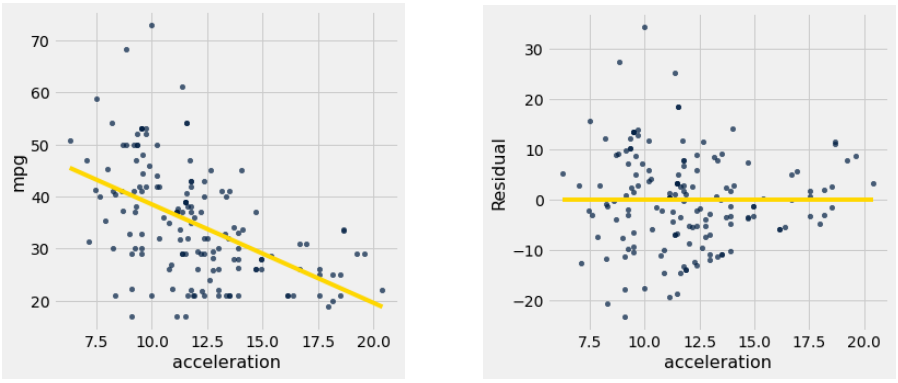
Data 8 Spring 2022

Project 3 Lab: Regression and Regression Inference

Residuals

In data science, we can use linear regression in order to make predictions. Moreover, we want to assess the accuracy of our predictions. To do so, we can examine the error between our actual data and the predictions; these errors are called *residuals*.

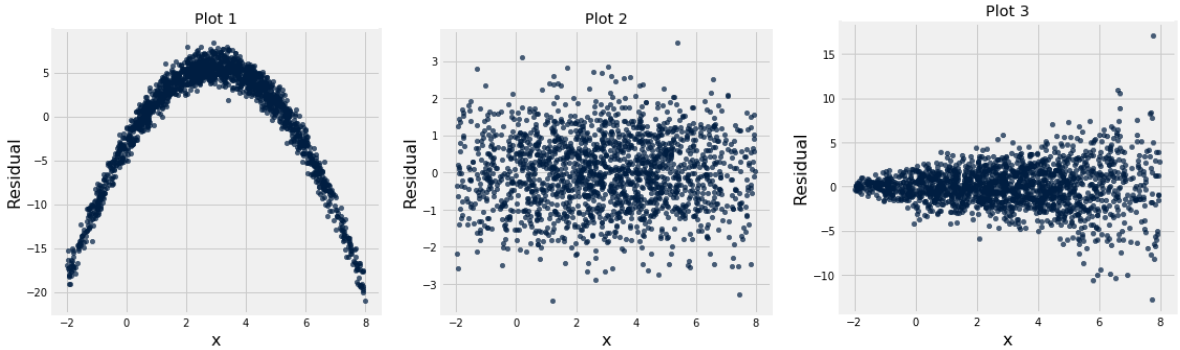
An example can be found below in the graph of miles per gallon compared to acceleration. The graph of the residuals is shown on the right. The yellow line is our regression line.



As a reminder:

- residual = y − estimated value of y = y − height of regression line at x
- The mean of residuals is zero and they show no trend (i.e. correlation is zero)

Question 1. Visual Diagnostic: Displayed below are three residual plots. For which of the following residual plots is using linear regression a reasonable idea, and why? What might the original graphs have looked like?



Question 2. Scooby Snacks: Will has a dataset consisting of a sample of 100 snacks. This dataset contains the calories from fat (`cal_fat`) and the calories total (`cal_total`) for each snack. He wants to use a snack's `cal_fat` to predict its `cal_total`. The correlation coefficient between the two variables is 0.6.

a. Will thinks that there is no correlation between `cal_fat` and `cal_total`, and that his sample was just biased. How can he test this hypothesis?

Null Hypothesis:

Alternative Hypothesis:

Describe Testing Method:

b. Will runs his hypothesis test and gets a 99% confidence interval of 0.24 to 0.89. Should he reject the null hypothesis?

c. Finally, Will wants to generate a line of best fit for his data. Should he use the method of least squares (i.e. minimizing RMSE) or the regression equations? Is there a difference between the two?

Question 3. Privacy Debrief

For the following questions, feel free to reference the Privacy Lecture slides!

a. What happened in the Cambridge Analytica Scandal?

b. What are disclosure, collection and inference, and can you come up with some examples for each?

c. What reactions did you have to the privacy lecture? Was anything surprising? Was anything frightening, hopeful, etc? As a data scientist, how can you help maintain privacy? Should you? Is inference ethical?