Corruption and Parking Violations

Akkineni, Hanna, Thorp

September 26, 2016

```r
setwd("C:/Users/kevin/OneDrive/School/MIDS/W203 - Statistics for Data Science/Lab 1/W203_lab1_corruptio
#library(car)
#library(grid)
library(ggplot2)
library(knitr)
library(kableExtra)

load("Corrupt.Rdata")

## Correct Data Problems
#Fix majoritymulsim where value = -1, should be 0
FMcorrupt[FMcorrupt$majoritymuslim == -1 & ! is.na(FMcorrupt$majoritymuslim), "majoritymuslim"] = 0

# Add missing counties.  Reference: https://www.worldatlas.com/aatlas/ctycodes.htm
FMcorrupt[FMcorrupt$wbcode == "ARE", "country"] = "UNITED ARAB EMIRATES"
FMcorrupt[FMcorrupt$wbcode == "CAF", "country"] = "CENTRAL AFRICAN REPUBLIC"
FMcorrupt[FMcorrupt$wbcode == "CAN", "country"] = "CANADA"
FMcorrupt[FMcorrupt$wbcode == "COL", "country"] = "COLUMBIA"
FMcorrupt[FMcorrupt$wbcode == "ECU", "country"] = "ECUADOR"
FMcorrupt[FMcorrupt$wbcode == "JAM", "country"] = "JAMAICA"
FMcorrupt[FMcorrupt$wbcode == "LVA", "country"] = "LATVIA"
FMcorrupt[FMcorrupt$wbcode == "NOR", "country"] = "NORWAY"
FMcorrupt[FMcorrupt$wbcode == "PAN", "country"] = "PANAMA"
FMcorrupt[FMcorrupt$wbcode == "SWE", "country"] = "SWEDEN"
FMcorrupt[FMcorrupt$wbcode == "TUR", "country"] = "TURKEY"

# Create named regions variable using region
FMcorrupt$region_name = NA
FMcorrupt[FMcorrupt$region == 1 & ! is.na(FMcorrupt$region), "region_name"] = "Caribbean"
FMcorrupt[FMcorrupt$region == 2 & ! is.na(FMcorrupt$region), "region_name"] = "South America"
FMcorrupt[FMcorrupt$region == 3 & ! is.na(FMcorrupt$region), "region_name"] = "Europe"
FMcorrupt[FMcorrupt$region == 4 & ! is.na(FMcorrupt$region), "region_name"] = "Asia" # "South Asia"
FMcorrupt[FMcorrupt$region == 5 & ! is.na(FMcorrupt$region), "region_name"] = "Oceania"
FMcorrupt[FMcorrupt$region == 6 & ! is.na(FMcorrupt$region), "region_name"] = "Africa"
FMcorrupt[FMcorrupt$region == 7 & ! is.na(FMcorrupt$region), "region_name"] = "Middle East" # "Western

FMcorrupt$region_name = factor(FMcorrupt$region_name)

# Remove 66 rows that do not have relevant data to the key analyses
corrupt = subset(FMcorrupt, !is.na(violations) & !is.na(mission) & !is.na(staff) )


# split data in to pre and post, before and after enforcement changes
cor_pre = subset(corrupt, prepost == "pre")
cor_pos = subset(corrupt, prepost == "pos")

# Merge both the above to one line with pre and pos appeneded to variable names (prepos removed)
```

```
cor_oneline = merge(cor_pre, cor_pos, by = "wbcode", suffixes = c(".pre", ".pos"))

# Grab only the variables that are needed.
cor_oneline = cor_oneline[, c("wbcode", "violations.pre",  "violations.pos", "fines.pre", "fines.pos",
                              "mission.pre", "staff.pre", "spouse.pre", "gov_wage_gdp.pre", "pctmuslim.p
                              "cars_total.pre", "cars_mission.pre", "pop1998.pre", "gdppcus1998.pre", "
                              "r_africa.pre", "r_middleeast.pre", "r_europe.pre", "r_southamerica.pre",
                              "country.pre", "distUNplz.pre",
                              "region.pre", "region_name.pre"
                              )]

# Remove suffix where not needed.
colnames(cor_oneline) =  c("wbcode", "violations.pre",  "violations.pos", "fines.pre", "fines.pos",
                              "mission", "staff", "spouse", "gov_wage_gdp", "pctmuslim", "majoritymuslin
                              "cars_total", "cars_mission", "pop1998", "gdppcus1998", "ecaid", "milaid"
                              "r_africa", "r_middleeast", "r_europe", "r_southamerica", "r_asia",
                              "country", "distUNplz",
                              "region", "region_name"
                              )

# Rename FMcorrupt to ensure we don't use it accidentally
cor_nas = FMcorrupt
remove(FMcorrupt)
```

Introduction

Research Question

Prior to 2002 diplomats at UN missions were exampt from parking violations and fines in New York City, by virtue of their diplomatic immunity. There was wide variation in diplomats' willingness to adhere to local parking laws. This analysis attempts to understand whether the variation in adherance to local parking law was related to cultural norms in the diplomats' home countries. For the period prior to 2002, we examine the relationship between perceptions of corruption in that country and that countries' diplomats' willingness to incur parking violations. In 2002 NYC parking enforcement acquired the right to confiscate license plates from vehicles belonging to foreign diplomats if they had accumulated unpaid parking violations, thus making payment an function of both cultral norms and legal enforcment. This had a notable compressing effect on parking law adherence.

Question: Does an index of perceived corruption in the diplomats' home country have explanatory power for a given diplomatic mission's compliance with local parking regulations?

Description of Dataset

Our data set has a total of 364 observations. Of the 364, 66 observations contain only economic data leaving NA for our dependent variable, violations. Considering the countries that are among these 66 and the variables for which they have valid data, we suspect these rows result from a merge of economic data with the violations data ($econmic \cup violations$). As such, we believe these 66 countries represent a data artefact from that data merge. As these observations do not contain valid values for key variables, we remove them from our dataset. Of note, those 66 observations appear to contain many which do not even have a mission or staff in New York City, and as such are not relevant for this study on diplomatic parking violations in New York City. Given these considerations, we feel comfortable that we are not biasing the results of the study by removing these observations. With those 66 rows removed, we're left with 298 observations (two observations for each of 149 countries where corruption data exists.) Each country has one observation from prior to the 2002 regulation change and one observation from after. Only the 'violations' and 'fines' variables differ between the two observations for a given country, while other variables remain constant.

