MS Module 20: Residuals and Standardized Residuals – practice problems

(The attached PDF file has better formatting.)

Exercise 20.1: Standardized residuals

An actuary regresses the five Y values on the five X values shown below.

X value	–2	–1	0	1	2
Y value	–4	0	2	2	0

A. What is S_{xx}?
B. What is S_{xy}?
C. What is β_1?
D. What is β_0?
E. What is the fitted value at x = -2?
F. What is the residual at x = -2?
G. What is the error sum of squares (SSE)?
H. What is s^2, the least squares estimate of σ^2?
I. What is the variance of the residual at x = -2?
J. What is the standard deviation of the residual at x = -2?
K. What is the standardized residual at x = -2?

Part A: The mean of the X values, $\bar{x}$, is (-2 + -1 + 0 + 1 + 2) / 5 = 0.

S_{xx} (the sum of squared deviations) is $(-2 - 0)^2 + (-1 - 0)^2 + (0 - 0)^2 + (1 - 0)^2 + (2 - 0)^2 = 10$

Part B: The mean of the Y values, $\bar{y}$, is (-4 + 0 + 2 + 2 + 0) / 5 = 0.

S_{xy} (the sum of cross deviations) is -2 × -4 + -1 × 0 + 0 × 2 + 1 × 2 + 2 × 0 = 10

Part C: $\beta_1 = S_{xy} / S_{xx} = 10 / 10 = 1$

Part D: $\beta_0 = \bar{y} - \beta_1 \times \bar{x} = 0 - 1 \times 0 = 0$

Part E: The fitted values are the same as the x values (since $\beta_0 = 0$ and $\beta_1 = 1$), so the fitted value at x = -2 is -2.

Part F: The residual is the actual y value minus the fitted y value. At x = -2, this is -4 - -2 = -2.

Part G: The error sum of squares (SSE) is the sum of squared residuals (page 692).

! We calculate the residuals for the five points as {-2, 1, 2, 1, -2}.
! The sum of squared residuals is $-2^2 + 1^2 + 2^2 + 1^2 + -2^2 = 14$.

Part H: The value of s^2 (the least squares estimate of σ^2) = 14 / (N - 2) = 14 / (5 - 2) = 4.66667

Part I: The variance of the residual is $\sigma^2 \times (1 - 1/n - (x_i - \bar{x})^2 / S_{xx})$.

For $\sigma^2 = 4.66667$; n = 5; $x_i = -2$; $\bar{x} = 0$; $S_{xx} = 10$, the variance of the residual is

$$4.66667 \times (1 - 1/5 - (-2)^2 / 10) = 1.86667$$

Question: The formula for the width of the confidence interval uses $(1 + 1/n + (x_i - \overline{x})^2 / S_{xx})$ instead of $(1 - 1/n - (x_i - \overline{x})^2 / S_{xx})$. Why do the formulas have the opposite signs for 1/n and $(x_i - \overline{x})^2 / S_{xx}$?

Answer: Regression analysis assumes that the variance is the same at all points. The variance is the expected value of the squared deviation of the actual value from the fitted value. The deviation has two parts:

- The deviation caused by the fitted regression line, which may differ from the true regression line.
- The deviation caused by the difference of the observed value minus the expected value (the error term).

The regression line passes through the mean x-value. The slope of the regression line depends on the observed y-values. A change in the slope has a larger effect on the fitted y-value if the x-value is farther from the mean x-value. The fitted value moves closer to the observed value, and the deviation of the residual (the observed value minus the fitted value) is smaller.

Question: How do random fluctuations affect the residual at points close to vs far from the mean x-value?

Answer: Points close to the mean of X have little influence on the slope of the regression line (β_1). A random fluctuation in the observed Y value at $x = \overline{x}$ causes the regression line to shift up or down but does not change its slope. The shift up or down is spread evenly over all points, with a variance of σ^2/n. The change in the variance of the residual is $\sigma^2 - \sigma^2 / n = \sigma^2 \times (1 - 1/n)$. [This explanation is heuristic (intuitive), not rigorous. The mathematics helps you remember the formulas; it does not prove the formulas.]

Points far from the mean of X affect the slope of the regression line (β_1). A random fluctuation in the observed Y value at a point far from $\overline{x}$ causes the regression line to change its slope (in addition to shifting up or down). The fitted value moves in the same direction as the observed value, so the residual has a smaller variance.

Part J: The standard deviation of the residual at x = -2 is $1.86667^{0.5} = 1.36626$

Part K: The standardized residual at x = -2 is -2 / 1.36626 = -1.46385

Question: I checked this practice problem with the Regression add-in Excel's Analysis Pack. Excel gives the same regression coefficients and residuals, but different standardized residuals.

