A Statistical Model to Identify Social Issues Affecting the United States

Homicide Rate

Brianna L. Palmisano (X03625986)
St. John's University, The Peter J. Tobin College of Business

Summer 2024

Table of Variables

Variable	Abbr.	R-Script
Homicide Rate	HR	homcideRate
Unemployment Rate	UR	unempRate
Income Level	IL	income
Bachelor's Degree	BD	bachDegree
High School Degree	HSD	hsDegree
Obesity Rate	OR	obesityRate
Robbery Rate	RR	robberyRate
Suicide Rate	SR	suicideRate
Teenage Pregnancy Rate	TPR	teenPregRate

Introduction

A “homicide” is defined by *Oxford Languages* as *“the killing of one person by another”* and is considered one of the most heinous levels of crime committable. **This evaluation uses the United States homicide rate against other social issues or indicators to examine the linear relationship between these variables**. This includes the unemployment rate, level of income, obesity rate, robbery rate, suicide rate, teen pregnancy rate, as well as the completion of a high school and/or bachelor’s degree.

Information regarding the homicide rate across the US could not only be valuable for analyzing the the social issues and economic factors that go into our understanding of predicting regional homicide rates, but for research within legal studies and other crime related fields as well.

Methodology

[SENTENCE ABOUT DATA SOUCRE HERE] This is secondary data, as these observations have been adjusted for the integrity of the data. Here we have cross-section data, including 48 states. All observations were taken on the last day of the year, 2021.

The variables used in this model can be considered social issues, as well as economic indicators in some context. **The dependent variable here is the homicide rate (HR), and the independent variables are the unemployment rate (UR), level of income (IL), obesity rate (OR), robbery rate (RR), suicide rate (SR), teen pregnancy rate (TPR), as well as the completion of a high school (HSD) and/or bachelor's degree (BD), for each of those specified 48 states**.

Graphical techniques including both histograms and scatterplots are used in this analysis. This data has been analyzed using descriptive statistics (for scalable variables), as well as correlation and regression statistical analysis, via both Microsoft Excel and R-Script.

Methodology, continued…

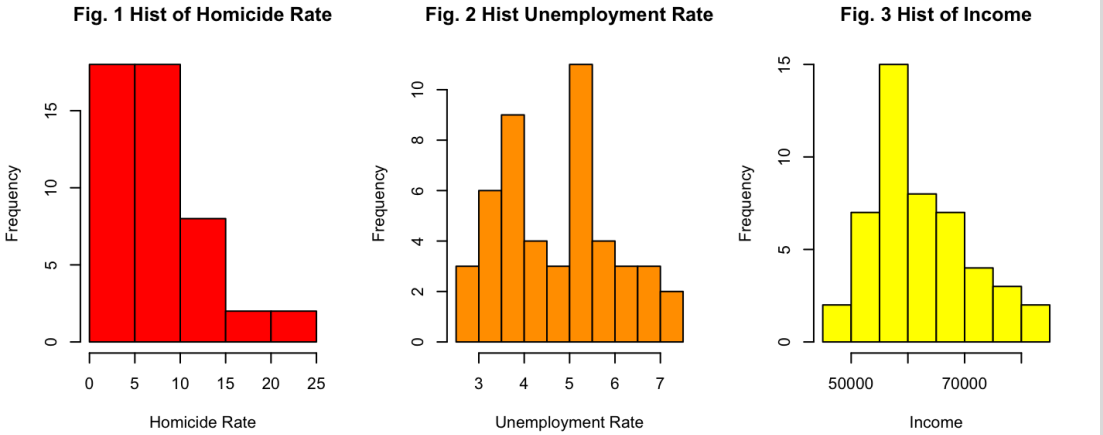
Listed below are the equations defining the functional specification (Eqn. 1), population regression equation (Eqn. 2) and sample regression equation (Eqn. 3).