Univariate Analysis of Key Variables

Key Variables:

- violations
- corruption
- staff
- trade (maybe?)

There are 66 observations that contain only economic data leaving NA for our dependent variable, violations. Considering the countries that are among these 66 we suspect these rows result from a merge of economic data with the violations data ($econmic \cup violations$) and the economic data set had countries that did not have embasies in Manhattan. Removing these superfluous observations results in the intersection of the two data sets ($economic \cap violations$), which is what we desire. With those 66 rows removed, we're left with 298 observations (two observations for each of 149 countries where corruption data exists.)

- show plots and summary data for key variables
  - Summary table
  - Frequency Distribution

```
# Use CSV version of Google Sheet 'Variable Description for Introduction': https://docs.google.com/spre
variable_description = read.csv("Lab 1 - Variable Descriptions for Introduction.csv", header = TRUE, sep
#summary(variable_description)

kable(variable_description, "latex", longtable = TRUE, booktabs = TRUE, caption = "Data Set Variables")
  kable_styling(full_width = TRUE, latex_options = c("HOLD_position", "striped", "repeat_header"), row_
```

Table 1: Data Set Variables

Variable	Description	Observations	Alterations
wbcode	Three Letter Country code		
violations.pre	The number of violations accumulated before enforcement changes.	Values have 6 decimal places, we don't quite understand why, however the instructions we were given with the data set suggest these values should be integer, so we will treat it as the sum.	
violations.pos	The number of violations accumulated after enforcement changes.	Values have 6 decimal places, we don't quite understand why, however the instructions we were given with the data set suggest these values should be integer, so we will treat it as the sum.	

Table 1: Data Set Variables *(continued)*

Variable	Description	Observations	Alterations
fines.pre	Summed cost of violations.pre adjusted for inflation.		
fines.pos	Summed cost of violations.pos adjusted for inflation.		
mission	If country has mission in NYC	All values are true.	
staff	# of employees		
spouse	# of spouses		
gov_wage_gdp			
pctmuslim	Percent of country that identifies as Muslim	2 countries have NA's: Bosnia-Herzegovina and Zaire	
majoritymuslim	Is country majority Muslim	This variable had some values of -1 which occured if and only if pctmuslim was 0.	We replaced all -1's with -0's.
trade	Total trade with the US in 1998 USD	2 countries have NA's: Bosnia-Herzegovina and Zaire, 1 country shows 0 Libia	
cars_total	The number of cars owned by both mission and the employees.	Total of 1455, 10 NA's	
cars_personal	Cars owned by staff and spouses of mission	Total of 740, 10 NA's	
cars_mission	Cars owned by mission	Total of 715, 10 NA's	
pop1998	Country's population in 1998		
gdppcus1998	GDP Percent in USD 1998(?)		
ecaid	US Economic Assistance grants and loans		
milaid	Military Aid		
region	7 Geographic regions labelled 1-7		
region_name	7 Geographic regions named		We created this variable using region variable above
corruption	Country corruption index	Min. = -2.58299 (Least corrupt), Max. = 1.58281 (Most corrupt)	
totaid	milaid + ecaid		
r_africa	Boolean if part of continent. Some countries belong to none, and some others have NA's.		
r_middleeast	"		
r_europe	"		

Table 1: Data Set Variables *(continued)*

Variable	Description	Observations	Alterations
r_southamerica	"		
r_asia	"		
country	country name	Some missing values, though wbcode are available.	Missing values were filled using wbcode variable.
distUNplz			

```
remove(variable_description)
```

```
# This is assuming staff is better than cars, staff+spouse, total_cars.  Though I don't know if that's

### Probably makes sense tomove this above
round_df <- function(x, digits) {
    # round all numeric variables
    # x: data frame
    # digits: number of digits to round
    numeric_columns <- sapply(x, mode) == 'numeric'
    x[numeric_columns] <-  round(x[numeric_columns], digits)
    x
}

summary_table_output = cor_oneline[, c("country", "staff", "violations.pre", "violations.pos", "corrupt
summary_table_output$mean_violations_per_staff.pre = summary_table_output$violations.pre/summary_table_
summary_table_output$mean_violations_per_staff.pos = summary_table_output$violations.pos/summary_table_

tmp_rounded = round_df(summary_table_output[, c("country", "mean_violations_per_staff.pre", "mean_viola
tmp_rounded = tmp_rounded[order(tmp_rounded$mean_violations_per_staff.pre, decreasing = TRUE), ]
#TODO Sort, Round

kable(tmp_rounded[1:20, ],
      "latex", longtable = TRUE, booktabs = TRUE,
      caption = "Top 20 Countries by Parking Violations (Key Variables)",
      col.names = c("Country", "Mean Violations per Staff Before 2002 Change", "Mean Violations per Sta
  kable_styling(full_width = TRUE, latex_options = c("HOLD_position", "striped", "repeat_header"), row_
```

Table 2: Top 20 Countries by Parking Violations (Key Variables)

	Country	Mean Violations per Staff Before 2002 Change	Mean Violations per Staff After 2002 Change	Corruption Index
77	KUWAIT	249.36	0.15	-1.07
41	EGYPT	141.37	0.33	0.25
129	CHAD	125.89	0.00	0.84
119	SUDAN	120.58	0.37	0.75
14	BULGARIA	119.03	1.64	0.50
93	MOZAMBIQUE	112.13	0.07	0.77
2	ALBANIA	85.54	1.85	0.92
1	ANGOLA	82.71	1.71	1.05
120	SENEGAL	80.21	0.21	0.45

Table 2: Top 20 Countries by Parking Violations (Key Variables)
(continued)

	Country	Mean Violations per Staff Before 2002 Change	Mean Violations per Staff After 2002 Change	Corruption Index
107	PAKISTAN	70.29	1.21	0.76
27	IVORY COAST	67.96	0.46	0.35
148	ZAMBIA	61.17	0.15	0.56
86	MOROCCO	60.77	0.40	0.10
45	ETHIOPIA	60.44	0.62	0.25
100	NIGERIA	59.40	0.44	1.01
128	SYRIA	53.32	1.36	0.58
11	BENIN	50.41	6.50	0.76
149	ZIMBABWE	46.15	0.86	0.13
28	CAMEROON	44.11	2.86	1.11
145	MONTENEGRO & SERBIA	38.52	0.05	0.97

