# CS031 Fall 2023

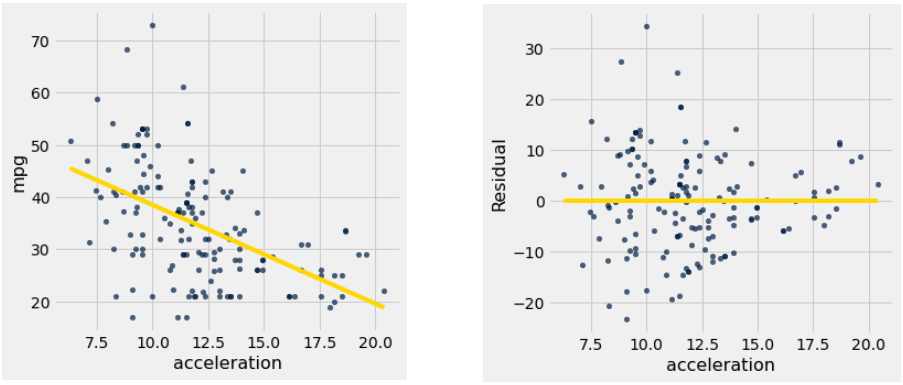
Project 3 Lab: Regression and Regression Inference

---

**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*.

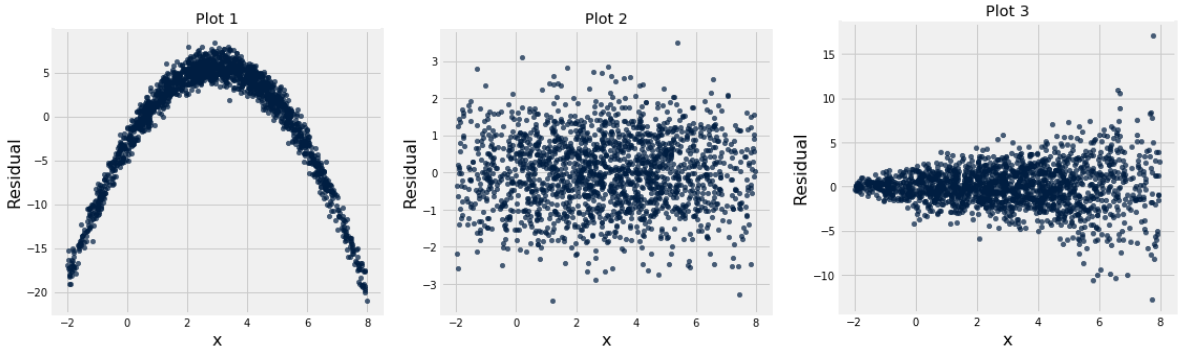
An example can be found below in the graph of miles per gallon compared to acceleration. The graph of the residuals is shown on the right. The yellow line is our regression line.




**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?




Plots 2 and 3 would be good datasets to use linear regression on, as there seem to be an equal number of points and magnitudes above and below the x-axis (though it would be wise to use error intervals in plot 3). Plot 2 would look like a typical cloud around the regression line, whereas plot 3 would look like a cloud that over time grows farther from the regression line in magnitude but still is patterned by the regression line.

**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:* The true regression line will have a slope of 0.

*Alternative Hypothesis:* The true regression line will have a nonzero slope.

*Describe Testing Method:* Bootstrap the sample many times and calculate the slope of the regression line of each sample, then use a confidence interval to estimate the slope of the true regression line.

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. Neither of these percentile values are even close to 0 in this case.

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?

He can use either, as there is no mathematical difference between the two (other than the rounding error present in computers). In other words, he should use whichever he feels is easier to implement.

**Question 3. Privacy Debrief**

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

a. What happened in the Cambridge Analytica Scandal?

In the Cambridge Analytica Scandal, an app called "This is Your Life" harvested data collected via survey of millions of Facebook users, and this data was used to aid in the election of Donald Trump and Ted Cruz. This data was harvested without explicit consent of its victims.

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

Disclosure: the sharing of personal information, either with or without consent. For example, creating a professional work email address typically requires disclosure of one's name.

Collection is the process of gathering personal information, either with or without consent. For example, Facebook collects analytics of users' usage based on things like location, age, race, gender, etc.

Inference is synthesizing unknown information based on what is known. For example, if one watches a lot of football videos on YouTube, the app will infer that the person likes watching football and will recommend more football videos. It can also infer that the watcher is male and might work more recommendations based on that (as well as age).

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?

I think the large-scale collection of data is both surprising and frightening, as it means privacy is dying in the 21st century. As a data scientist, I can help maintain privacy by refusing to use or create data with people's personal information that is not explicitly collected. It might be bad for profitability and being the best at data science, but if every data scientist abided by this code then privacy would be alive and well in the 21st century. Inference is ethical, but only if the data upon which it is based is ethical (meaning it is explicitly and transparently collected). This is because, at the end of the day, if one gives their information to a company, they should be able to freely think about and process it, so long as it is not shared elsewhere or collected indefinitely without explicitly and transparently collected consent. I was especially surprised that anonymized data can still be useful to track people and their bad habits.