Blarry Wang
Rebecca U. Thorpe
December 13, 2019

Homework 4 Report

Developing a Research Question

Question 2 - Democracy Data RQ

The democracy data provides a great resource to test the factors that impact the level of civil liberty. Specifically, why do come countries have greater level civil liberty while some don't? One of the possible explanation is the economy — a greater GDP per worker (GDPW) implies the country prospers as a whole and has a strong middle class. After all, no activism can take place when the workers constantly worry about keeping the family from starving and live paycheck to paycheck. Prosperity and a wealthy middle class makes it possible that when rights and liberties are taken away by the government, people can take it to the streets and/or go out to vote these politicians out of office. Therefore, we can expect GDPW to be a positive factor impacting civil liberty.

Formally, the hypothesis is as follows.

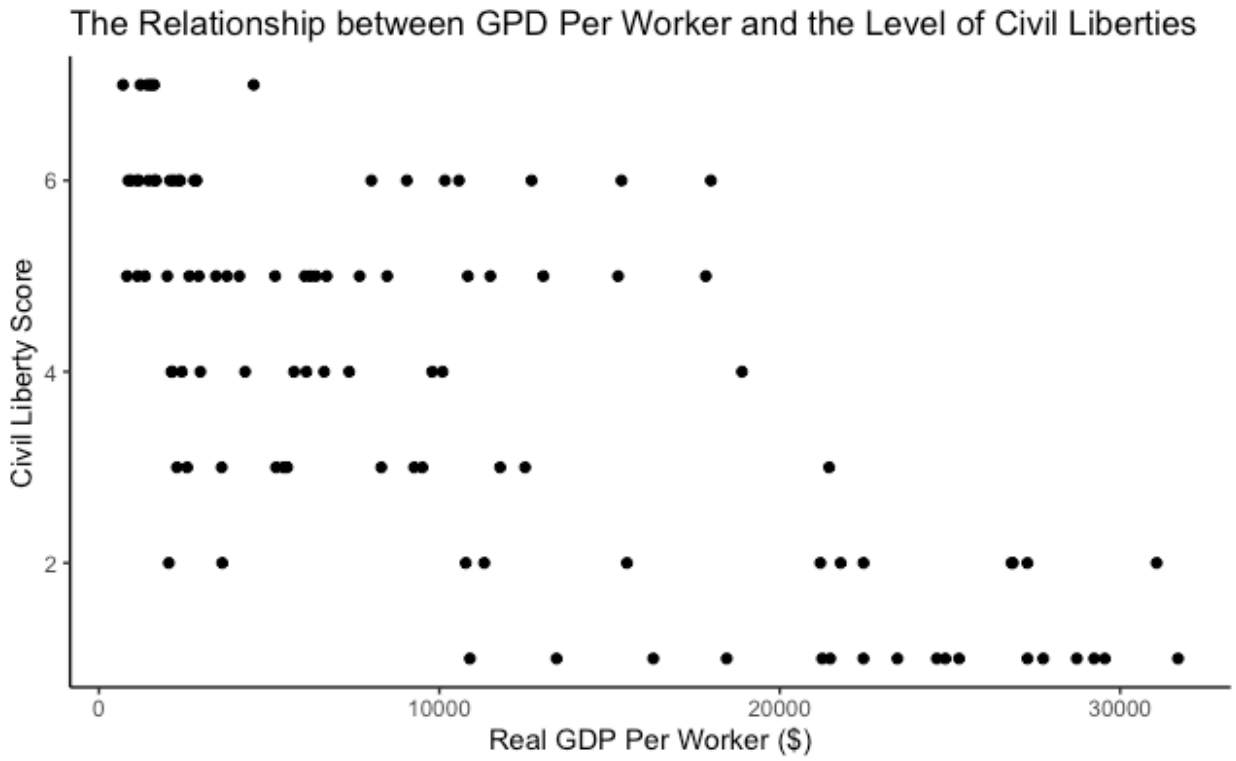
H1: There is a positive relationship between the GDP per worker and the level of civil liberties.

The null hypothesis would be fail to find a significant relation between the two variables. The p-value cutoff is set to $p \leq .05$ to test for statistical significance.

Of course, GDPW isn't the only factor that impact the level of civil liberty. To control for the hypothesis, there are some additional factors we should consider. 1) DICTATOR (whether the country is under regime). To prolong a dictator' regime, the government suppresses certain civil liberties such as freedom of expression, freedom of religion, and freedom of the press, because all of these civil liberties limits the dictator's control over the country. 2) BRITCOL (1 if former British colony; 0 if not). British influence may have an impact on civil liberty. While not a country, Hong Kong is an example that demonstrates how the British control brought to democratic traditions a colony. 3) EDT (years of education of the average member of the labor force). To put it simply, a more educated workforce are less likely to tolerate their civil liberties

being taken away. They are the backbone of the country's economy. Their power of disruption gives them a great leverage to protect civil rights and liberties.