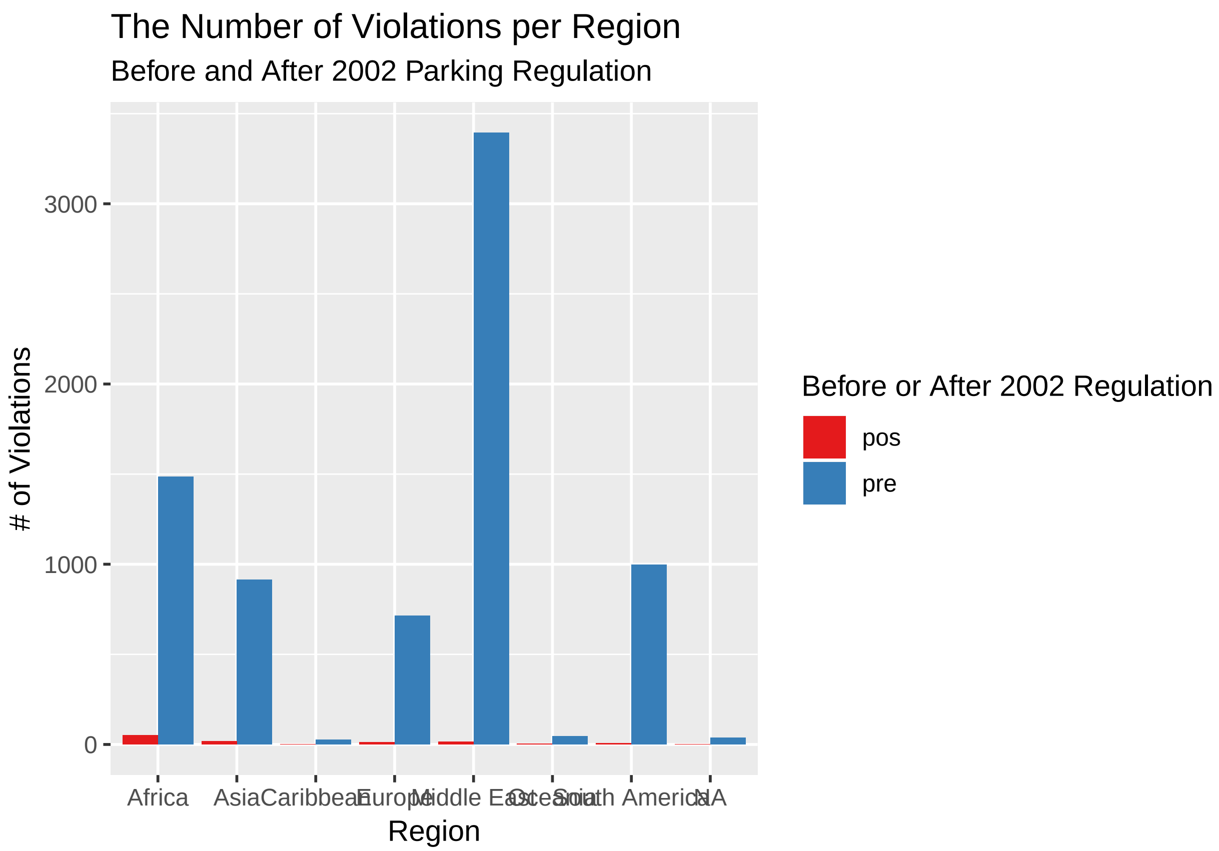
```
remove(summary_table_output, tmp_rounded)
```

Analysis of Key Relationships

Here we turn to analysis of key relationships among the variables. For the exploratory phase, we are interested in determining which of the variables are correlated, and if so, how strongly. For this analysis, where applicable, and based on the discussion in the Univariate section above, we use the log transformation of variables, as opposed to the untransformed version of the variable(s). This helps improve our understanding of relationships as positive skew that is present in some of the variables is compressed.

1) What is the relationship between violations before and after the 2002 introduction of the new parking regulation?

```
correlation_matrix_input = cor_oneline[, c("corruption", "violations.pre", "fines.pre", "violations.pos
                                                        "staff", "spouse", "majoritymuslim", "pctmuslim
                                                        )]

p <- ggplot(corrupt, aes(factor(region_name), violations, fill = factor(prepost))) +
  geom_bar(stat="identity", position = "dodge") +
  scale_fill_brewer(palette = "Set1")
p + labs(title = 'The Number of Violations per Region', subtitle = 'Before and After 2002 Parking Regul
```

