Multiple Linear Regression Analysis of Homocide Rates

Vincient Arnold, Noah Labs, Linda Hu

February 2019

STAT3220 Regression Analysis

University of Virginia

1 Introduction

In this project, we will use statistical analysis to answer the following question: To what extent is it possible for us to predict the crime rate of a nation? Our motivation for researching this question is to understand predictive factors for homicide, since this is a major societal issue in every nation. The analysis utilized in this study was multiple linear regression with both quantitative and qualitative variables, screened via variable screening procedures.

2 Data Summary

2.1 Data Description

2.1.1 Sourcing and Explanation

The data is in the form of a comma-separated values (CSV) file. This data comes from The Human Freedom Index 2018, a report co-authored by Ian Vasquez and Tanja Porcnik and co-published by the Cato Institute, the Fraser Institute, and the Liberales Institut at the Friedrich Naumann Foundation for Freedom. A full description of the study can be found here. The expressed purpose of the report is to attempt to quantitatively measure human freedom, of various kinds. The research includes 1,458 data points collected from 162 nations (certain failed states, i.e. Democratic People's Republic of Korea and Somalia, were omitted) in 123 different categories, using data collected from the year 2016 (as misleading as the title may be, this is the last year in which full data was available).

2.1.2 Data Dictionary

Name	Description	Class
pf_ss_homocide	Homocide Rate	Quantitative
ef_government	Size of Government	Quantitative
ef_legal_enforcement	Legal Enforcement of Contracts	Quantitative
pf_rol	Rule of Law	Quantitative
ef_legal	Legal System and Property Rights	Quantitative
ef_legal_police	Reliability of Police	Quantitative
ef_legal_protection	Protection of Property Rights	Quantitative
ef_legal_police2	Squared Reliability of Police	Quantitative
ef_legal2	Squared Legal System and Prop Rights	Quantitative
blackmarket	if black market score not 10, then =1, else 0	Categorical
hi_integrity	if integrity score between 3 and 6, then =1, else 0	Categorical
lo_integrity	if integrity score less than 3, then =1, else 0	Categorical
trade_police	Black Markets and Police Interaction	Interaction

Table 1: Table of pertinent variables.

All of the above quantitative variables' units are given in a simple score, on a scale of one to ten. On these scales, a score of ten is the best scenario, with one being the worst. Thus, it should be noted that the homicide rate indicator is not a true percentage number, but rather a score on a scale of one to ten, ten being one of the lowest homicide rates on earth.
Note: Qualitative variables β_9 through β_{11} will be encoded as follows. After carefully observing all the scatter plots of the variable, we found that the scatter plots of blackmarket and integrity had some different patterns than others, and they fit for categorical data. For blackmarket, there are a lot of points concentrated on 10, so we decided to choose the base code at 0 if the value is 10, and 1 if not. Unlike other scatter plots with a lot of continuous points, integrity's scatter plot is more concentrated on certain values, and we decided to separate the value range to three sections. From range 0 to 3, the high integrity is 0, and the low integrity is 1. From range 3 to 6, the high integrity is 1, and the low integrity is 0. From range 6 to 10, the high integrity and the low integrity are both 0.

2.2 Advantages and Disadvantages

One of the strongest advantages of this data is its robustness. Due to the large variety of data (162 nations) and indicators, we have a very strong set of data with plenty of data points to work with. Further, this is population data, not sample data, if viewed as only for a certain year. Inasmuch, this data has a large n-value and adequately representative observations. However, one of the largest disadvantages of this data set is that each column, with the exception of the initial nation and region columns, are given as scores on a scale. While this is extremely useful for being able to compare seemingly disparate indicators without units, one sacrifices some accuracy and context for work-ability. For instance, being able use a scale of one to ten for homicide rate and legal enforcement of contracts is highly preferable as an analyst, however, it leads to an issue in that the report does a poor job of detailing how the researchers arrived at these scores. Thus, we do not have a great picture of what a given nation's homicide rate or black market exchange rate actually is. Finally, another notable disadvantage of this data is that, like much political science or economic data, there are numerous missing values. This can hardly be avoided in the nature of the research, however it does pose a barrier to analysis and detracts from the accuracy of any findings.

