DATA 8

Spring 2022

Lecture 32

Residuals

Announcements

- Homework 10 due Thursday, 04/14
  - Turn in on Wednesday for a bonus point
- Project 2 due Friday, 04/15
  - OH Party this Friday 1-5pm in SOCS 581
- Check out the staff-created tutoring videos

Weekly Goals

- **Today**
    - Least squares: finding the "best" line for a dataset
    - Residuals: analyzing mistakes and errors
- Wednesday
    - Regression inference
    - Uncertainty in the slope & intercept
- Friday
    - Data and privacy

Least Squares

(Demo)

Error in Estimation

- **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)

(Demo)

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)

(Demo)

Least Squares Line

- Minimizes the root mean squared error (rmse) among all lines

- Equivalently, minimizes the mean squared error (mse) among all lines

- Names:
  - “Best fit” line
  - Least squares line
  - Regression line

Errors and Residuals

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

(Demo)

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?