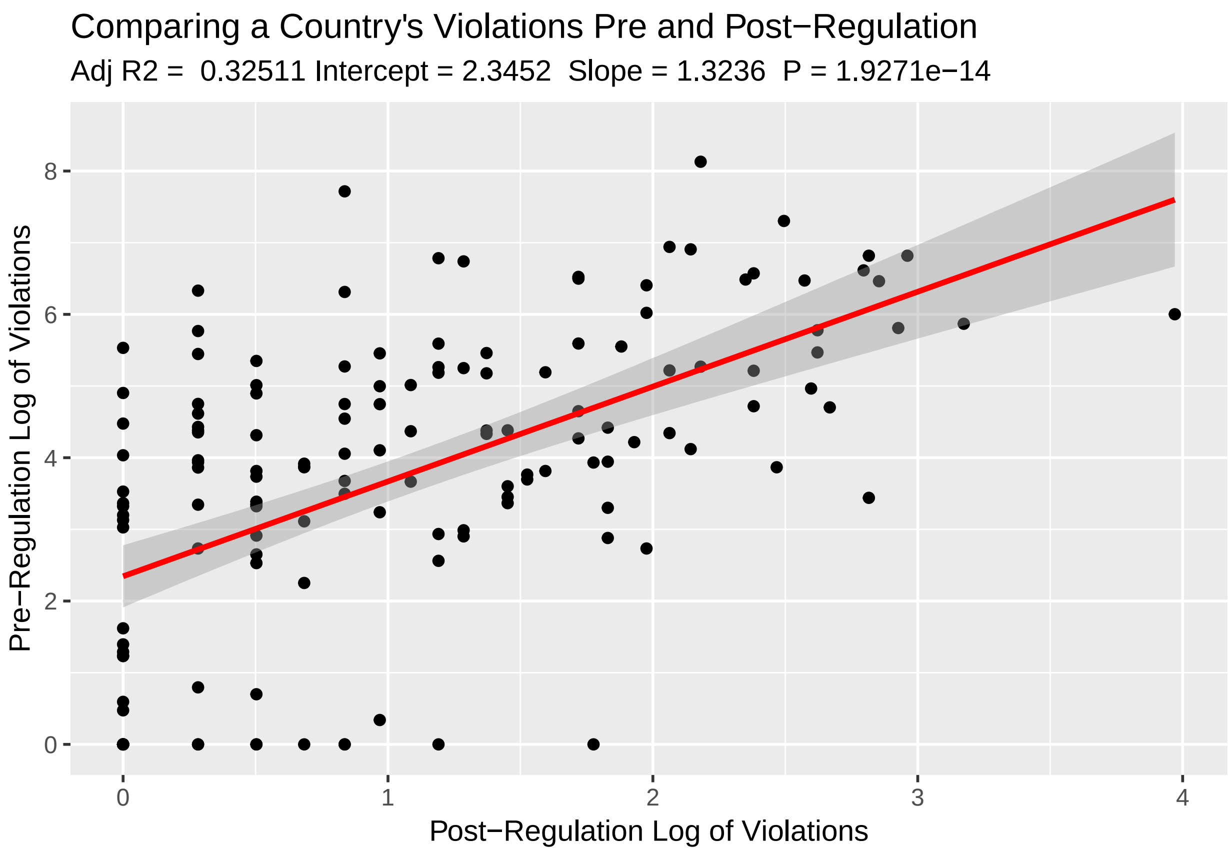


This shows us that: There is a dramatic difference between the number of violations before and after the 2002 regulation for all regions. (Regions are used here as an expedient bucketing method across the observations to reduce complexity. This is useful for explanation, though not necessarily useful for model-building.)

However, it is difficult to see the rate-of-change relationship between the two variables given their different scales. This relationship can be explored by taking a log transformation of both variables .

```
#add log transforms
#(with the addition of 1 to the log transform to circumvent issues with the 0 values)
correlation_matrix_input$violations.pos.log = log(correlation_matrix_input$violations.pos+1)
correlation_matrix_input$violations.pre.log = log(correlation_matrix_input$violations.pre+1)

#plot the relationship
ggplotRegression <- function (fit, title, x, y) {

  require(ggplot2)

  ggplot(fit$model, aes_string(x = names(fit$model)[2], y = names(fit$model)[1])) +
    geom_point() +
    stat_smooth(method = "lm", col = "red") +
    labs(title = title, subtitle = paste("Adj R2 = ",signif(summary(fit)$adj.r.squared, 5),
                       "Intercept =",signif(fit$coef[[1]],5 ),
                       " Slope =",signif(fit$coef[[2]], 5),
                       " P =",signif(summary(fit)$coef[2,4], 5)),
                        x = x,
                        y = y)
}
```

```
lm1 <- lm(violations.pre.log ~ violations.pos.log, data=correlation_matrix_input)
ggplotRegression(lm1, 'Comparing a Country\'s Violations Pre and Post-Regulation', 'Post-Regulation Log
```



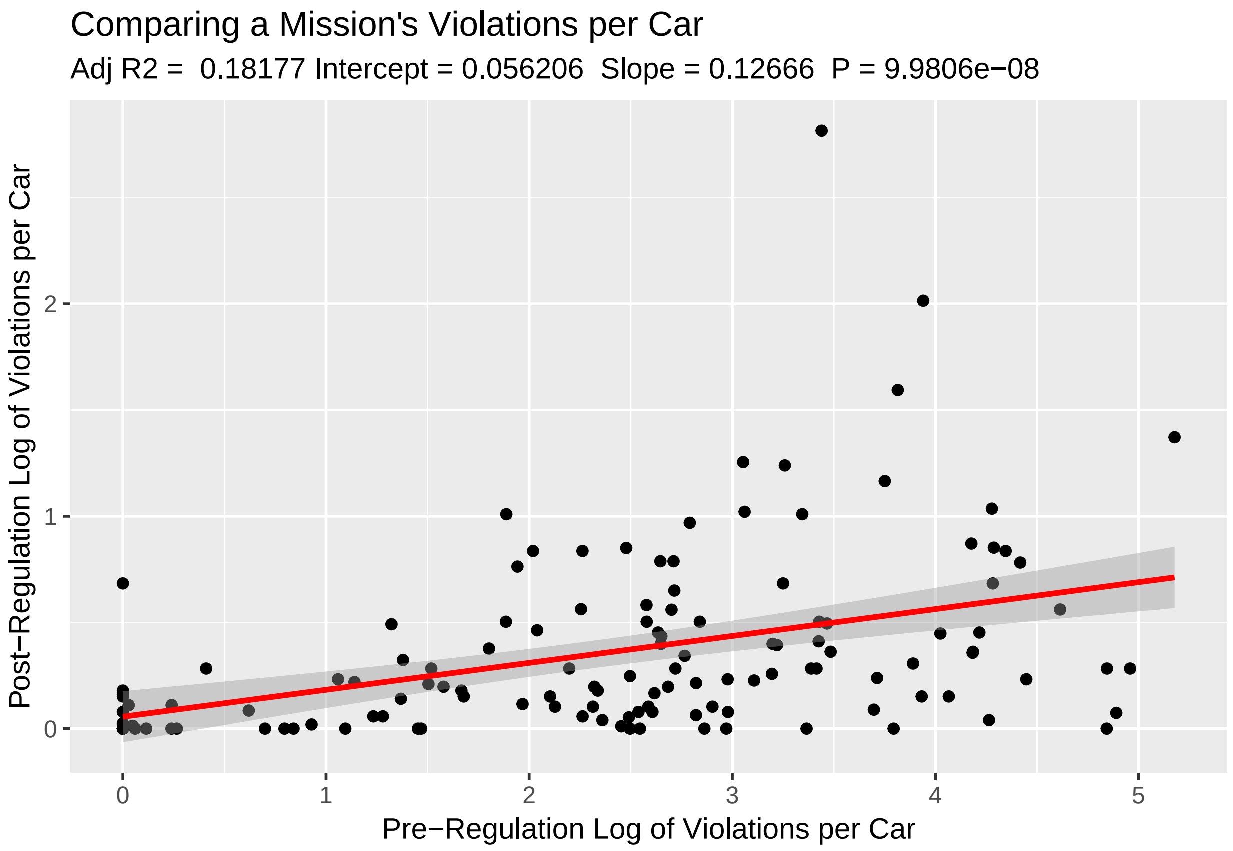
This plot roughly shows that for a 1 percent increase in the number of violations before the regulation, the data reveals a roughly 0.24 percent increase in the number of violations seen after the regulation. At this exploratory stage, there are two important facets to takeaway: 1) The correlation between these two series at 0.33 is notable, 2) The relationship between the two variables is clearly positive.

