population and these men's height and weight are recorded. A regression of weight on height yields

$$\widehat{Weight} = -79.24 + 4.16 \times Height, \quad R^2 = 0.72, \quad SER = 12.6,$$

where *Weight* is measured in pounds and *Height* is measured in inches.

(a) What is the regression's weight prediction for someone who is 64 inches tall? 68 inches tall? 72 inches tall?

(b) A man has a late growth spurt and grows 2 inches over the course of a year. What is the regression's prediction for the increase in this man's weight?

(c) Suppose that instead of measuring weight and height in pounds and inches, these variables are measured in centimeters and kilograms. What are the regression estimates from this new centimeter-kilogram regression? (Give all results, estimated coefficients, R^2, and *SER*.)

Solution. (a) Simply inserting given values of height into the sample regression equation, we get 187, 203.64, and 220.28 as weight prediction in order.

(b) Since $\hat{\beta}_1 = 4.16$, prediction of weight increase is $2 \times 4.16 = 8.32$.

(c) Let's abbreviate *Weight* and *Height* as W and H, respectively. The original sample regression equation can be rewritten as $\hat{W}_{\text{lb},i} = -79.24 + 4.16\, H_{\text{in},i}$. Suppose that the new sample linear regression equation which measured in cm, kg is $\hat{W}_{\text{kg},i} = \hat{\alpha}_0 + \hat{\alpha}_1 H_{\text{cm},i}$. Since $1\,\text{in} = 2.54\,\text{cm}$ and $1\,\text{lb} = 0.453592\,\text{kg}$,

$$H_{\text{cm},i} = 2.54\, H_{\text{in},i}, \quad W_{\text{kg},i} = 0.453592\, W_{\text{lb},i}, \quad \hat{W}_{\text{kg},i} = 0.453592\, \hat{W}_{\text{lb},i}.$$

The followings are equivalent.

$$\begin{aligned}
\hat{W}_{\text{kg},i} &= \hat{\alpha}_0 + \hat{\alpha}_1 H_{\text{cm},i} \\
0.453592\, \hat{W}_{\text{lb},i} &= \hat{\alpha}_0 + 2.54\, \hat{\alpha}_1 H_{\text{in},i} \\
\hat{W}_{\text{lb},i} &= \frac{\hat{\alpha}_0}{0.453592} + \frac{2.54\, \hat{\alpha}_1}{0.453592} H_{\text{in},i} \\
\hat{W}_{\text{lb},i} &= -79.24 + 4.16\, H_{\text{in},i}
\end{aligned}$$

Thus,

$$\hat{\alpha}_0 = -79.24 \times 0.453592 = -35.94263, \quad \hat{\alpha}_1 = \frac{4.16 \times 0.453592}{2.54} = 0.7428908.$$

Note

$$\overline{W}_{\text{kg}} = \frac{1}{n} \sum_{i=1}^{n} W_{\text{kg},i} = \frac{1}{n} \sum_{i=1}^{n} 0.453592\, W_{\text{lb},i} = 0.453592\, \overline{W}_{\text{lb}}.$$

By the definition of the new regression R^2,

$$R^2 = \frac{\sum_{i=1}^n (\hat{W}_{\text{kg},i} - \overline{W}_{\text{kg}})^2}{\sum_{i=1}^n (W_{\text{kg},i} - \overline{W}_{\text{kg}})^2} = \frac{0.453592^2 \sum_{i=1}^n (\hat{W}_{\text{lb},i} - \overline{W}_{\text{lb}})^2}{0.453592^2 \sum_{i=1}^n (W_{\text{lb},i} - \overline{W}_{\text{lb}})^2} = 0.72.$$

The standard error of new regression is

$$SER = \sqrt{\frac{1}{n-2}\sum_{i=1}^n (W_{\text{kg},i} - \hat{W}_{\text{kg},i})^2} = \sqrt{\frac{1}{n-2}\sum_{i=1}^n (0.453592\, W_{\text{lb},i} - 0.453592\, \hat{W}_{\text{lb},i})^2}$$
$$= \sqrt{\frac{0.453592^2}{n-2}\sum_{i=1}^n (W_{\text{lb},i} - \hat{W}_{\text{lb},i})^2} = 0.453592 \times 12.6 = 5.715259.$$

