

Data 8, Project 3 Lab

Residuals and Regression Inference

Hubert Luo

Spring 2020

24 April 2020

Review: Linear Regression Overview

- Predict the value of a **continuous** random variable
 - Example: predict the number of students attending lab (dependent variable) using the air quality (independent variable)
- Dependent variable on the x axis, independent variable on the y axis
- Fit a regression line to represent the points on the graph
- Can use regression line to **predict** the y value for points on the x axis

Review: Linear Regression Equation

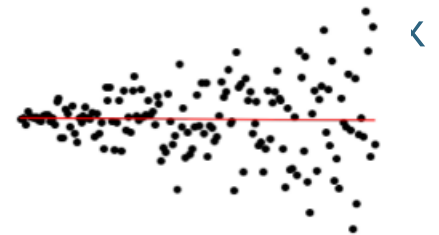
estimate of $y = \text{slope} * x + \text{intercept}$

$$\text{Slope} = r * \frac{\text{Standard Deviation of } y}{\text{Standard Deviation of } x}$$

Intercept = Average of y – Slope * Average of x

Residuals

- Residual = Actual Value of y – Predicted Value of y
- The residual plot of a good regression (a linear model) shows no patterns in the graph of the residuals
 - Will look like formless cloud if linear regression is a good model
- Average of residuals is always 0
- Heteroscedastic: Uneven variation of residuals around the horizontal line at 0
 - This means that the regression estimates are not equally accurate! For example, this model is better at values are small than when x values are large:



Standard Deviation of Residuals

- SD of the Residuals = $\text{np.sqrt}(1-r^{**2}) * \text{SD_of_y}$
- The SD tells us how good the linear predictor is
 - The smaller the SD of Residuals, the closer the residuals are to their mean
 - Mean of Residuals is always 0
 - Example: If $r=1$ (perfect correlation), there SD of residuals is 0 since we have a perfect linear relationship between X and y
- SD of Predicted Values = $|r| * \text{SD of } y$

Announcements

- Project 3:
 - Checkpoint is due this Friday 4/24
 - Project is due 5/1
- Watch year's lecture on privacy [here](#)
 - Material will be on the final!