The interpretation of this is that in general countries with higher amounts of violations before the regulation are likely to also have relatively higher amounts of violations after.

However, there is likely much more going on in this data generation process. At this early stage, we speculate that violations is likely related to the size of a mission. This facet was examined through the car-based and the people-based variables. We found that the strongest relationship (consdiering both correlation and slope) between pre- and post-regulation percent changes in violations was for the number of violations normalized for the number of cars. As such, we will be using this during the remainder of our exploration. Should we be beyond the exploration phase into a phase where we were trying to determine causation or to predict violations, this would be insufficient, but given that we are still in exploration phase, it will suffice. Alternate formulations that were explored to control for size (staff members, total people, staff cars) are included in the Appendix.

```
# add violations treated with total cars
correlation_matrix_input$violations_weighted.cars_total.pre = correlation_matrix_input$violations.pre/c
correlation_matrix_input$violations_weighted.cars_total.pos = correlation_matrix_input$violations.pos/c

#add log transformations
#The addition of 1 to the normalized number of violations would not be acceptable for predictive or cau
# will help us circumvent the 0 - 1 log transformation issues while maintaining the sequence of observa
```

```
correlation_matrix_input$violations_weighted.cars_total.pre.log = log(correlation_matrix_input$violatio
correlation_matrix_input$violations_weighted.cars_total.pos.log = log(correlation_matrix_input$violatio

lm5 <- lm(violations_weighted.cars_total.pos.log~violations_weighted.cars_total.pre.log, data=correlati
ggplotRegression(lm5,'Comparing a Mission\'s Violations per Car','Pre-Regulation Log of Violations per 
```

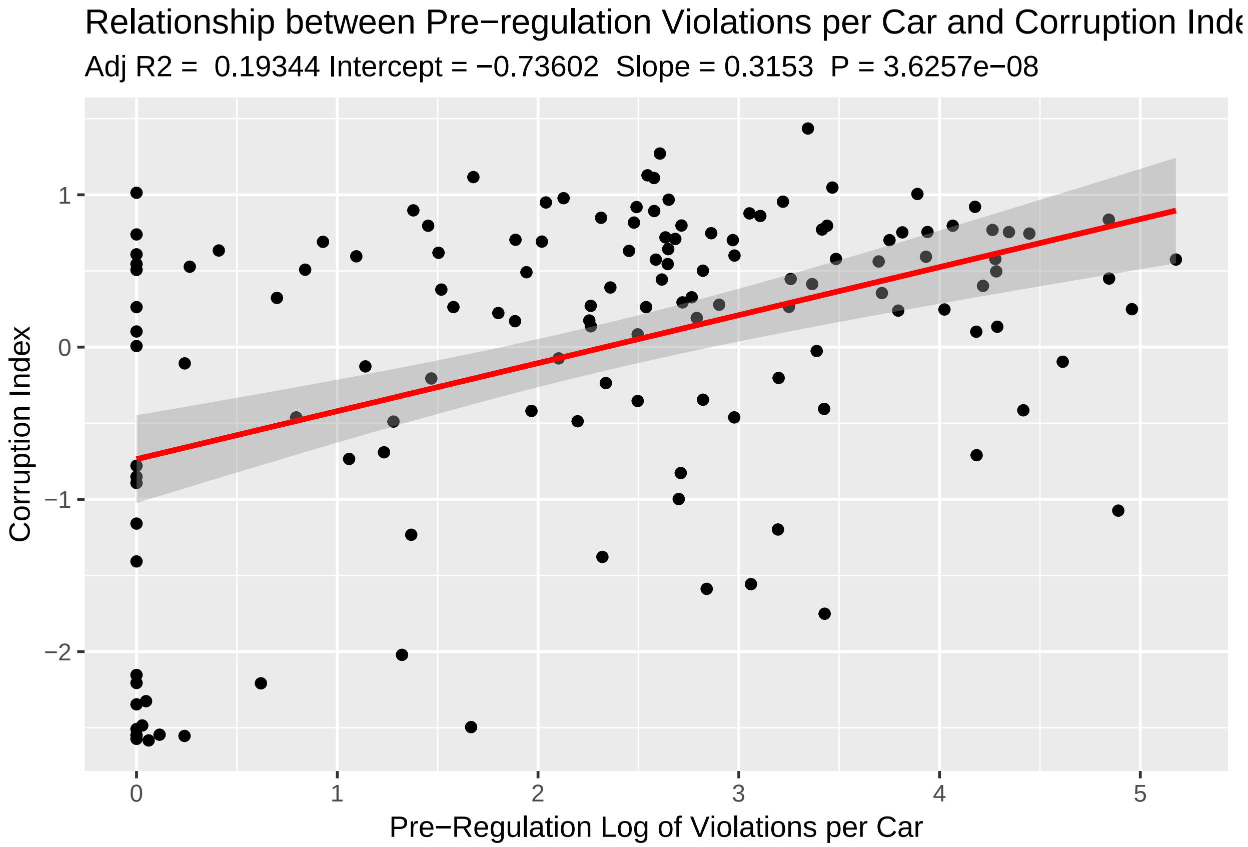


From this chart, we can again take away a sense that those countries which had higher violations prior to the regulations were also likely to have relatively higher violations after the regulation, even after controlling for the conflating of size impacts by normalizing violations with the most appropriate proxy for the mission size-type variables. And, importantly, this formulation of size normalization appears to offer the best preservation of signal on country differences in violations even after controlling for size. This selection also makes intuitive sense as only cars can incur parking fines - people without cars do not.

Now we turn to understanding how other variables in our data set might help explain the differences in violations per car. We will begin with an examination of the relationship between the provided corruption index and violations per car, with other variables addressed in subsequent sections.

