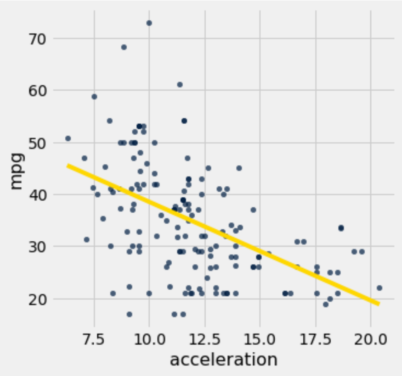
Data 8 Spring 2022

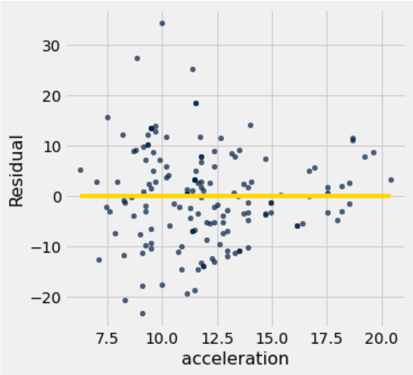
Project 3 Lab: Regression and Regression Inference

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.

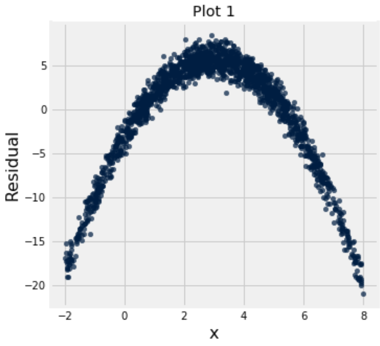


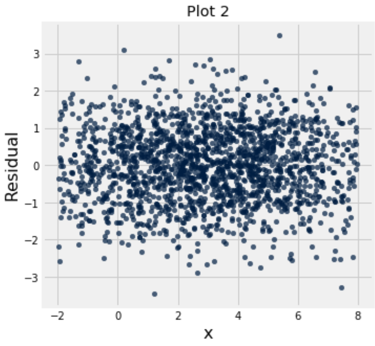


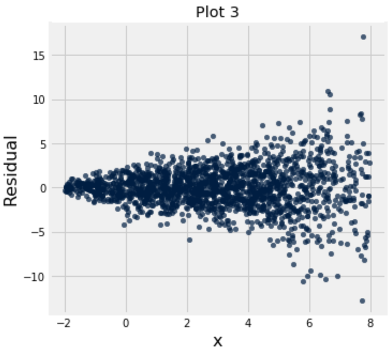
As a reminder:

- 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?

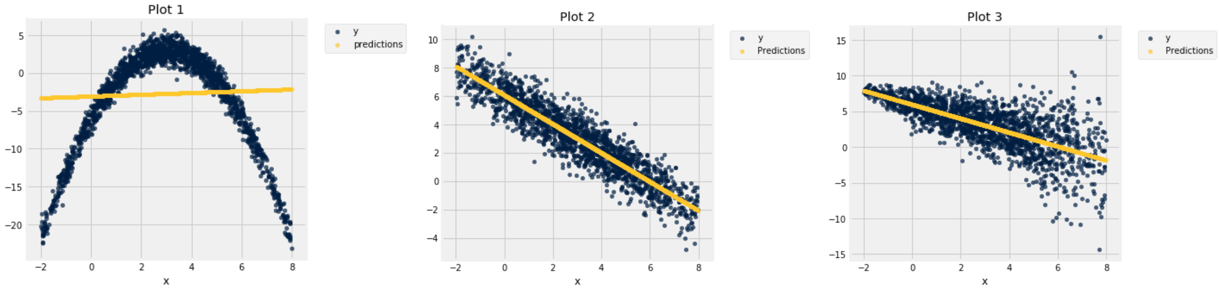






Plot 2 and Plot 3 are residual plots for which using linear regression on the original data is reasonable. Plot 2 is the best residual to use linear regression for, since the residuals have the pattern of a formless cloud. Note that using linear regression for plot 3 is reasonable, but the residual plot is heteroscedastic, which means that the residuals are unevenly spread for different x values (check out the textbook for examples!).

Here are the original graphs:



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: There is no association between fat calories and total calories, the true correlation is 0.
Alternative Hypothesis: The association between fat calories and total calories is not zero.

Describe Testing Method: Will should bootstrap his sample repeatedly, generate a confidence interval for the correlation and check to see if zero is in the CI.

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

Yes, Will should reject the null hypothesis, because 0 is not in the CI.

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?

It doesn't matter which method Will uses; they both result in the same line.

Question 3. Privacy Debrief

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

a. What happened in the Cambridge Analytica Scandal?

CA got private data from 50 million facebook users and used it to support different political movements. Facebook apps could access the private info of your friends if you installed it.

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

Disclosure: You explicitly tell them (ie filling out your age to prove that you are allowed to make an account)

1. &**Collection:** You didn't explicitly tell them but they observed you (ie web based advertisement tracking)

Inference: Inferring new data points from other data to determine values of new data (machine learning algorithms for political party for example)

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?

Open discussion, try and facilitate a conversation but discourage people from being judgmental of anyone's opinions/comments.