Answer: See the explanation at the web page:

https://stats.stackexchange.com/questions/166533/how-exactly-are-standardized-residuals-calculated

(The explanation is attached as a PDF file to this posting.)

Statisticians use several standardized residuals and studentized residuals. Excel does not explain its formula, and the comments on this web-site say it is not correct. You are responsible for the version in the textbook.

Exercise 20.2: Standardized residuals

A regression model has the independent variable X values {1, 2, …, 10, 11}.

- At the point x = 2, the residual is 0.500 and the standardized residual is 0.300.
- The residual at the point x = 11 is –0.750.

A. What is $\bar{x}$, the mean X value?
B. What is S_{xx}, the sum of squared deviations of the X values?
C. What is the ratio of the residual to the standardized residual at the point x = 2?
D. What is *s*, the least squares estimate for σ?
E. What is the standardized residual at the point x = 11?

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

$(1 – 6)^2 + (2 – 6)^2 + (3 – 6)^2 + (4 – 6)^2 + (5 – 6)^2 + (6 – 6)^2 + (7 – 6)^2 + (8 – 6)^2 + (9 – 6)^2 + (10 – 6)^2 + (11 – 6)^2 = 110$

Part C: The ratio of the residual to the standardized residual at the point x = 2 is 0.5 / 0.3 = 1.66667

Part D: The variance of the residual is $\sigma^2 \times (1 - 1/n - (x_i - \bar{x})^2 / S_{xx})$ and the mean residual is zero, so

the standardized residual = (the residual – the mean residual) / the standard deviation of the residuals ⇛
the standardized residual = the residual / $[\sigma^2 \times (1 - 1/n - (x_i - \bar{x})^2 / S_{xx})]^{½}$ ⇛
$\sigma \times [(1 - 1/n - (x_i - \bar{x})^2 / S_{xx})]^{½}$ = the residual / the standardized residual

At the point x = 2, $[(1 - 1/n - (x_i - \bar{x})^2 / S_{xx})]^{½} = (1 - 1/11 - (2 - 6)^2 / 110)^{0.5} = 0.87386$ ⇛

$0.87386 \times s = 5 / 3$ ⇛ $s = (5 / 3) / 0.87386 = 1.90725$

Part E: We derive the ratio of the residual to the standardized residual at the point x = 11.

At the point x = 11, $[(1 - 1/n - (x_i - \bar{x})^2 / S_{xx})]^{½} = (1 - 1/11 - (11 - 6)^2 / 110)^{0.5} = 0.82572$ ⇛

$0.82572\ s = 0.82572 \times 1.90725 = 1.57485$

The standardized residual at the point x = 11 is -0.750 / 1.57485 = -0.47624

Exercise 20.3: Variance of residuals

A linear regression uses the N points $X_i = \{1, 2, \ldots, 11\}$

The error sum of squares (SSE) is 36.

A. What is s^2, the least squares estimator of σ^2?
B. What is S_{xx}, the sum of squared deviations for the X variable?
C. What is the variance of the residual at x = 2?

Part A: The value of s^2, the estimate of σ^2, is SSE/(N-2) = 36 / (11-2) = 4.

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

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

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

Part C: The variance of the residual is $\sigma^2 \times (1 - 1/n - (x_i - \overline{x})^2 / S_{xx})$.

We use $s^2 = 4$ as the estimate for σ^2; n = 11; $x_i = 2$; $\overline{x} = 6$; and $S_{xx} = 110$. The variance of the residual is

$4 \times (1 - 1/11 - (2 - 6)^2 / 110) = 3.05455$

Question: The variance of the width of the prediction interval is proportional to $\sigma^2 \times (1 + 1/n + (x_i - \overline{x})^2 / S_{xx})$.

The width is narrowest at the mean X value $\overline{x}$ and wider at points farther away from the mean. The formula for the variance of the residuals has minus signs instead of the plus signs. If the prediction interval is wider, shouldn't the variance of the residual be greater?

Answer: This practice problem examines the variance of the residuals $(y_i - \hat{y}_i)$, not the variance of y_i.

Points farther from the mean X value have more influence on the slope of the regression line.

! If the Y value at the point $x = \overline{x}$ is higher than expected by a random fluctuation of 1 unit, the regression line is shifted up by 1/n units, but its slope does not change. The residual is (1 - 1/n), and the variance of the residual is $\sigma^2 \times (1 - 1/n)$.

! If the Y value at the point x much greater than $\overline{x}$ is higher than expected by a random fluctuation of 1 unit, the regression line is shifted up by 1/n units, and its slope increases, pulling the fitted value of y toward the observed value. The expected residual is $(1 - 1/n - (x_i - \overline{x})^2 / S_{xx})$., and the variance of the residual is $\sigma^2 \times (1 - 1/n - (x_i - \overline{x})^2 / S_{xx})$.