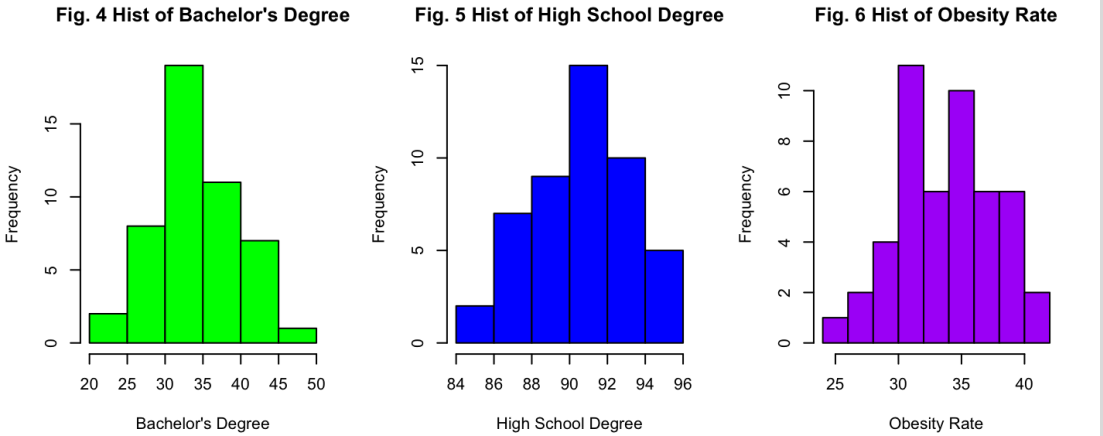
Eqn. 1 Homicide Rate = f(UR, IL, BD, HSD, OR, RR, SR, TPR)

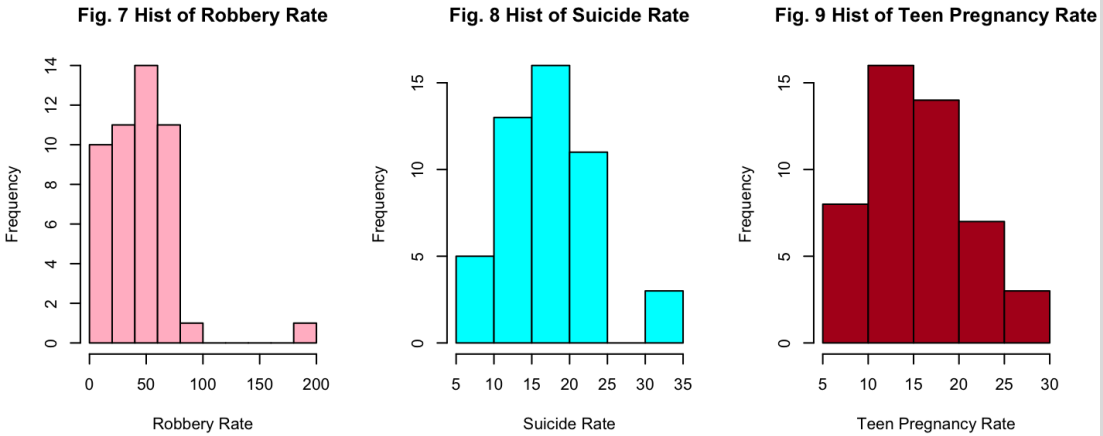
Eqn. 2 Homicide Rate = a + ßur UR + ßil IL + ßbd BD + ßhsd HSD + ßor OR + ßrr RR + ßsr SR + ßtpr TPR

Eqn. 3 Homicide Rate = a + bur UR + bil IL + bbd BD + bhsd HSD + bor OR + brr RR + bsr SR + btpr TPR

The objective of this model is to evaluate the association or lack thereof between these independent variables and the dependent variable.







Histograms

Fig. 1 through *Fig. 9* show the histograms of each variable being analyzed.

All histograms are roughly normally distributed, excluding the graphs of HR (the dependent variable), UR, OR and RR.

Fig. 1 and *Fig. 7* (graphs of of HR and RR) are both negative and skewed right.

There is less of a recognizable, weak pattern in *Fig. 2* and *Fig. 6 (*graphs of UR and OR), having more peaks and irregular symmetry.

Scatterplots

The relationships visible via the scatterplots in *Fig. 10* through *Fig. 17* are the independent variables, each relative to the dependent variable.

All graphs are nonlinear, lack a clear positive or negative pattern, and thus hold a weak relationship. A lack of trend across all variables could be due to outliers.

The relationships seen in *Fig. 11 – 12* are more negative, although they are weak. Whereas in *Fig. 14 – 15 and 17,* the pattern is positive and weak. There are fewer outliers as well.

★ The graphs with the strongest and most linear relationships here are *Fig. 11 - 12* (IL and BD) and *Fig. 17 (*TPR) when plotted against the HR. The graphs of IL and BD were negatively correlated, while the graph of TPR was positive.