Let's check $SE(\hat{\alpha}_1)$ and $SE(\hat{\alpha}_0)$, too. Recall that variance of $\hat{\alpha}_1$ is

$$\sigma^2_{\hat{\alpha}_1} = \frac{1}{n}\frac{\text{Var}[(H_{\text{cm},i} - \mu_{H_{\text{cm}}})u_i]}{[\text{Var}(H_{\text{cm},i})]^2}.$$

Then we get

$$\begin{aligned} SE(\hat{\alpha}_1) &= \sqrt{\frac{1}{n}\frac{\frac{1}{n-2}\sum_{i=1}^n (H_{\text{cm},i} - \overline{H}_{\text{cm}})^2 \hat{u}_i^2}{\left[\frac{1}{n}\sum_{i=1}^n (H_{\text{cm},i} - \overline{H}_{\text{cm}})^2\right]^2}} \\ &= \sqrt{\frac{1}{n}\frac{\frac{1}{n-2}\sum_{i=1}^n (2.54H_{\text{lb},i} - 2.54\overline{H}_{\text{lb}})^2 (W_{\text{kg},i} - \hat{\alpha}_0 - \hat{\alpha}_1 H_{\text{cm},i})^2}{\left[\frac{1}{n}\sum_{i=1}^n (2.54H_{\text{lb},i} - 2.54\overline{H}_{\text{lb}})^2\right]^2}} \\ &= \sqrt{\frac{1}{n}\frac{\frac{1}{n-2}\sum_{i=1}^n (H_{\text{lb},i} - \overline{H}_{\text{lb}})^2 0.453592^2 (W_{\text{cm},i} - 79.24 - 4.16H_{\text{in},i})^2}{2.54^2\left[\frac{1}{n}\sum_{i=1}^n (H_{\text{lb},i} - \overline{H}_{\text{lb}})^2\right]^2}} \\ &= \frac{0.453592}{2.54}\, SE(\hat{\beta}_1). \end{aligned}$$

Now for $SE(\hat{\alpha}_0)$,

$$\sigma^2_{\hat{\alpha}_0} = \text{Var}(\overline{W}_{\text{kg},i} - \hat{\alpha}_1 \overline{H}_{\text{cm}}) = \overline{H}^2_{\text{cm}}\, \text{Var}(\hat{\alpha}_1) = \overline{H}^2_{\text{cm}} \sigma^2_{\hat{\alpha}_1},$$
$$SE(\hat{\alpha}_0) = \overline{H}_{\text{cm}}\, SE(\hat{\alpha}_1) = 2.54\, \overline{H}_{\text{in}} \frac{0.453592}{2.54}\, SE(\hat{\beta}_1) = 0.453592\, SE(\hat{\beta}_0).$$

Note that $\hat{\alpha}_0$, $SE(\hat{\alpha}_0)$, SER change identically, and $\hat{\alpha}_1$, $SE(\hat{\alpha}_1)$ change identically, but R^2 is invariant to both units.

Exercise 4.9. (a) A linear regression yields $\hat{\beta}_1 = 0$. Show that $R^2 = 0$.

(b) A linear regression yields $R^2 = 0$. Does this imply that $\hat{\beta}_1 = 0$?

Solution. (a) Since $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i = \overline{Y} - \hat{\beta}_1 \overline{X} + \hat{\beta}_1 X_i = \overline{Y} + \hat{\beta}_1 (X_i - \overline{X}) = \overline{Y}$,

$$ESS = \sum_{i=1}^n (\hat{Y}_i - \overline{Y})^2 = \sum_{i=1}^n (\overline{Y} - \overline{Y})^2 = 0 \quad \Rightarrow \quad R^2 = 0.$$

(b) $R^2 = 0 \Rightarrow ESS = 0 \Rightarrow \hat{Y}_i = \overline{Y}\ \forall i$. Subtracting $\overline{Y}$ from $\overline{Y} + \hat{\beta}_1 (X_i - \overline{X}) = \overline{Y}$, we get $\hat{\beta}_1 = 0$ or $X_i = \overline{X}\ \forall i$. However if $X_i = \overline{X}$ for all i, that is X_i are constant, $\hat{\beta}_1$ is undefined. Thus, $\hat{\beta}_1 = 0$

Exercise 4.11. Consider the regression model $Y_i = \beta_0 + \beta_1 X_i + u_i$.