Question 3 - Plots



Question 4 - Multivariate Regression

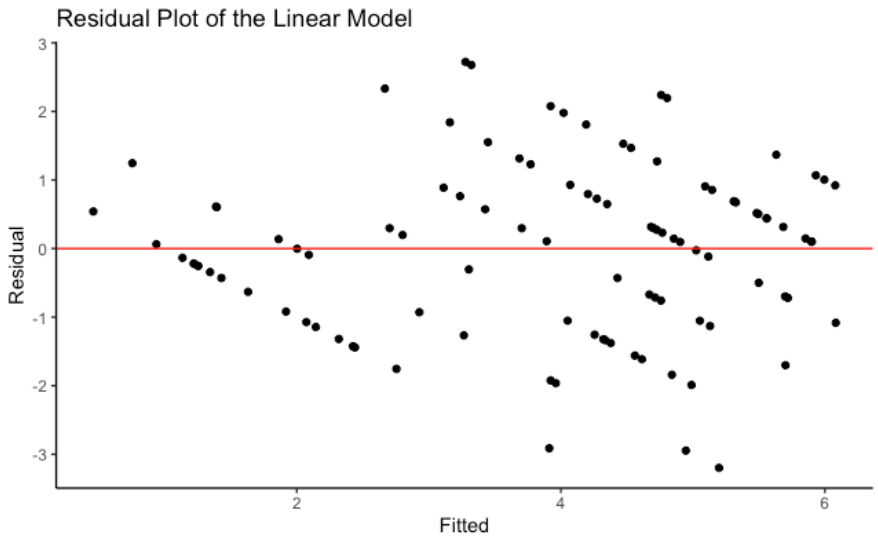
The model with 2 control variables (BRITCOL & EDT) showed a signifiant negative relationship between GDPW and CIVLIB ($p < .05$). Therefore we must reject our hypothesis because our model showed the opposite relation is true and significant. Not only this is the case, all the control variables showed significant negative relationship. The R^2 value is 0.5749. This means 57.49% of the variation is explained by the independent and control variables.

Question 5 - OSL Assumptions

1. Relationship between DV and IV is linear

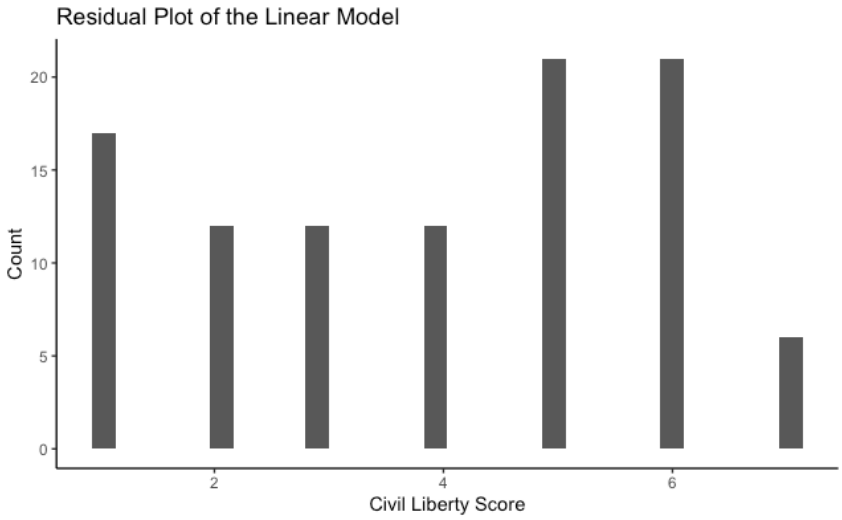
Referring to the scatter plot of GDPW and CIVLIB, the relationship is somewhat linear, because we can see a general downward trend of the association.

2. Homoscedastic: residuals are equal across length of regression line



The residual plot showed somewhat normally distributed error near the regression line. This fits the assumption of the model.

3. Normally distributed data in your dependent variable



The scatter plot showed that the normal distribution assumption is violated. Instead of a normal distribution, the data is binomial. There is a lot more countries with low and high civil liberty scores, but fewer in the middle.

4. Independent and control variables aren't highly correlated

To test for independent variable's correlation with the control variables, I calculated their correlations and found that there is small correlation between GPDW and BRITCOL, and between EDT and BRITCOL. However, there is a correlation score of 0.8268 between GDPW and EDT. This violates the assumption made for OLS.

Correlation Matrix for Independent and Control Variables

	GDPW	BRITCOL	EDT
GDPW	1.0000	-0.2173	0.8268
BRITCOL	-0.2173	1.0000	-0.1158
EDT	0.8268	-0.1158	1.0000

5. No major outliers in the data

According to Cook's distances, there are 5 outliers in our data. This should be acceptable since we have over 100 samples.