2.3 Exploratory Data Analysis

Below we include notable scatter plots for the first variables we considered and provide our interpretations for each variable.

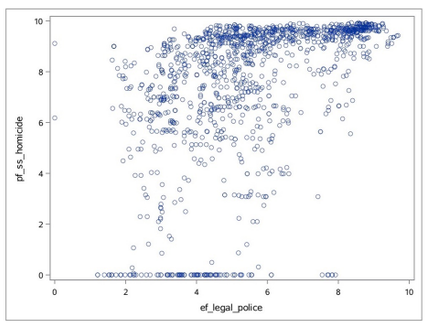

Figure 1: Reliability of Police vs Homicide

Note: Homicide is rated on a scale of one to ten, with ten being the best score (low homicide rate). In this case we noted a relatively strong positive linear relationship and chose to include this variable, we also thought it looked quadratic so we also considered it as a quadratic variable.

Legal System and Property Rights vs Homicide: We noted a strong positive relationship and decided to include this variable. We also thought that this could be a quadratic term so we also considered it as quadratic.

Rule of Law: We noted a strong positive relationship and decided to include this variable.

Legal Enforcement of Contracts: We noted a strong positive relationship and decided to include this variable.

Legal Integrity: We noted a strong positive relationship and decided to include this variable. However, we also noted that these were mostly measured in 'steps' so we decided to code this variable as qualitative.

Black Market Trade: This is perhaps the most unique plot and a clear candidate for categorical analysis. Thus, we coded this variable in two levels: either equal to ten or not.

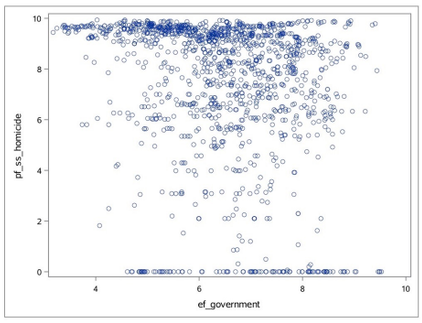

Figure 2: Size of Government vs Homicide

It appeared that there was no relationship, so we decided not to include this variable in this analysis. Because this was unusual, we thought it appropriate to readily provide the grahpic.

We then looked at the coded categorical variables. We chose a standard x-variable and grouped by qualitative class to get a grouped scatter plot. One largely representative plot is shown below; the others can readily be found in the appendix.

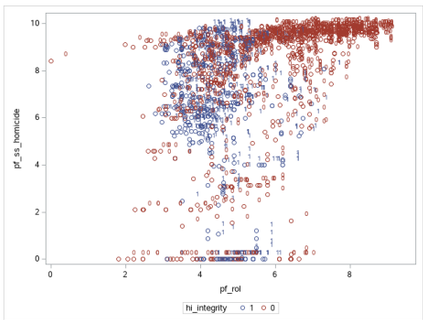

Figure 3: High Integrity Score, Qualitative

2.4 Exploratory Proposed Models

$$\mathrm{E(y)} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \beta_6 (x_4)^2 + \beta_8 (x_3)^2$$

Where: x_1 = ef_legal_enforcement, x_2 = pf_rol, x_3 = ef_legal, x_4 = ef_legal_police, x_5 = ef_legal_protection

3 Analysis

3.1 Stage I: Quantitative

We ran a stepwise variable screening process on all the (lower-order) quantitative variables in the data dictionary. We also ran backward selection and forward selection, and ended with the same four predictors. The process resulted in the following model.

Model 1: $\mathrm{E(y)} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4$,

where x_1 = ef_legal_protection, x_2 = ef_legal_enforcement, x_3 = ef_legal_police, x_4 = pf_rol

3.2 Stage II: Quantitative and Qualitative

We ran a nested F-test on the three categorical predictors given in the data dictionary. The nested F-test came back positive, showing that the addition of the three qualitative terms was significant. We arrived at the following model.