(a) Suppose you know that $\beta_0 = 0$. Derive a formula for the least squares estimator of β_1.

(b) Suppose you know that $\beta_0 = 4$. Derive a formula for the least squares estimator of β_1.

Solution. (a) Since $u_i = Y_i - \beta_1 X_i$, the OLS estimator of β_1 is

$$\hat{\beta}_1 = \arg\min_{\beta_1} \sum_{i=1}^{n} (Y_i - \beta_1 X_i)^2.$$

The derivative of sum of square error is

$$\frac{d}{d\beta_1} \sum_{i=1}^{n} (Y_i - \beta_1 X_i)^2 = -2 \sum_{i=1}^{n} (Y_i - \beta_1 X_i) X_i,$$

and first order condition says that this should be 0 when $\beta_1 = \hat{\beta}_1$, that is,

$$\sum_{i=1}^{n} (Y_i - \hat{\beta}_1 X_i) X_i = 0.$$

Hence, we get

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} X_i Y_i}{\sum_{i=1}^{n} X_i^2}.$$

(b) Similarly, when $u_i = Y_i - 4 - \beta_1 X_i$, $\hat{\beta}_1$ should satisfy

$$\sum_{i=1}^{n} (Y_i - 4 - \hat{\beta}_1 X_i) X_i = 0.$$

Thus, we get

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} X_i (Y_i - 4)}{\sum_{i=1}^{n} X_i^2} = \frac{\sum_{i=1}^{n} X_i Y_i - 4n\overline{X}}{\sum_{i=1}^{n} X_i^2}.$$

Exercise 4.12 (optional). (a) Show that the regression R^2 in the regression of Y on X is the squared value of the sample correlation between X and Y. That is, show that $R^2 = r_{XY}^2$.

(b) Show that the R^2 from the regression of Y on X is the same as the R^2 from the regression of X on Y.

(c) Show that $\hat{\beta}_1 = r_{XY}(s_Y / s_X)$, where r_{XY} is the sample correlation between X and Y, and s_X and s_Y are the sample standard deviations of X and Y.

Solution. (a) Note that sample variances and sample covariance are

$$s_X^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \overline{X})^2, \quad s_Y^2 = \frac{1}{n-1} \sum_{i=1}^{n} (Y_i - \overline{Y})^2, \quad s_{XY} = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \overline{X})(Y_i - \overline{Y}).$$

Again, since $\hat{Y}_i - \overline{Y} = \hat{\beta}_1 (X_i - \overline{X})$,

$$\begin{aligned} R^2 &= \frac{\sum_{i=1}^{n} (\hat{Y}_i - \overline{Y})^2}{\sum_{i=1}^{n} (Y_i - \overline{Y})^2} = \hat{\beta}_1^2 \frac{\sum_{i=1}^{n} (X_i - \overline{X})^2}{\sum_{i=1}^{n} (Y_i - \overline{Y})^2} = \left[\frac{\sum_{i=1}^{n} (X_i - \overline{X}) Y_i}{\sum_{i=1}^{n} (X_i - \overline{X})^2} \right]^2 \frac{\sum_{i=1}^{n} (X_i - \overline{X})^2}{\sum_{i=1}^{n} (Y_i - \overline{Y})^2} \\ &= \frac{\left[\sum_{i=1}^{n} (X_i - \overline{X})(Y_i - \overline{Y}) \right]^2}{\sum_{i=1}^{n} (X_i - \overline{X})^2 \sum_{i=1}^{n} (Y_i - \overline{Y})^2} = \frac{\left[\frac{1}{n-1} \sum_{i=1}^{n} (X_i - \overline{X})(Y_i - \overline{Y}) \right]^2}{\frac{1}{n-1} \sum_{i=1}^{n} (X_i - \overline{X})^2 \frac{1}{n-1} \sum_{i=1}^{n} (Y_i - \overline{Y})^2} \\ &= \frac{s_{XY}^2}{s_X^2 s_Y^2} = \left(\frac{s_{XY}}{s_X s_Y} \right)^2 = r_{XY}^2. \end{aligned}$$