We begin by examining the relationship between pre-regulation # of violations per car and the provided corruption index.

```
lm6 <- lm(corruption~violations_weighted.cars_total.pre.log, data=correlation_matrix_input)
ggplotRegression(lm6,'Relationship between Pre-regulation Violations per Car and Corruption Index','Pre-
```

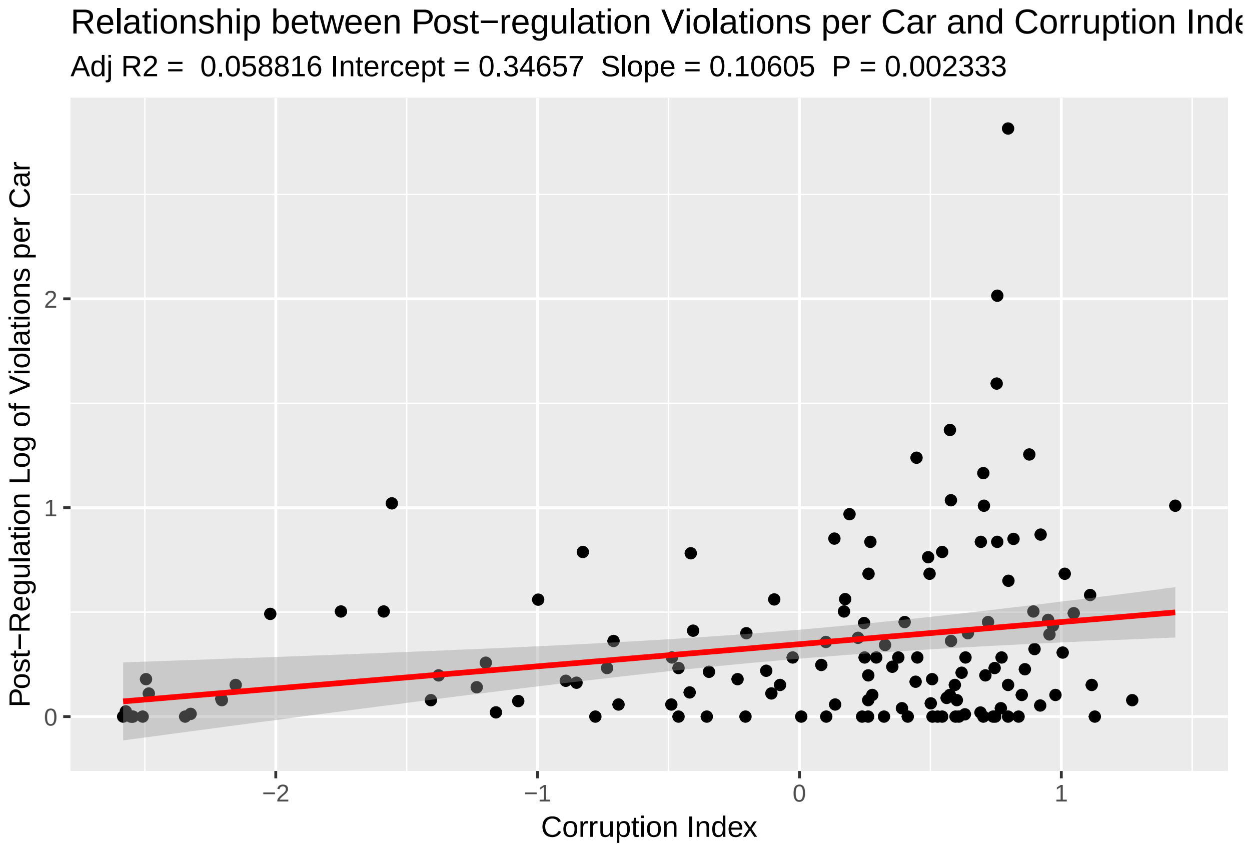


From this chart, we can see a strong positive relationship between these two variables appears clear, although there is much more going on than is characterized here as the corelationship is notable but far from linear. This chart can be interpreted as - for a 1 unit increase in the country's ranking on this corruption index, we expect to see a 63 percent increase in that country's number of parking violations per car.

This seems to confirm our initial inclination that corruption may be a good indicator of a country's willingness to incur parking violations.

Was this relationship changed by the introduction of new parking regulations?

```
lm7 <- lm(violations_weighted.cars_total.pos.log~corruption, data=correlation_matrix_input)
ggplotRegression(lm7,'Relationship between Post-regulation Violations per Car and Corruption Index','Co
```



This chart shows that after the regulation, the positive relationship between a country's corruption index and its willingness to incur parking violations still exists, however it is much weaker and the magnitude much smaller. After the introduction of the parking regulation, we expect a one unit increase in a country's corruption index to result in only an 11 percent increase in parking violations per car. Also of note, the correlation between the two variables decreased markedly, with the relationship seen here clearly positive but non-linear.

It is also important to note that, while our analysis reveals a positive relationship between a country's corruption index and their willingness to incur parking violations, both before and after the regulation, there are numerous countries with a high corruption index that also incurred 0 parking violations. So, while this relationship may help describe the population, individual countries within the population can and do deviate markedly from the statistical relationship seen above.

Analysis of Secondary Effects

- show that pctmuslim doesn't have much relevance
- analyze trade relationship
- analyze gov_wage_gdp (particular after, when fines are being paid)
  - Countries with higher violations + higher corruption index might be low gov_wage_gdp (might also mean small governemnt, so population might also be a factor)
- does region have any *meaningful* relationship to violations?