Model 2: $\mathrm{E(y)} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \beta_6 x_6 + \beta_7 x_7$,

where x_1 = ef_legal_protection, x_2 = ef_legal_enforcement, x_3 = ef_legal_police, x_4 = pf_rol, x_5 = blackmarket , x_6 = hi_integrity , x_7 = lo_integrity

3.3 Stage III: Quantitative, Qualitative, and Interactions & Higher Order Terms

We ran a nested F-test for legal_police squared and the interaction between black-market and legal_police. The nested F-test resulted in a p-value of 0.0773, which was significantly larger than our acceptable alpha level of 0.05. Thus, we failed to reject the null (that at least one of the added betas is not equal to zero) and arrived at the same model as Model 2.

$$\text{Model 3: } E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \beta_6 x_6 + \beta_7 x_7,$$

where x_1 = ef_legal_protection, x_2 = ef_legal_enforcement, x_3 = ef_legal_police, x_4 = pf_rol, x_5 = blackmarket, x_6 = hi_integrity , x_7 = lo_integrity

3.4 Stage IV: Analysis of Models

Model	Adj R-Square	P	P-value	Mean VIF	MCp
1	0.273	4	<.0001	3.153	5.000
2	0.355	7	<.0001	2.578	8.000
3	0.355	7	<.0001	2.578	8.000

Table 2: Table of pertinent statistics of models. P-value is from Global F-tests. P is number of predictors. MCp is Mallow's Cp. All values are rounded to three decimal points.

Due to the significant increase in Adjusted R-Squared with the addition of only three terms, we decided that Model 2 was preferable. Notably, the mean Variance Inflation Factor actually decreased.

3.5 Stage V: Residual Analysis

Having selected Model 2, we now will analyze residual plots and individual VIF terms to verify regression assumptions and potential multicollinearity. No notable VIF issues were present, as the average was far less than 10 and the individual VIF numbers were unremarkable, suggesting the lack of a multicollinearity issue. We checked the residual plots for the effectiveness of our final model. The residual by predicted plot showed a strong linear trend and did not have constant variance, so we concluded that the errors violated the assumption of lack of fit and that the errors were heteroscedastic. Observing the histogram and QQ plot of residuals, we found that the histogram had an approximate bell-shape, but the QQ plot had a curved trend, which violated the assumption of normality. In order to fix all of those problems, we transformed the independent variables and dependent variables. Below are the original residual plot, QQ plot, residual histogram, and outlier/leverage point plot.

