Data 8, Lab 9

Linear Regression and Residuals

Hubert Luo

Spring 2020

17 April 2020

Berkeley
UNIVERSITY OF CALIFORNIA

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

Linear Regression Equation

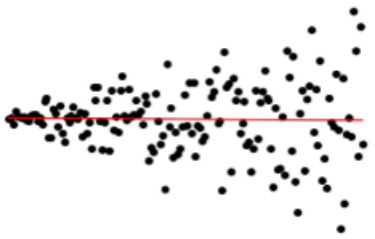
estimate of y = slope * x + intercept

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

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
- 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 predicting when x values are small than when x values are large:



Standard Deviation of Residuals

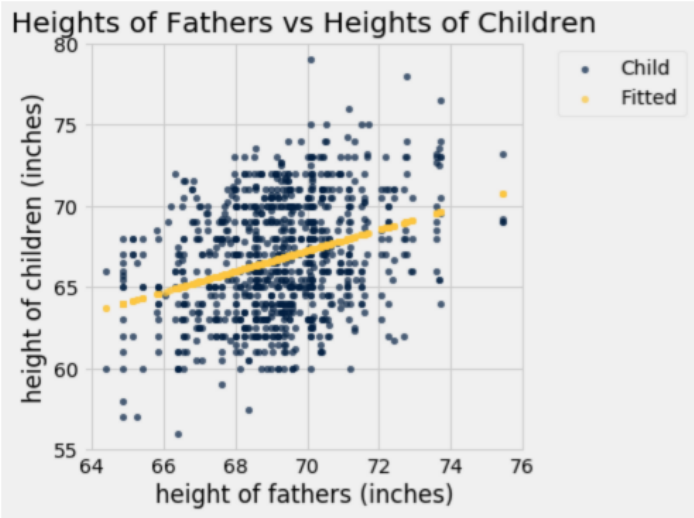
- SD of the Residuals = np.sqrt(1-r**2) * SD_of_y
- The SD tells us how good the linear predictor is
  - The smaller the SD of Residuals, the closer the residuals are to their mean
  - Mean of Residuals is always 0
  - Example: If r=1 (perfect correlation), there SD of residuals is 0 since we have a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

Announcements

- Project 2 is due today (4/17)
- Watch year's lecture on privacy here
  - Material will be on the final!
- For labs 7 through 12, we'll be adding one extra lab drop, and changing the policy for how to get credit: you only need to pass at least 80% of public tests to get full credit, and anything extra will be bonus points. See this Piazza post for more details.

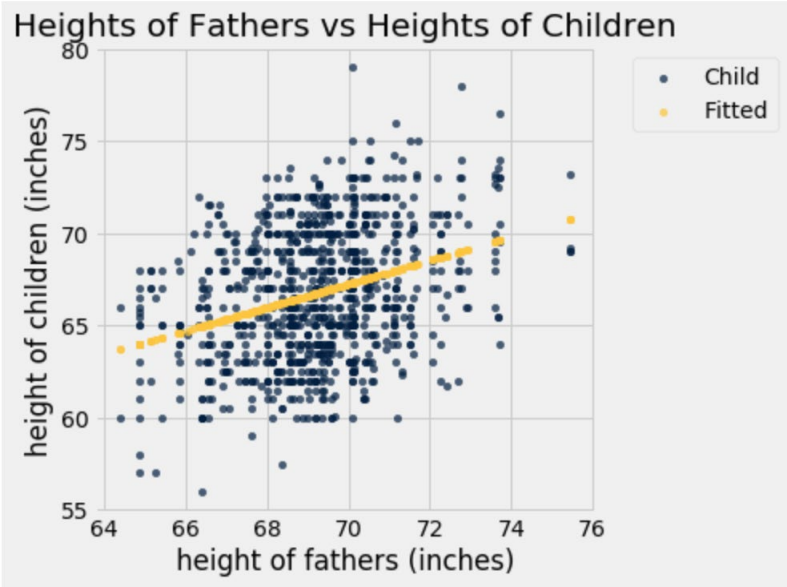
Worksheet Question 1

Question 1. Suppose you are given the scatter diagram shown below that shows the relationship between the height of fathers and the height of children. You have calculated the line of best fit (shown in red). Suppose you encounter a new family where the father has a height of 70 inches. How would you predict the height of the children in that family?



Worksheet Question 1 (Soln)

- Use the yellow regression line on the graph to the right
- Father's height is 70 inches so we are given that x=70
- At x=70, the corresponding y value on the regression line is about 67
- Therefore, predicted child height is around 67 inches



Worksheet Q2a

Question 2. We want to investigate the correlation between the daily ounces of coffee consumed by an individual and the number of hours the individual stayed awake on that day. It is our intention to use the ounces of coffee consumed to predict the number of hours the individual stayed awake. The data from our sample of 500 people has the following characteristics:

- The number of ounces of coffee consumed has a mean of 12 ounces and SD of 4
- The number of hours stayed awake has a mean of 16 and an SD of 2
- The correlation between the number of ounces of coffee consumed and number of hours spent awake is 0.5.
- Suppose the scatter plot is roughly linear.

What is the slope of the line of best fit?

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

$$= 0.5*2/4 = 0.25$$

Worksheet Q2b

Question 2. We want to investigate the correlation between the daily ounces of coffee consumed by an individual and the number of hours the individual stayed awake on that day. It is our intention to use the ounces of coffee consumed to predict the number of hours the individual stayed awake. The data from our sample of 500 people has the following characteristics:

- The number of ounces of coffee consumed has a mean of 12 ounces and SD of 4
- The number of hours stayed awake has a mean of 16 and an SD of 2
- The correlation between the number of ounces of coffee consumed and number of hours spent awake is 0.5.
- Suppose the scatter plot is roughly linear.

What is the intercept of the line of best fit?

Intercept = Average of y – Slope * Average of x

= 16 – 0.25*12 = 13

Worksheet Q2c

Question 2. We want to investigate the correlation between the daily ounces of coffee consumed by an individual and the number of hours the individual stayed awake on that day. It is our intention to use the ounces of coffee consumed to predict the number of hours the individual stayed awake. The data from our sample of 500 people has the following characteristics:

- The number of ounces of coffee consumed has a mean of 12 ounces and SD of 4
- The number of hours stayed awake has a mean of 16 and an SD of 2
- The correlation between the number of ounces of coffee consumed and number of hours spent awake is 0.5.
- Suppose the scatter plot is roughly linear.

Suppose your friend is in this population. She told you that she consumed 24 ounces of coffee that morning. Use your line of best fit to predict how many hours she will stay awake today.

Predicted y value = Slope * x + Intercept

= 0.25*24 + 13 = 19