Fig. 10 Unemployment Rate vs. Homicide Rate

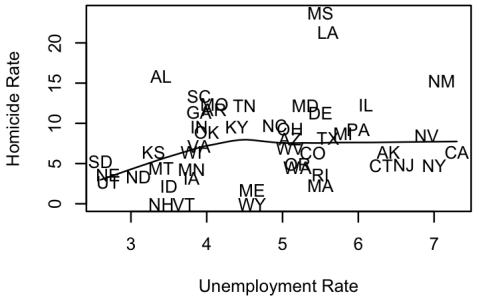


Fig. 11 Income vs. Homicide Rate

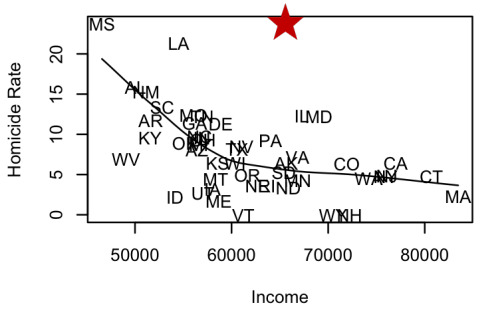


Fig. 12 Bachelor's Degree vs. Homicide Rate

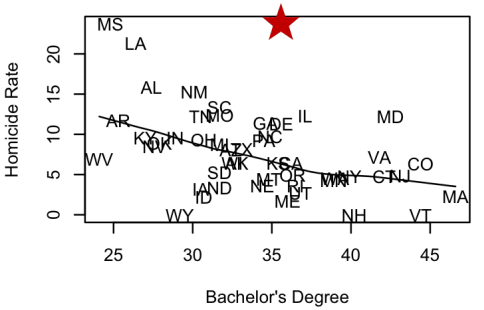


Fig. 13 High School Degree vs. Homicide Rate

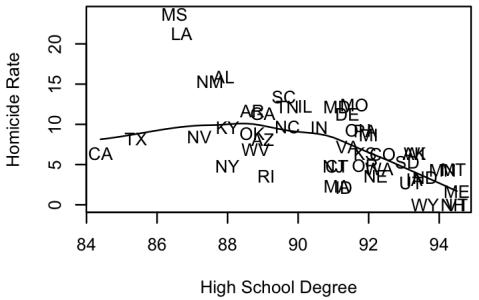


Fig. 14 Obesity Rate vs. Homicide Rate

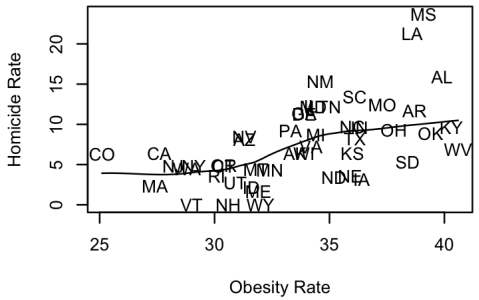


Fig. 15 Robbery Rate vs. Homicide Rate

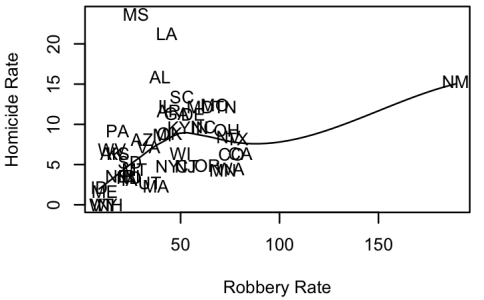


Fig. 16 Suicide Rate vs. Homicide Rate

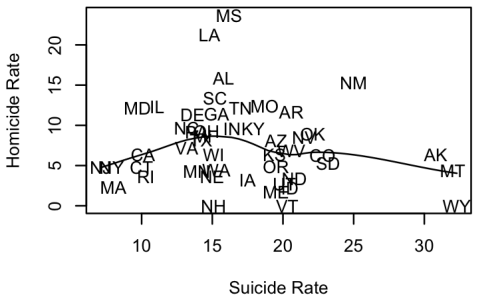
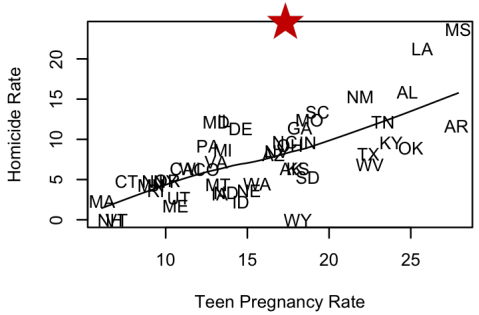