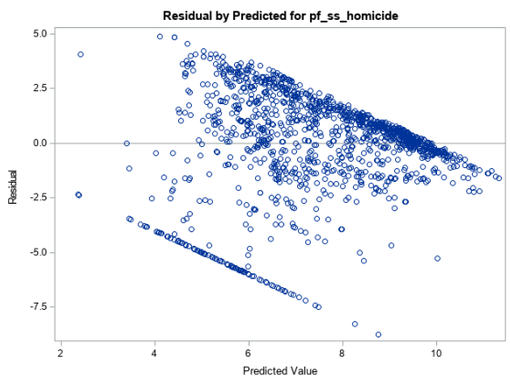

Figure 4: Residual Plot

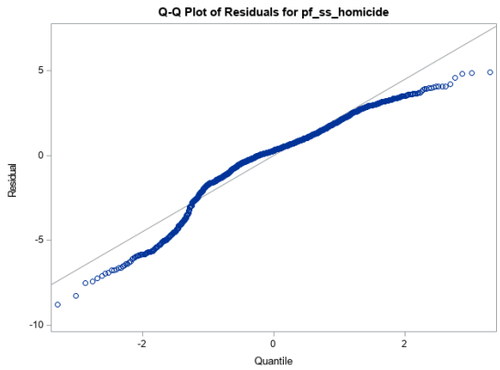

Figure 5: QQ Plot

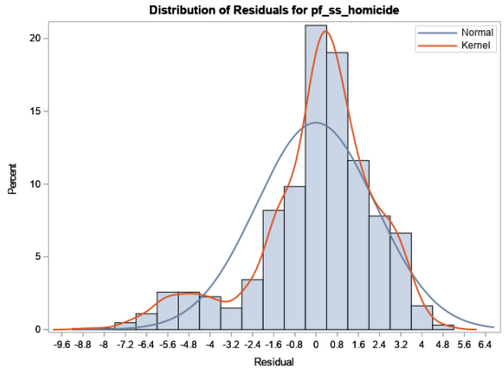

Figure 6: Residual Histogram

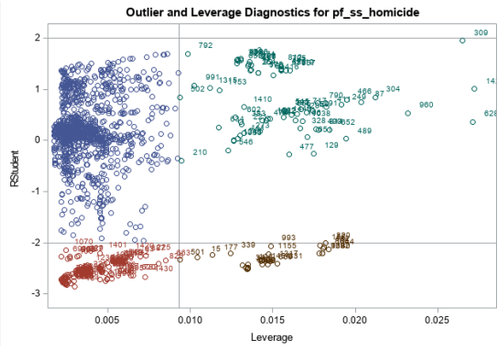

Figure 7: Outlier Plot

3.5.1 Transformation of y to $\ln(y)$

After having transformed the response variable via a natural logarithm stabilizing transformation, the residual by predicted plot did not have a clear trend, yet it still had a problem of heteroscedasticity, notably fanned in. However, it significantly improved compared to our previous model (original model without variable transformation). The histogram was bell-shaped, and the Q-Q plot had a clear linear trend. In addition, the numbers of outliers and leverages were significantly reduced, and this model had a variance inflation less than 3. It demonstrated an good transformation, but legal protection and black market became insignificant under this model.

3.5.2 Transformation of y to $\sqrt{y}$

After having transformed the response variable via a square root stabilizing transformation, the residual by predicted plot still had a strong linear relationship a fanning-in

pattern. The histogram was bell-shaped, but the Q-Q plot did not show a linear relationship. The residual plot problems in the original model still existed. However, the VIF test result was still ideal; all the predictors were significant and it reduced a significant number of outliers.

3.5.3 Transformation of a Predictor

In order to fix the problem of lack of fit in the original model, we decided to transform an independent variable as well. Since the transformation of $\ln(y)$ gave us an ideal result, we decided to also transform the independent variable with such a stabilization technique. After several attempts, we decided the legal protection was the best candidate for transformation. The residual by predicted plot did not have a clear pattern and did not show strong evidence of violating constant variance. The histogram was bell-shaped, and the Q-Q plot had a clear linear trend. The number of outliers and leverage points was further reduced. However, under this model, rule of law and legal protection became insignificant and thus we removed them from the model.

3.5.4 A Proposed Final Model

As stated above, we removed the insignificant variables (legal protection and rule of law) from the model, and all the variables became significant. It further improved the Adjusted R Squared from 0.1957 to 0.1965. It still solved all the problems of residual plots and reduced the number of outliers and leverage points.

Final Model: $\mathrm{E(y)} = \mathrm{e}^{(}\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 \ln_{(} x_4) + \beta_5 x_5$

Final Prediction Equation: $\ln(\hat{y}) = 1.662 - 0.093\mathrm{x}_1$ - $0.208x_2$ - $0.253x_3 + 0.665 \ln_{(} x_4)$ - $0.020x_5$

,where where $x_1 =$ ef_legal_protection, $x_2 =$ hi_integrity, $x_3 =$ lo_integrity, $x_4 =$ ef_legal, $x_5 =$ ef_legal_enforce

β_0: The estimate of log of mean of homicide rate is 1.662 when protection of property rights, high legal integrity, low legal integrity, ln(legal system and property rights) and legal enforcement of contracts are 0 units.

β_1: For every one unit increase in protection of property rights, the estimate of log of mean of homicide rate decreases by 0.093, while all other variables stay constant.

β_2: The estimate of log of mean of homicide rate decreases by 0.208 when the average difference between the category high integrity changes from 0 to 1, while all other variables stay constant.

β_3: The estimate of log of mean of homicide rate decreases by 0.253 when the average difference between the category low integrity changes from 0 to 1, while all other variables stay constant.

β_4: For every one unit increase in the log of legal system and property rights, the estimate of log of mean of homicide rate increases by 0.665, while all other variables stay constant.

β_5: For every one unit increase in legal enforcement of contracts, the estimate of log of mean of homicide rate decreases by 0.020 while all other variables stay constant.

