MS Module 16: Regression estimates – practice problems

(The attached PDF file has better formatting.)

Exercise 16.1: Least squares estimator for β_1

! A linear regression uses the N points X_i = {1, 2, …, 10, 11}
! The least squares estimator for β_1 is a linear function of the Y values = $\Sigma \gamma_i Y_i$

(The textbook uses the notation $\beta_1 = \Sigma c_i Y_i$)

A. What is $\overline{x}$, the mean X value?
B. What is S_{xx}, the sum of squared residuals for the X values?
C. What is γ_2, the coefficient of the Y value corresponding to X=2, in the estimate of β_1?

Part A: The mean X value is (1 + 2 + 3 + 4 + 5 + 6 + 7 + 8 + 9 + 10 + 11) / 11 = 6

Part B: S_{xx}, the sum of squared residuals for the X values, is

$(1\text{-}6)^2 + (2\text{-}6)^2 + (3\text{-}6)^2 + (4\text{-}6)^2 + (5\text{-}6)^2 + (6\text{-}6)^2 + (7\text{-}6)^2 + (8\text{-}6)^2 + (9\text{-}6)^2 + (10\text{-}6)^2 + (11\text{-}6)^2 = 110$

Part C: $\gamma_i = (x_i - \overline{x}) / S_{xx} = (2 - 6) / 110 = -0.03636$

Question: How does this formula relate to the formula $\beta_1 = S_{xy} / S_{xx}$?

Answer: Expand the formula $\beta_1 = S_{xy} / S_{xx} = \Sigma (x_i - \overline{x}) (y_i - \overline{y}) / S_{xx}$ =

$$\Sigma (x_i - \overline{x}) \times y_i / S_{xx} - \Sigma (x_i - \overline{x}) \times \overline{y} / S_{xx} = \Sigma \gamma_i Y_i - 0$$

The value of $\overline{y} / S_{xx}$ is independent of the subscript *i*, so $\Sigma (x_i - \overline{x}) \times \overline{y} / S_{xx} = [\Sigma(x_i - \overline{x})] \times [\overline{y} / S_{xx}] = 0$, and

$$\Sigma (x_i - \overline{x}) \times y_i / S_{xx} - \Sigma (x_i - \overline{x}) \times \overline{y} / S_{xx} = \Sigma \gamma_i Y_i$$

Exercise 16.2: Summary statistics

A regression analysis on 10 data points has summary statistics

- $\Sigma x_i = 40$
- $\Sigma y_i = 20$
- $\Sigma x_i^2 = 4{,}000$
- $\Sigma y_i^2 = 1{,}200$
- $\Sigma x_i y_i = 1{,}600$

A. What is $\overline{x}$, the average X value?
B. What is $\overline{y}$, the average Y value?
C. What is S_{xx}, the sum of squares of the X values?
D. What is S_{yy}, the sum of squares of the Y values?
E. What is S_{xy}, the cross sum of squares of the X and Y values?
F. What is the least squares estimate for β_1?
G. What is the least squares estimate for β_0?
H. What is the error sum of squares SSE?
I. What is s^2, the least squares estimate for σ^2?
J. What is the correlation ρ between X and Y?
K. What is the least squares estimate for R^2?

Part A: The average X value is $\overline{x} = \Sigma x_i / N = 40 / 10 = 4$

Part B: The average Y value is $\overline{y} = \Sigma y_i / N = 20 / 10 = 2$

Part C: S_{xx}, the sum of squared deviations of the X values, is $\Sigma x_i^2 - N \times \overline{x}^2 = \Sigma x_i^2 - (\Sigma x_i)^2/N =$

$$4{,}000 - 10 \times 4^2 = 3{,}840$$

Part D: S_{yy}, the sum of squares of the Y values (the total sum of squares SST), is $\Sigma y_i^2 - N \times \overline{y}^2 =$

$$1{,}200 - 10 \times 2^2 = 1{,}160$$

Part E: S_{xy}, the cross sum of squares of the X and Y values, is $\Sigma x_i y_i - N \times \overline{x} \times \overline{y} =$

$$1{,}600 - 10 \times 4 \times 2 = 1{,}520$$

Part F: The least squares estimate for β_1 is $S_{xy} / S_{xx} = 1{,}520 / 3{,}840 = 0.395833$

Part G: The least squares estimate for β_0 is $\overline{y} - \beta_1 \times \overline{x} = 2 - 0.39583333 \times 4 = 0.416667$

Part H: The error sum of squares SSE is $\Sigma y_i^2 - \beta_0 \times \Sigma y_i - \beta_1 \times \Sigma x_i y_i =$

$1{,}200 - 0.41666667 \times 20 - 0.39583333 \times 1{,}600 = 558.333339$

Answer: Do we need so many significant digits?

Answer: Extra significant digits are not used in real problems, since they give a false sense of accuracy. The practice problems show many significant digits so that when you work the problems on a spread-sheet or a calculator you can check your answers.

Some terms have very small numbers multiplied by very large numbers. If you round 0.00149×200 to 0.001×200, your solution may be incorrect.

The textbook says "in computing $\hat{\beta}_0$, use extra digits in $\hat{\beta}_1$, because, if $\overline{x}$ is large in magnitude, rounding may affect the final answer." See page 619 for an example.

Part I: The value of s^2, the least squares estimate for σ^2, is SSE / (N-2) = 558.3333 / (10 – 2) = 69.7917

Part J: The correlation ρ between X and Y is S_{xy} / ($S_{xx} \times S_{yy})^{½}$ =

$$1{,}520 / (3{,}840 \times 1{,}160)^{0.5} = 0.720193$$

Part K: R^2 is 1 – SSE/SST; SST is the same as S_{yy}.

$$R^2 = 1 - 558.333333 / 1{,}160 = 0.518678$$

Note that R^2 is the square of the correlation between X and Y: $0.720193^2 = 0.518678$

Exercise 16.3: Estimating σ^2

A statistician estimating σ^2 for a regression analysis mistakenly uses divides SSE (the error sum of squares) by (n-1) instead of (n-2), where n is the number of observations.

If the population is normally distributed, n = 17, and $\sigma^2 = 4$:

A. What is the expected value of the statistician's estimator?
B. What is the bias of the statistician's estimator?

Part A: Let s^2 be the unbiased estimator of σ^2, using a denominator of (n-2). The mistaken estimator using a denominator of (n-1) is $s^2 \times (n-2)/(n-1)$, and its expected value is $\sigma^2 \times (n-2)/(n-1)$.

Part B: The bias of the mistaken estimator is $\sigma^2 \times (n-2)/(n-1) - \sigma^2 = -\sigma^2/(n-1)$. For n = 17 and $\sigma^2 = 4$, the bias is –4/(17-1) = –4/16 = –0.250.

(See Example 7.6 on pages 339-340 of the textbook (second edition) or page 333 of the first edition)