YData: An Introduction to Data Science

Lecture 32: Residuals

Jessi Cisewski-Kehe
Statistics & Data Science, Yale University
Spring 2020

Credit: data8.org

Announcements

Error in Estimation (Review)

- **error = actual value – estimate**
- Typically, some errors are positive and some negative
- To measure the rough size of the errors
  - **square** the **errors** to eliminate cancellation
  - take the **mean** of the squared errors
  - take the square **root** to fix the units
  - **root mean square error** (rmse)

Residuals

- Error in regression estimate
- One residual corresponding to each point (x, y)
- **residual**
  = observed y - regression estimate of y
  = observed y - height of regression line at x
  = vertical distance between the point and the best line

(DEMO)

Residual Plot

A scatter diagram of residuals

- Should look like an unassociated blob for linear relations
- But will show patterns for non-linear relations
- Used to check whether linear regression is appropriate

Regression Diagnostics

Dugong

(DEMO)

Properties of Residuals

Discussion Questions

- What should the average of the residuals be?
- Does your answer depend on whether the scatter diagram looks linear or shows a nonlinear pattern?

Average of Residuals

- The average of the residuals is always 0
- No matter what the scatter looks like

- Just as the average of the deviations from mean is always 0
- No matter what the data look like

(DEMO)

A Measure of Clustering

Correlation, Revisited

- "The correlation measures how clustered the points are about a straight line."
- We can now quantify this statement.

(DEMO)

SD of Fitted Values

$$\frac{\text{SD of fitted values}}{\text{SD of y}} = |r|$$

$$\text{SD of fitted values} = |r| * (\text{SD of y})$$

Variance of Fitted Values

$$\begin{aligned}\text{Variance} &= \text{Square of the SD} \\ &= \text{Mean Square of the Deviations}\end{aligned}$$

Variance has bad units, but good math properties

$$\frac{\text{Variance of fitted values}}{\text{Variance of y}} = r^2$$

A Variance Decomposition

$$\frac{\text{Variance of fitted values}}{\text{Variance of y}} = r^2$$

$$\frac{\text{Variance of residuals}}{\text{Variance of y}} = 1 - r^2$$

Residual Average and SD

- The average of residuals is always 0
- $\frac{\text{Variance of residuals}}{\text{Variance of y}} = 1 - r^2$
- SD of residuals $= \sqrt{1 - r^2}$ SD of y

(DEMO)

Discussion Question

Midterm: Average 70, SD 10
Final: Average 60, SD 15
$r = 0.6$

Fill in the blank:
For at least 75% of the students, the regression estimate of final score based on midterm score will be correct to within _____ points.