4 Conclusion

Our final model was chosen based on the significance of individual independent variables, Adjusted R Square, lack of multicollinearity, the number of outliers and leverage points, and ideal residual plots. Each independent variable is significant at $\alpha = 0.05$ (also at 0.02), and the overall model has a high utility with a Global F Test p-value $<.0001$. The model does not have the problems of multicollinearity, lack of fit, heteroscedasticity and abnormality. The number of outliers and leverage points is reasonable for a sample size of 1,458 data points. The trade-off however, is that our final model only has an Adjusted R Squared of 0.1965, which means that only 19.65% of the variation in the response variable can be explained by the model. This value is regrettably significantly less than our initial model, which boasted an Adjusted R

Squared of 0.3662. In general, the more variables we included, the higher Adjusted R Squared we observed. Yet due to the restrictions of various requirements, we deleted many variables in the process. It is far from a perfect model; yet some of this predictive inadequacy is due to the size and variability of the data. In the original data, it has 123 variables, we only chose a small portion of it for the data analysis, largely because many had an unacceptably large number of missing data points. Nonetheless, despite its shortcomings, we are confident that this model has the most significant predictive power possible given constraints of the data set and the selected candidate predictors.

5 SAS Code

```
* Here we will upload the CSV file for our project;
proc import datafile = '/folders/myfolders/Homeworks/hfi_cc_2018.csv'
 out = work.hfi_cc_2018
 dbms = CSV
 ;
run;   *the file is successfully imported;

* Some quantitative scatterplots to follow;
proc sgplot data=work.hfi_cc_2018;
scatter y=pf_ss_homicide x=ef_legal_police;
run;   * scatterplot for homicide vs police reliability;

proc sgplot data=work.hfi_cc_2018;
scatter y=pf_ss_homicide x=ef_legal;
run;   * scatterplot for homicide vs legal system and property rights;

proc sgplot data=work.hfi_cc_2018;
scatter y=pf_ss_homicide x=ef_legal_enforcement;
run;   * scatterplot for homicide vs legal enforcement of contracts;

proc sgplot data=work.hfi_cc_2018;
scatter y=pf_ss_homicide x=ef_legal_integrity;
run;   * scatterplot for homicide vs integrity of legal system;

proc sgplot data=work.hfi_cc_2018;
scatter y=pf_ss_homicide x=ef_government;
run;   * scatterplot for homicide vs size of government;

proc sgplot data=work.hfi_cc_2018;
scatter y=pf_ss_homicide x=pf_rol;
run;   * scatterplot for homicide vs Rule of Law (strength of);

proc sgplot data=work.hfi_cc_2018;
scatter y=pf_ss_homicide x=ef_legal_protection;
run;   * scatterplot for homicide vs protection of property;



* QUALITATIVE VARIABLES;
data bmdata; * coding dummy variables for black market ;
set work.hfi_cc_2018;
blackmarket = 0;
if ef_trade_black < 10 then blackmarket = 1;
run;

data bmdata2; *coding dummy variables for legal integrity scores;
set bmdata;
hi_integrity = 0;
lo_integrity = 0;
if ef_legal_integrity => 3 and ef_legal_integrity < 6 then hi_integrity=1;
if ef_legal_integrity < 3 then lo_integrity = 1;
run;

*Create grouped scatterplots;
proc sgplot data=bmdata2;
scatter y=pf_ss_homicide x=pf_rol / group=hi_integrity datalabel=hi_integrity;
run; * grouped scatter for hi_integrity;
```

```
proc sgplot data=bmdata2;
scatter y=pf_ss_homicide x=pf_rol / group=lo_integrity datalabel=lo_integrity;
run; * grouped scatter for lo_integrity;

proc sgplot data=bmdata2;
scatter y=pf_ss_homicide x=pf_rol / group=blackmarket datalabel=blackmarket;
run; * grouped scatter for blackmarket;

* Use Iterative selection procedures;
proc reg data=final_data_4 plots=none;