(b) Consider the new sample regression equation $\hat{X}_i = \hat{\alpha}_0 + \hat{\alpha}_1 Y_i$. Proceeding as before, we get new regression $R^2 = r_{YX}^2$ by interchanging Y with X and replacing $\hat{\beta}_1$ with $\hat{\alpha}_1$, which is exactly same with the original regression R^2.

(c)

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (X_i - \overline{X})Y_i}{\sum_{i=1}^n (X_i - \overline{X})^2} = \frac{\frac{1}{n-1}\sum_{i=1}^n (X_i - \overline{X})(Y_i - \overline{Y})}{\frac{1}{n-1}\sum_{i=1}^n (X_i - \overline{X})^2} = \frac{s_{XY}}{s_X^2} = \frac{s_{XY}}{s_X s_Y}\frac{s_Y}{s_X} = r_{XY}\frac{s_Y}{s_X}.$$

Exercise 4.14. Show that the sample regression line passes through the point $(\overline{X}, \overline{Y})$.

Solution. Since the sample regression line is $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i = \overline{Y} - \hat{\beta}_1 \overline{X} + \hat{\beta}_1 X_i = \overline{Y} + \hat{\beta}_1 (X_i - \overline{X})$, it is trivial that the point $(X_i, \hat{Y}_i) = (\overline{X}, \overline{Y})$ lies on the sample regression line.

2 Additional questions

Exercise 4.5 (3rd edition). A professor decides to run an experiment to measure the effect of time pressure on final exam scores. He gives each of the 400 students in his course the same final exam, but some students have 90 minutes to complete the exam while others have 120 minutes. Each student is randomly assigned one of the examination times based on the flip of a coin. Let Y_i denote the number of points scored on the exam by the i^{th} student ($0 \leq Y_i \leq 100$), let X_i denote the amount of time that the student has to complete the exam ($X_i = 90$ or 120), and consider the regression model $Y_i = \beta_0 + \beta_1 X_i + u_i$.

(a) Explain what the term u_i represents. Why will different students have different values of u_i?

(b) Explain why $E(u_i|X_i) = 0$ for this regression model.

(c) Are the other assumptions in Key Concept 4.3 satisfied? Explain.

(d) The estimated regression is $\hat{Y}_i = 49 + 0.24 X_i$.

(i) Compute the estimated regression's prediction for the average score of students given 90 minutes to complete the exam. Repeat for 120 minutes and 150 minutes.

(ii) Compute the estimated gain in score for a student who is given an additional 10 minutes on the exam.

Solution. (a) u_i is a value that represents the difference between the observed value and value of the population regression function given X_i, which refers to other variables that explain the scores other than time.

(b) Because X is randomly assigned in an experiment, first assumption, $E(u_i|X_i) = 0$, hods.

(c) It depends upon how we set the population. If population is set to all the students who took the class that the professor teaches, these 400 students should be considered random-extracted students from population to satisfy the second assumption. Since both X_i and Y_i are finite values, both have finite fourth moment. Therefore, third assumption is satisfied.

(d) (i) $49 + 0.24 \times 90 = 70.6$, $49 + 0.24 \times 120 = 77.8$, $49 + 0.24 \times 150 = 85$.

(ii) $0.24 \times 10 = 2.4$.

