DATA 8

Spring 2020

Lecture 32

Residuals

Regression roadmap

- Last Monday:
  - Association and correlation
- Last Wednesday
  - Prediction, scatterplots and lines
- Last Friday:
  - Least squares: finding the “best” line for a dataset
- **Today:**
  - **Residuals: analyzing mistakes and errors**
- Wednesday
  - Regression inference: understanding uncertainty

Errors and Residuals

Error in Estimation

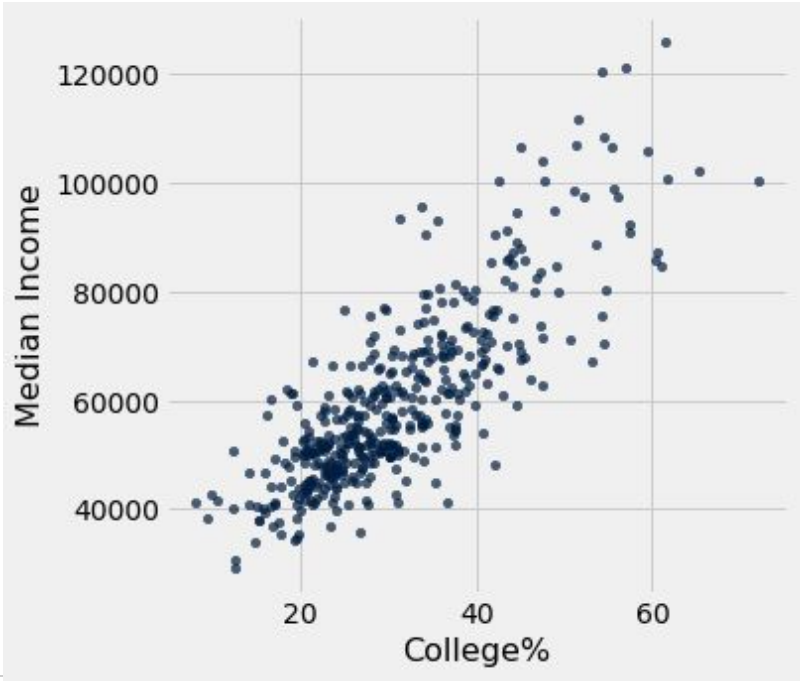
- **error = actual value - estimate**
- Some errors are positive and some negative
- To measure the rough size of the errors
  - **square** the **errors** to eliminate cancellation
  - take the **mean** of the squared errors
  - take the square **root** to fix the units
  - **root mean square error**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

Discussion Question

Based only on the graph, which must be true? Pick all that apply.

1. Going to college causes people to get higher incomes.
2. For any district, having more college-educated people live there causes median incomes to rise.
3. For any district, having a higher median income causes more college-educated people to move there.

USA Congressional Districts, 2016



Numerical Optimization

- Numerical minimization is approximate but effective
- Lots of machine learning uses numerical minimization
- If the function **`mse(a, b)`** returns the mse of estimation using the line "estimate = $ax + b$",
  - then **`minimize(mse)`** returns array $[a_0, b_0]$
  - a_0 is the slope and b_0 the intercept of the line that minimizes the mse among lines with arbitrary slope a and arbitrary intercept b (that is, among all lines)

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)

Regression Diagnostics

Example: Dugongs

(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
- Look for curves, trends, changes in spread, outliers, or any other patterns

Properties of residuals

- Residuals from a linear regression **always** have
  - **Zero** mean
    - (so **rmse = SD of residuals**)
  - **Zero** correlation with x
  - **Zero** correlation with the fitted values

- These are all true **no matter what the data look like**
  - Just like deviations from mean are zero on average

(Demo)

Discussion Questions

How would we adjust our regression line…

- if the average residual were 10?
- if the residuals were positively correlated with x?
- if the residuals were above 0 in the middle and below 0 on the left and right?

A Measure of Clustering

Correlation, Revisited

- “The correlation coefficient measures how clustered the points are about a straight line.”

- We can now quantify this statement.

(Demo)

SD of Fitted Values

- SD of fitted values

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

- SD of fitted values = $|r|$ * (SD of y)

Variance of Fitted Values

- Variance = Square of the SD
  = Mean Square of the Deviations
- Variance has weird units, but good math properties

- Variance of fitted values
  ---------- = r^2
  Variance of y

A Variance Decomposition

By definition,

$$y = \text{fitted values} + \text{residuals}$$

Tempting (but wrong) to think that:

~~$$SD(y) = SD(\text{fitted values}) + SD(\text{residuals})$$~~

But it **is** true that:

$$Var(y) = Var(\text{fitted values}) + Var(\text{residuals})$$

(a result of the **Pythagorean theorem!**)

A Variance Decomposition

Var(*y*) = Var(fitted values) + Var(residuals)

- Variance of fitted values

  ---------- = r^2

  Variance of *y*

- Variance of residuals

  ---------- = $1 - r^2$

  Variance of *y*

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 1

Midterm: Average 70, SD 10

Final: Average 60, SD 15

$r = 0.6$

Fill in the blank:

The SD of the residuals is _____.

Discussion Question 2

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.