Analysis Challenge 5

Due Wednesday, March 20

Background

Failure to understand ***regression to the mean*** can lead to all kinds of erroneous inferences with data. Election fraud is one of the more pernicious kinds of bad inferences that we can conceive of because, as we've witnessed, the belief that the 2020 US Presidential election was stolen from the Republican incubment, Donald J. Trump, culminated in political violence on January, 6 2021.

Eggers, Garro, and Grimmer (2021) conduct a series of rigorous statistical tests using data from the 2020 election to demonstrate that, if fraud did occur in 2020, the evidence for it is slim at best. One claim in particular that they tested comes from a YouTube video by Shiva Ayyaduri with over 1 million views that claims to find statistical evidence that voting machines in Michigan switched people's votes from Trump to Biden. The analysis compared the difference in Trump's straight ticket to split ticket ballot share per precinct to Trump's straight ticket ballot share. Ayyaduri claimed that, in the absence of fraud, there is no good reason to expect Trump's performance among straight ticket voters to differ systematically compared to his performance among split ticket voters *in the same precinct*. The way he tests this is to take the difference in vote shares and regress it on straight ticket vote shares:

$$\text{Split}_i - \text{Straight}_i = \beta_0 + \beta_1 \text{Straight}_i + \epsilon_i$$

Ayyaduri claims that in the absence of fraud, we should expect $\beta_1 = 0$. Essentially, he proposes that we should fail to reject the null hypothesis.

However, this isn't what Ayyaduri finds. Instead, the estimate for β_1 is negative and statistically significant, meaning he can reject the null hypothesis. For Ayyaduri, this constitutes evidence of illegal vote switching.

To the average viewer, this evidence might seem convincing, *but not to you!* As a budding data analyst, you understand the concept of *regression to the mean*. From the study by Eggers and colleagues we know that a important statistical fact about cases where we observe regression to the mean is that if you do a regression just like the one Ayyaduri performs, you should expect to get an estimate less than zero. Why? Because of regression to the mean, outlier events in one period of time or for one kind of ballot are more likely to be closer to the average in the next time period or for another kind of ballot. A by-product of this phenomenon is that if you take past values and use them to predict future values (or if you use outcomes with one ballot to predict outcomes for another kind of ballot), the slope of the regression line will be less than 1. By extension, that also means that in a regression model like Ayyaduri's, the regression slope will be less than 0.

Here's a simple simulation to demonstrate. The below code generates some data that demonstrates regression to the mean. It then estimates a regression model similar to the one used by Ayyaduri. Sure enough, we get a negative statistically significant regression slope.

```
## packages
library(tidyverse)
library(estimatr)
options(digits = 2)

## fake data
tibble(
  u = rnorm(1000),
  x = u + rnorm(1000),
  y = u + rnorm(1000)
) -> sim_data

## check for regression to the mean
lm_robust(
  y - x ~ x,
  data = sim_data,
  se_type = "stata"
)
```

```
            Estimate Std. Error t value Pr(>|t|) CI Lower CI Upper  DF
(Intercept)  -0.0086      0.039   -0.22  8.3e-01   -0.085    0.068 998
x            -0.5016      0.027  -18.38  3.5e-65   -0.555   -0.448 998
```

The Challenge

You can use this approach to see whether there is evidence of regression to the mean in a range of data. For this analysis challenge, I want you to use the `county_data` we used in class to see if there is evidence of regression to the mean for Hillary Clinton's vote shares in 2016 compared to Obama's vote shares in 2012. Here's some code to help you get started:

```r
## open {tidyverse}
library(tidyverse)
library(estimatr)

## get county-level voting data
socviz::county_data |>
  drop_na() -> Data
```

To demonstrate regression to the mean:

1. Produce a scatter plot with a regression slope.
2. Perform a regression analysis similar to example in the previous section.
3. Based on the model output and the data visualization, answer whether you have evidence of regression to the mean.

Submit your work as a rendered Quarto document and submit to Canvas.