Capstone 1 - Inferential Statistics

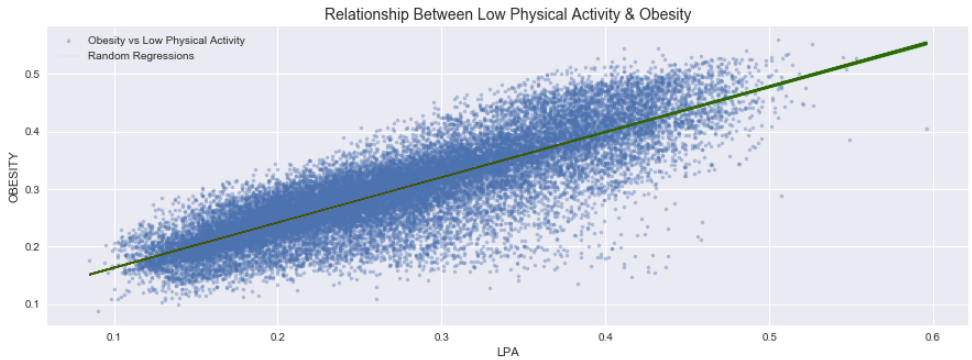
By Richard Wolff

After performing exploratory data analysis, there were 4 different series of data that seemed like they had a relationship with my target series of Obesity. These include:

* Low physical activity
* Less than 7 hours of sleep
* Population size
* Geography/State

For the first three variables, I checked correlation to the target.

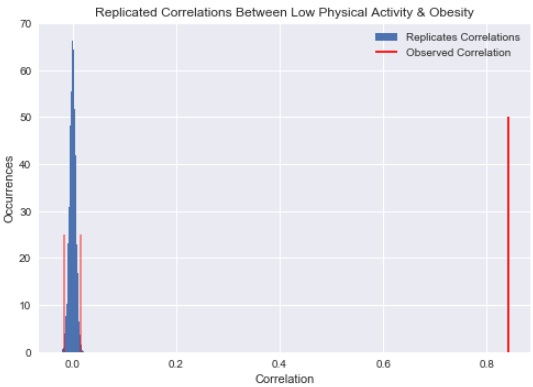
# Obesity Rate & Its Relationship to Low Physical Activity, Sleep, & Population Size



The relationship between low physical activity and obesity is very positive and has an observed correlation of 0.841. To make sure this wasn’t random chance, we tested that the correlation was in fact real.

**Null Hypothesis:** There is no correlation between Obesity and LPA.

**Test Statistic:** Pearson Correlation



I tested 10000 random permutations of the low physical activity set and calculated the correlation coefficient to obesity for each. We can see our obsereved correlation is much farther from the random noise and has a p-value of very close to 0. With this, we can reject the null hypothesis that there is no correlation between Obesity and LPA.

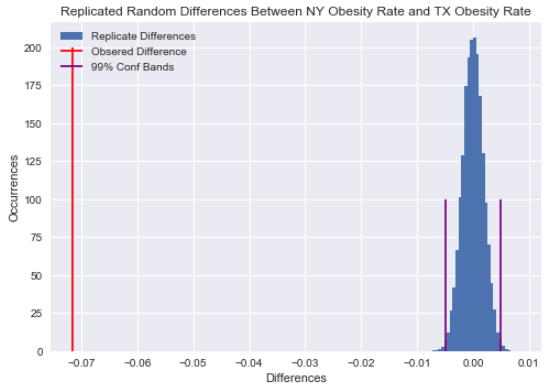
Other series I tested this on include low sleep and population count. We saw a similar outcome (p value very close to 0) for low reported sleep but see a slightly different correlation against population.

There is a slightly negative correlation between population count and obesity of -0.139. The p value is again very close to 0 and this metric may not prove important in our model predicting obesity rate.

# Obesity Rate & Its Relationship to Geography

I also questioned how different geographies may affect the obesity rate. There is speak of “people from x are more obese than those from y” and this seems like a great data set to test this.

I tested the states of TX, CA, and CT against the state of NY to see if there were any differences. The null hypothesis was “There is no difference between the difference of mean obesity rates in New York and California. The test statistic was the mean obesity rate among census tracts of NY less the mean obesity rate among census tracts of the other states.

In Texas, I observed more than 700 basis point difference in the difference of means. The 99 % confidence bands were (-0.005, 0.005) when assuming no difference in obesity rates. We can reject the null hypothesis in this case that there was no difference between obesity rates means and accept that there is a difference between these geographies. In Connecticut, we see a similar outcome.

When looking at California, our observed difference was 190 basis points with 99% confidence bands between (-0.004,0.004). Again we can reject the null hypothesis that there is no difference between obesity rates in California and New York, although the obesity rate of CA is a lot closer to NY than TX.

With this said, there does seem to be differences in obesity rates in different states and states should be worked into the prediction model.

# Summary

By using inferential stats, we are able to prove out relationships between various series in our data. This gives us a direction to look in while building our model and should prove to help guide the overall narrative of our Causes of Obesity story.