model pf_ss_homicide = ef_legal_protection ef_legal ef_legal_police pf_rol ef_government ef_legal_enfo
selection=stepwise SLentry=0.05 SLstay=0.10 details;
run;


proc reg data=final_data_4 plots=none;
model pf_ss_homicide = ef_legal_protection ef_legal ef_legal_police pf_rol ef_government ef_legal_enfo
selection=forward SLentry=0.05;
run;


proc reg data=final_data_4 plots=none;
model pf_ss_homicide = ef_legal_protection ef_legal ef_legal_police pf_rol ef_government ef_legal_enfo
selection=backward SLstay=0.05 details=all;
run;

* This is the regression;

proc reg data=final_data_4 plots=none;
model pf_ss_homicide = ef_legal_protection ef_legal_enforcement
ef_legal_police pf_rol / selection=CP best=3 ADJRSQ RMSE;
Run;

Model 1

* Now testing categorical;

proc reg data=final_data_4 plots=none;
model pf_ss_homicide = ef_legal_protection ef_legal_enforcement ef_legal_police pf_rol blackmarket hi_
test blackmarket,hi_integrity,lo_integrity;
run;

*This becomes model 2 and model 3;

proc reg data=final_data_4 plots=none;
model pf_ss_homicide = ef_legal_protection ef_legal_enforcement ef_legal_police pf_rol
blackmarket hi_integrity lo_integrity / selection=CP ADJRSQ RMSE
run;

* Now higher-order and interaction;

proc reg data=final_data_4 plots=none;
model pf_ss_homicide = ef_legal_protection ef_legal_enforcement
ef_legal_police pf_rol blackmarket hi_integrity lo_integrity ef_legal_police2 trade_police;
test ef_legal_police2, trade_police;
run;

proc reg data=final_data_4 plots=none;
model pf_ss_homicide = ef_legal_protection ef_legal_enforcement
ef_legal_police pf_rol blackmarket hi_integrity lo_integrity
ef_legal_police2 trade_police;
run;

*This right here showed that the quadratic and interaction were insignificant;

* Now checking assumptions;

proc reg data=final_data_4 plots(only)=(residualbypredictedresidualplot qqplot residualhistogram);
model pf_ss_homicide = ef_legal_protection pf_rol blackmarket hi_integrity lo_integrity
```

```
ef_legal_police ef_legal_enforcement/ vif;
run;

* ++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++;

data final;
set final_data_4;
ln_y = log(pf_ss_homicide);
run;

proc reg data=final plots(only)=(residualbypredictedresidualplot qqplot residualhistogram);
model ln_y = ef_legal_protection pf_rol blackmarket hi_integrity lo_integrity
ef_legal_police ef_legal_enforcement / dwprob;
run;

data final_2;
set final;
sqrty = sqrt(pf_ss_homicide);
run;

proc reg data=final_2 plots(only)=(residualbypredicted residualplot qqplot residualhistogram);
model sqrty = ef_legal_protection pf_rol blackmarket
hi_integrity lo_integrity ef_legal_police ef_legal_enforcement / dwprob;
run;

data final_3;
set final_2;
log_legal = log(ef_legal_police);
run;

proc reg data=final_3 plots(only)=(residualbypredicted residualplot qqplot residualhistogram);
model pf_ss_homicide = ef_legal_protection pf_rol blackmarket
hi_integrity lo_integrity log_legal ef_legal_enforcement;
run;


proc reg data=final plots=none;
model ln_y = ef_legal_protection pf_rol blackmarket
hi_integrity lo_integrity ef_legal_police ef_legal_enforcement / vif;
run;

proc reg data=final_2 plots=none;
model sqrty = ef_legal_protection pf_rol blackmarket
hi_integrity lo_integrity log_legal ef_legal_enforcement / vif;
run;


proc reg data=final_3 plots=none;
model ln_y = ef_legal_protection pf_rol blackmarket
hi_integrity lo_integrity log_legal ef_legal_enforcement / vif;
run;

proc reg data=final_3 plots=none;
model ln_y = ef_legal_protection
hi_integrity lo_integrity log_legal ef_legal_enforcement / vif;
run;
```

6 Optional Appendix

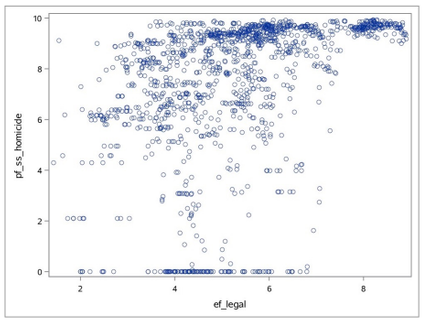

Figure 8: Legal System and Property Rights vs Homicide

Legal System and Property Rights vs Homicide: We noted a strong positive relationship and decided to include this variable. We also thought that this could be a quadratic term so we also considered it as quadratic.