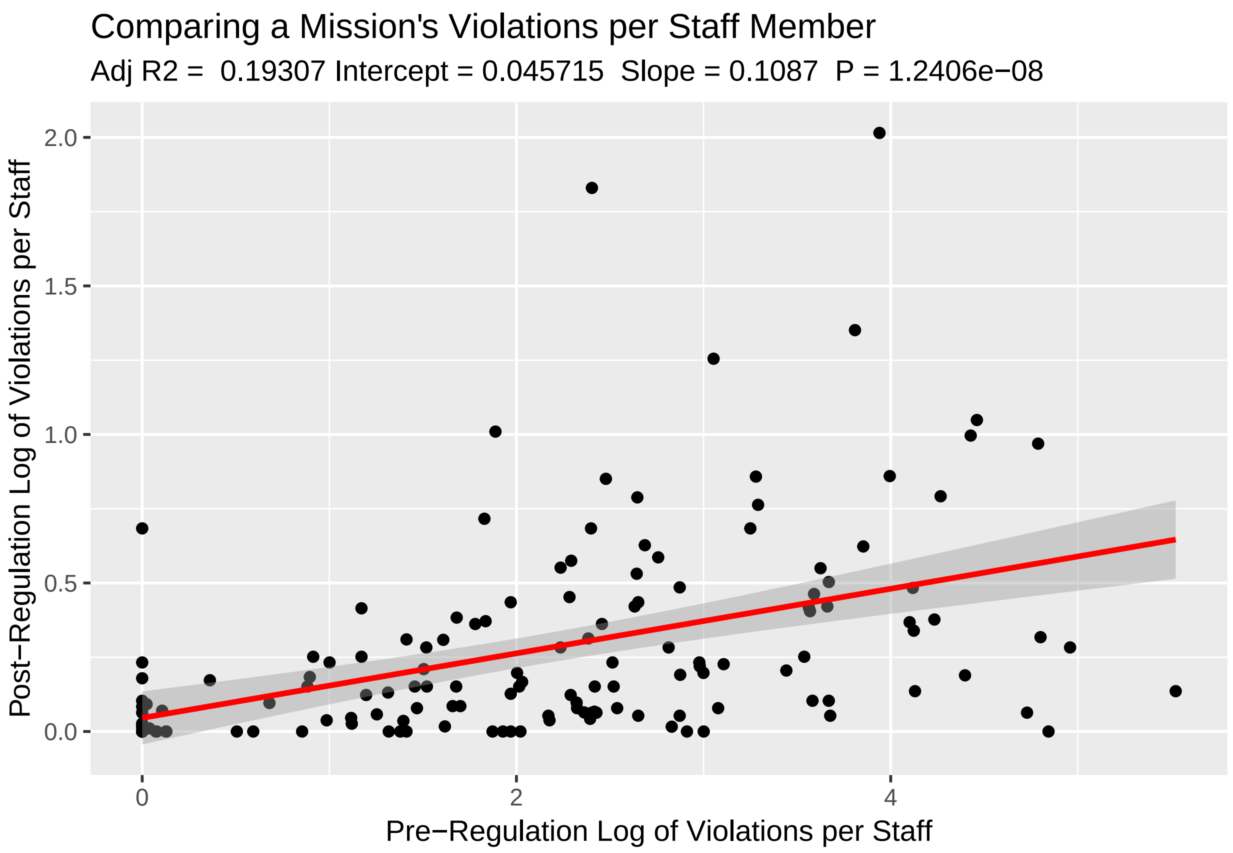
Conclusion

Some poople used to be jerks. But now they're not. Enforcing fines works on Jerks. Canadian's are not jerks (bias: author: Hanna, is a Canadian)

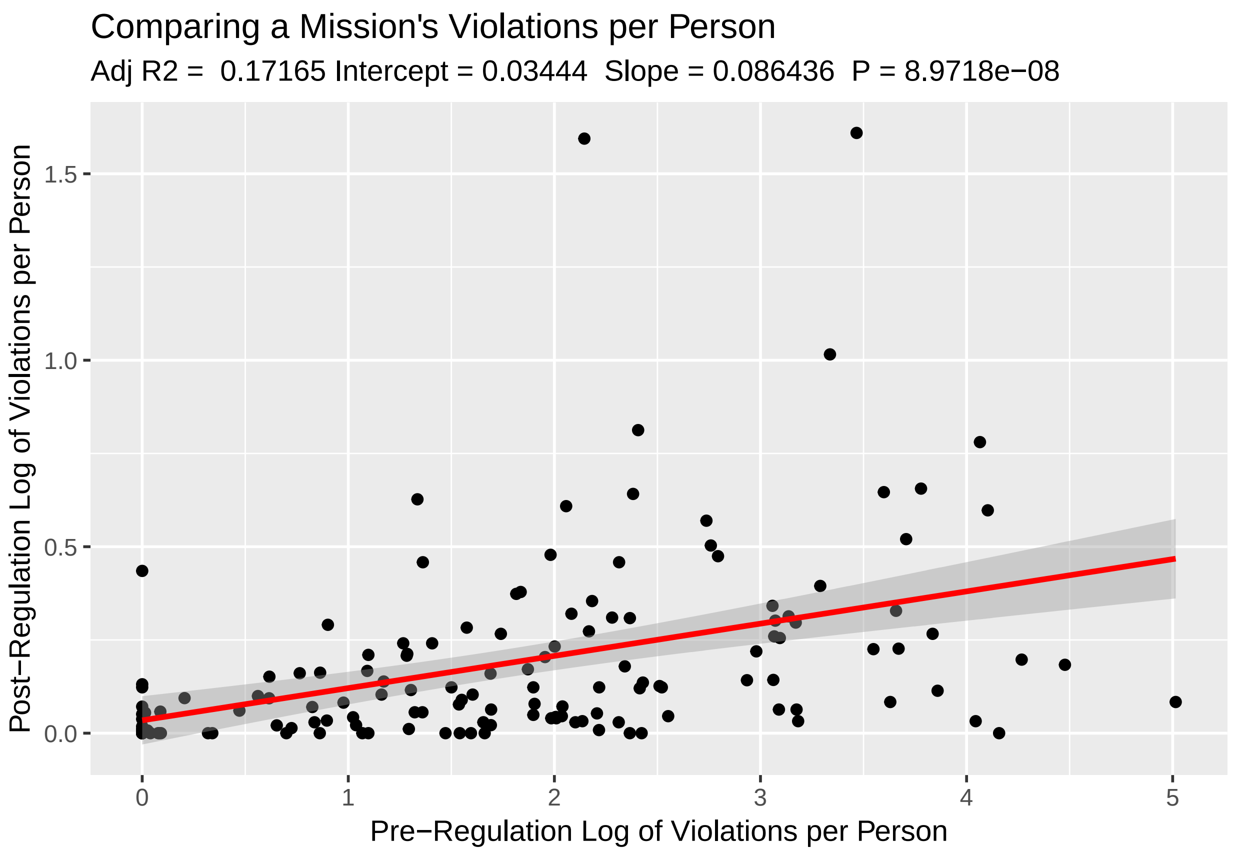
Appendix

From the Analysis of Key Relationships section. We also explored alternate means of controlling for the impact of mission size on the relationship between violations pre and post regulation. These formulations are documented here for the curious reader.

