Corruption and Parking Violations

Akkineni, Hanna, Thorp

September 26, 2016

```
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.
                              "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", "majoritymusli
                              "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 there were no measures New York City (NYC) was able to take to enforce payment of parking violations from UN officials who have diplomatic emmunity. Therefore, payment was largely a function of cultural norms. 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.

In our exploratory data analysis, we will describe how cultural norms affect the payment of parking violations by researching the number of violations diplomats had accumulated both before and after the enforcement changes using the corruption index of the country of the embassy they are accociated with.

How do a country's corruption and fine enforcement affect its diplomats' parking behavior?

Description of Dataset

Our data set has a total of 364 observations, there are two observation for each country, where only the number of violations and value of the fines changes, one before NYC gained the ability to enforce non-payment, and another after.

Univariate Analysis of Key Variables

Key Variables:

- violations
- corruption
- staff
- trade

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

```
# 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.	
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			

Table 1: Data Set Variables *(continued)*

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

Alex notes to be incorporated above

“”“Data quality issues:
~~1) We have several extra observations that have NaN values for key variables. These data likely reflect a merging process wherein the desired dataset regarding parking violations and fines was merged with economic data. +wbcode is impacted by this. There are 213 unique values for wbcode, whereas there are only 151 countries with pre and post-legislation observations ++++Suggestion: We drop the N/A variables and note the number of countries for which we have no data.~~

2) The violations and fines data are odd. The instructions do not provide adequate background regarding these variables. This is a particular problem because these variable are at the core of the proposed analysis. Oddities: +violations has 7 decimal places. This is odd because we would expect a positive integer (in the case of a single year) or at least a figure that is recognizable as an average of some sort (in the case of multiple years)

3) The mission, staff, and spouse variables also contain oddities. For example, the entries for HKG and PRI are 0. These are the only zeros in those variables. Looking at the values for other variables for those two observations, some other oddities emerge. For example: +HKG’s majoritymuslim (which should be a dummy variable of either 1 or 0) is -1 ++++Suggestion: Given that our research question regards the impact of the corruption on willingness to incur parking violations and fines in NYC, we probably do not care about observations without a mission in NYC. As such, taking all of these oddities and the low desirability of the data, I recommend excluding. where mission equals 0 or NaN

4) The variable gov_wage_gdp is not specified in the instructions, and no source for the data is provided. We don’t know if this comes from an official sector body or from some potentially less rigorous institution. Furthermore, the data appears somewhat disconnected from our core violations and fines variables, as there are numerous NaNs for gov_wage_gdp where valid values exist for violations and fines. I do think the concept that this variable hints at - the impact of government workers’ wages on their willingness to occur fines - would be interesting to examine in relation to the degree of decline from the pre to the post subsets. The thinking being that those government workers with lower average salaries would be less willing to incur fines out of their own pocket. This is potentially problematic though as it is not clear that the workers themselves would definitely be the ones paying for the fines. ++++Suggestion: Given that we don’t know the provenance of this dataset nor how it was calculated, I would not treat results using this dataset with a high degree of confidence. Rather, I would be inclined to largely exclude this variable from the analysis At most, it should be used with caution in it’s own separate section.

5) The variable pctmuslim is among the more complete variables outside the core violations and fines dataset. We are not told the origin of this variable either, so we have no idea of its veracity other than a common sense check. The thought behind including this variable may be that religion and perhaps, Islam in particular, would have an impact on ethical or ethical (corrupt) behavior. It is not clear to me why Islam would be included and all other religions excluded. A more appropriate variable would be the percent of population which practices religion. It is also not clear to me that the ethics-religion association is necessarily as strong as some might think it is, though a more valid non-biased variable could be used to test that relationship.
“”"

```r
remove(variable_description)
```

```r
# 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'
```

```r
    x[numeric_columns] <-  round(x[numeric_columns], digits)
    x
}

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

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_l
```

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
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

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

Analysis of Key Relationships

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)
- 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)