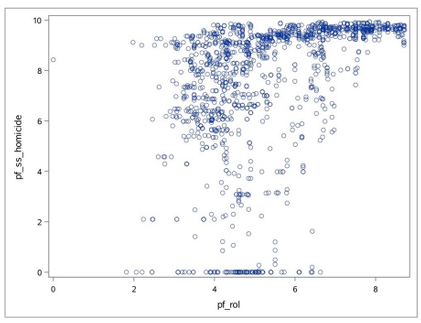

Figure 9: Rule of Law vs Homicide

Rule of Law: We noted a strong positive relationship and decided to include this variable.

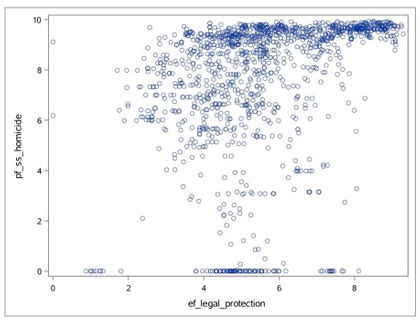

Figure 10: Legal Protection vs Homicide

We noted a strong positive relationship and decided to include this variable.

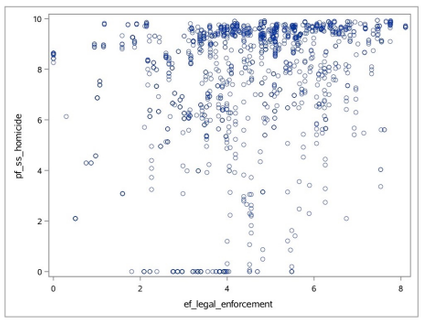

Figure 11: Legal Enforcement of Contracts vs Homicide

Legal Enforcement of Contracts: We noted a strong positive relationship and decided to include this variable.

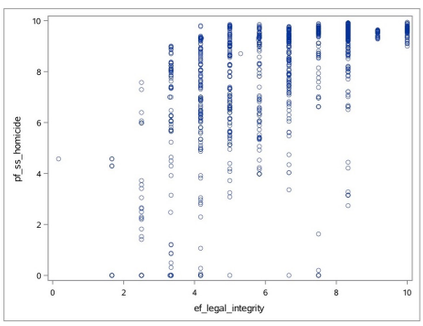

Figure 12: Legal Integrity vs Homicide

Legal Integrity: We noted a strong positive relationship and decided to include this variable. However, we also noted that these were mostly measured in 'steps' so we decided to code this variable as qualitative.

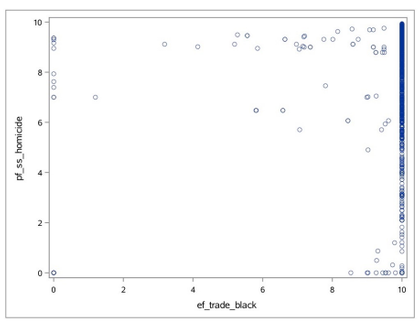

Figure 13: Black Market Trade vs Homicide

Black Market Trade: This is perhaps the most unique plot and a clear candidate for categorical analysis. Thus, we coded this variable in two levels: either equal to ten or not.

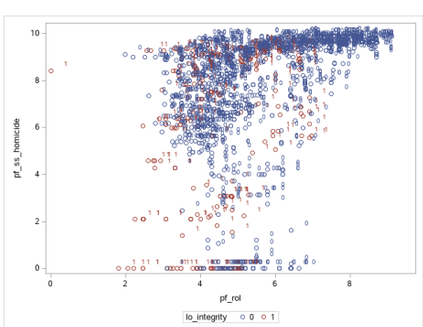

Figure 14: Low Integrity Score, Qualitative

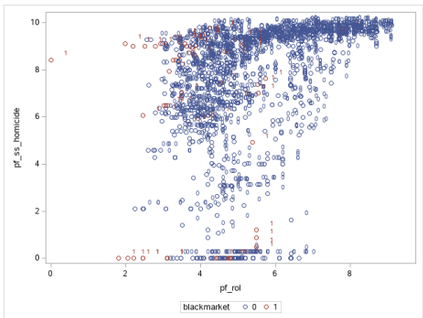

Figure 15: Black Market Trade, Qualitative