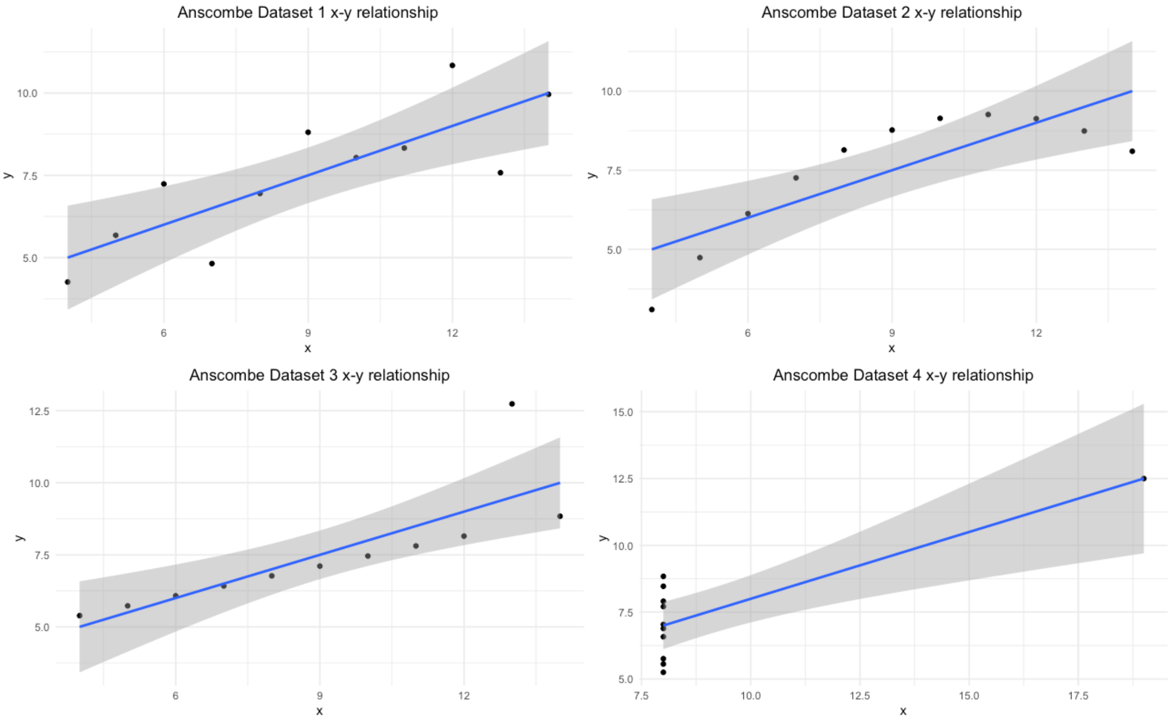
Blarry Wang
Rebecca U. Thorpe
December 5, 2019

# Homework 3 Report

## The Anscombe Dataset & Control Variables

## Question 4 - Plots for the Anscombe Datasets



## Question 5 - Similarities

The 4 datasets share same statistics. This means the means (averages) and standard deviations of x and y in each dataset are exactly the same. At the same time, the line we fit through these datasets with a bi-variate linear model also produced very similar results.

At a first glance, we could be tempted to say that the data points are sampled from the same distribution because these summary statistics and linear model results are very close. However, this assumption is falsified when we plot each of these datasets. They share very little commonalities and exhibit completely different patterns.

This serves as a warning that linear model's assumptions doesn't always hold for real data. Dataset 2 and 4 are clearly not linear by examining the plot, while Dataset 3's outlier completely shifted the results of the linear regression. In our own research, we need to be mindful of 1) the actual distribution of our samples (we need to plot our data to see if a linear model makes sense), and 2) outliers (and possibly removing them from analysis).

## Question 6 - Bivariate Regression

In short, a bivariate regression assumes a linear relationship between two variables — say, one unit of increase in global temperature results in a fixed number of units of increase in sea level. In the real world, we cannot possibly record all global temperatures and sea level across the world and throughout history and the future, instead we collect several recording at different locations and time. Bivariate regression tries to find that fixed number of units of increase in sea level by checking what number fits our samples. In addition, the result also tells us how confident the number is given how well the number did for our samples.

## Question 7 - Control Variables

The control variables can be collected from the Wikipedia page for the 2014 gubernatorial elections (https://en.wikipedia.org/wiki/2014_United_States_gubernatorial_elections?oldformat=true). It contains a table for all candidates who ran and their state, party, and whether they are an incumbent. It also contains a table of competitive races — namely races that are predicted as "tossups" by any of the 6 polling agencies (Cook, Daily Kos, Governing, Real Clear Politics, Rothenberg, Sabato).

For the variable attacked, we need the attack tweets from each candidate first. For each tweet, we label it with a binary variable indicating whether it was tweeted after an attack from their opponent.

| Control Variable | Description | Data Source |
|---|---|---|
| **state** | the state which candidate is from | Wikipedia |
| **party** | the party of candidate | Wikipedia |

| Control Variable | Description | Data Source |
|---|---|---|
| **incumbency** | whether or not the candidate is incumbent | Wikipedia |
| **attacked** | Existing literature points to a higher likelihood of attacking when attacked by their opponent. This variable shows whether the candidate has been attacked prior to the tweet. | Twitter |
| **competitiveness** | Existing work shows that candidates are more likely to tweet in competitive races. This variable shows if the race is competitive. | Wikipedia |