Appendix 4.2. SW Appendix 4.2 provides the derivation of the OLS estimators $\hat{\beta}_0$ and $\hat{\beta}_1$ in the simple regression model, $Y_i = \beta_0 + \beta_1 X_i + u_i$, $i = 1, \cdots, n$. To get the final expression for $\hat{\beta}_1$ in

$$\hat{\beta}_1 = \frac{\frac{1}{n}\sum_{i=1}^n X_i Y_i - \overline{X}\,\overline{Y}}{\frac{1}{n}\sum_{i=1}^n X_i^2 - \overline{X}^2} = \frac{\sum_{i=1}^n (X_i - \overline{X})(Y_i - \overline{Y})}{\sum_{i=1}^n (X_i - \overline{X})^2} \tag{4.25}$$

show the following equalities:

(a) $\sum_{i=1}^{n} X_i^2 - n\overline{X}^2 = \sum_{i=1}^{n}(X_i - \overline{X})^2$

(b) $\sum_{i=1}^{n} X_i Y_i - n\overline{X}\,\overline{Y} = \sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})$

Solution.

$$\begin{aligned}
\sum_{i=1}^{n} X_i^2 - n\overline{X}^2 &= \sum_{i=1}^{n} X_i^2 - 2n\overline{X}^2 + n\overline{X}^2 \\
&= \sum_{i=1}^{n} X_i^2 - 2\overline{X}\sum_{i=1}^{n} X_i + \sum_{i=1}^{n} \overline{X}^2 \\
&= \sum_{i=1}^{n}(X_i^2 - 2\overline{X}X_i + \overline{X}^2) = \sum_{i=1}^{n}(X_i - \overline{X})^2, \\
\sum_{i=1}^{n} X_i Y_i - n\overline{X}\,\overline{Y} &= \sum_{i=1}^{n} X_i Y_i - n\overline{X}\,\overline{Y} - n\overline{X}\,\overline{Y} + n\overline{X}\,\overline{Y} \\
&= \sum_{i=1}^{n} X_i Y_i - \overline{Y}\sum_{i=1}^{n} X_i - \overline{X}\sum_{i=1}^{n} Y_i + \sum_{i=1}^{n} \overline{X}\,\overline{Y} \\
&= \sum_{i=1}^{n}(X_i Y_i - \overline{Y}X_i - \overline{X}Y_i + \overline{X}\,\overline{Y}) \\
&= \sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y}).
\end{aligned}$$

Appendix 4.3. SW Appendix 4.3 shows that $\hat{\beta}_1 = \frac{s_{XY}}{s_X^2}$ is an unbiased estimator of β_1 in the simple linear regression model, $Y_i = \beta_0 + \beta_1 X_i + u_i$, $i = 1, \cdots, n$, given the three least squares assumptions (LSA) (in Key Concept 4.3). Take a close look at

$$\hat{\beta}_1 = \beta_1 + \frac{\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})u_i}{\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2} \tag{4.28}$$

and

$$\begin{aligned}
E(\hat{\beta}_1 | X_1, \cdots, X_n) &= \beta_1 + E\left[\left.\frac{\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})u_i}{\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2}\right| X_1, \cdots, X_n\right] \\
&= \beta_1 + \frac{\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})E(u_i | X_1, \cdots, X_n)}{\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2},
\end{aligned} \tag{4.29}$$

and make sure you understand where in the proof LSA #1 and #2 are used.

Solution. The second least squares assumption says that (X_i, Y_i) are i.i.d., so u_i is distributed independently of X for all observations other than i. That is,

$$E(u_i | X_1, \cdots, X_n) = E(u_i | X_i) = 0$$

where second equality due to the first least squares assumption directly. We then get

$$\beta_1 + \frac{\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})E(u_i | X_1, \cdots, X_n)}{\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2} = \beta_1 + \frac{\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})E(u_i | X_i)}{\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2} = \beta_1,$$

and by the law of iterated expectations,

$$E(\hat{\beta}_1) = E(E(\hat{\beta}_1 | X_1, \cdots, X_n)) = E(\beta_1) = \beta_1.$$

Thus $\hat{\beta}_1$ is an unbiased estimator of β_1.