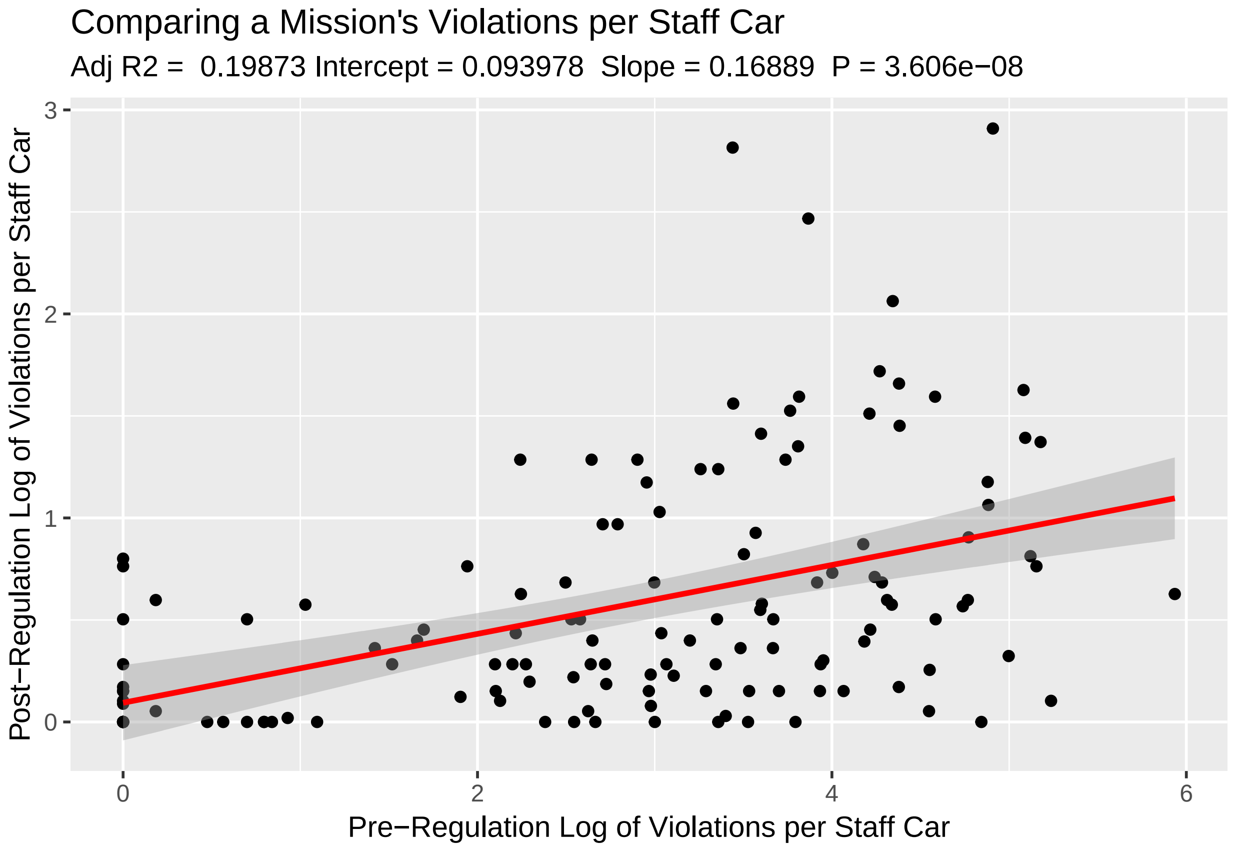
```
# add violations treated with staff
correlation_matrix_input$violations_weighted.staff.pre = correlation_matrix_input$violations.pre/correla
correlation_matrix_input$violations_weighted.staff.pos = correlation_matrix_input$violations.pos/correla

# add violations treated with total people (staff + spouse)
correlation_matrix_input$total_people = correlation_matrix_input$staff + correlation_matrix_input$spous

correlation_matrix_input$violations_weighted.total_people.pre = correlation_matrix_input$violations.pre
correlation_matrix_input$violations_weighted.total_people.pos = correlation_matrix_input$violations.pos

# add violations treated with cars_mission
correlation_matrix_input$violations_weighted.cars_mission.pre = correlation_matrix_input$violations.pre
correlation_matrix_input$violations_weighted.cars_mission.pos = correlation_matrix_input$violations.pos


#add log transformations
#The addition of 1 to the normalized number of violations would not be acceptable for predictive or cau
# will help us circumvent the 0 - 1 log transformation issues while maintaining the sequence of observa

correlation_matrix_input$violations_weighted.staff.pre.log = log(correlation_matrix_input$violations_we
correlation_matrix_input$violations_weighted.staff.pos.log = log(correlation_matrix_input$violations_we
correlation_matrix_input$violations_weighted.total_people.pre.log = log(correlation_matrix_input$violat
correlation_matrix_input$violations_weighted.total_people.pos.log = log(correlation_matrix_input$violat
correlation_matrix_input$violations_weighted.cars_mission.pre.log = log(correlation_matrix_input$violat
correlation_matrix_input$violations_weighted.cars_mission.pos.log = log(correlation_matrix_input$violat


lm2 <- lm(violations_weighted.staff.pos.log~violations_weighted.staff.pre.log, data=correlation_matrix_
ggplotRegression(lm2,'Comparing a Mission\'s Violations per Staff Member','Pre-Regulation Log of Violat
```



```
lm3 <- lm(violations_weighted.total_people.pos.log~violations_weighted.total_people.pre.log, data=correl
ggplotRegression(lm3,'Comparing a Mission\'s Violations per Person','Pre-Regulation Log of Violations pe
```



```
lm4 <- lm(violations_weighted.cars_mission.pos.log~violations_weighted.cars_mission.pre.log, data=correl
ggplotRegression(lm4,'Comparing a Mission\'s Violations per Staff Car','Pre-Regulation Log of Violation
```



It makes sense that the set of normalizations which best preserved signal from violations normalized by size-type variables was the total cars variable. The three other formulations above are potential substitutes, but appear to preserve less of the signal than the total cars formulation. As mentioned above, it makes intuitive sense for the cars variables to carry more signal than the number of staff variables as only cars can incur parking violations. Also, we considered using the number of violations per staff car variable as it also appear to preserve signal well, however the normalization method used here (violations divided by staff car) is compromised where staff cars == 0, as is the case for four observations.