Fig. 17 Teen Pregnancy Rate vs. Homicide Rate

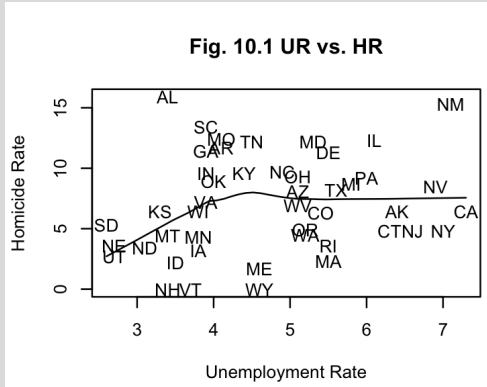


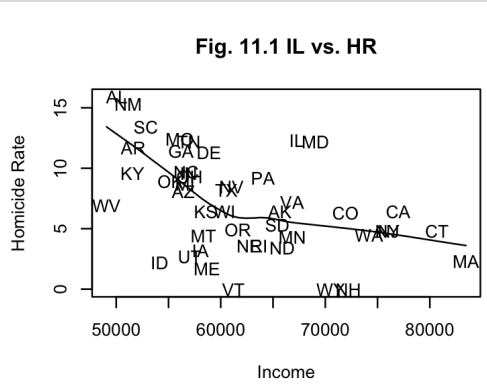
Post-Removal of Outliers

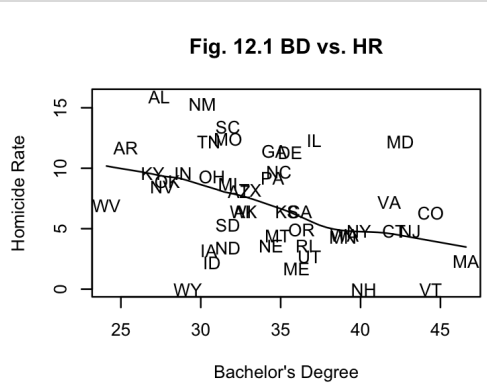
A lack of trend across all variables could be due to outliers in the data for states including Mississippi and Louisiana. **These states could have less economic activity in general, and for many reasons (low population levels, consumer norms, etc.), resulting in extreme data in either direction.** *Fig. 10.1* through *17.1* are the same scatterplots as earlier, following the removal of these states.

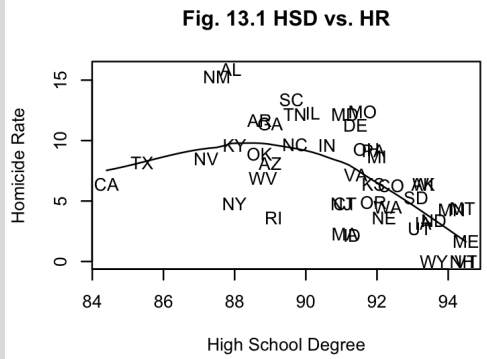
The data in *Fig. 15.1 (RR)* is congested and clustered at the beginning of the graph and only stretches out because of the data for New Mexico. This state was not removed, as it is only an outlier when looking at the RR and is not an outlier when looking at our other crime related independent variables.

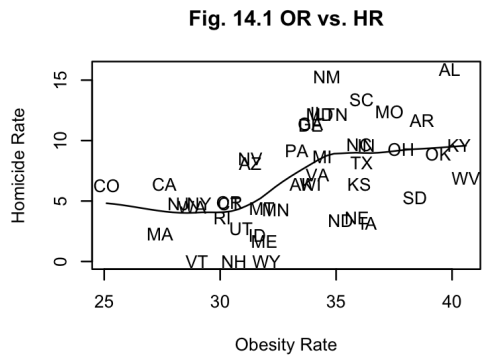
Removing outliers was not impactful of the results in this part of the model. **The outliers in question have been included in the following statistical analysis** because they may not be outliers, as well as to avoid compromising the number of observations included here or the integrity of this model.

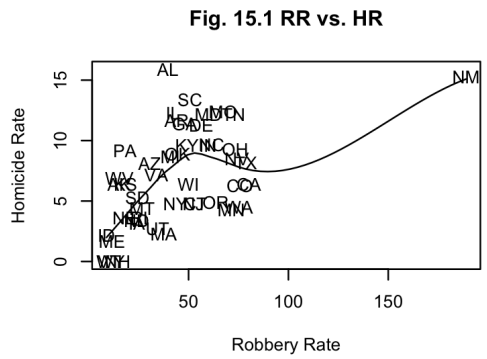


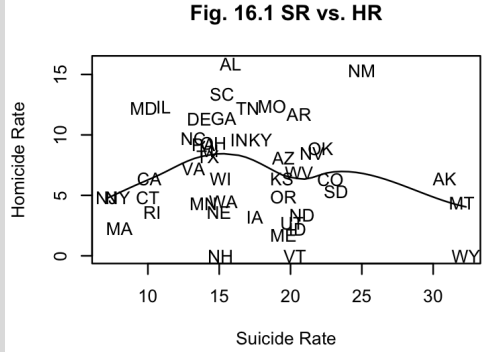












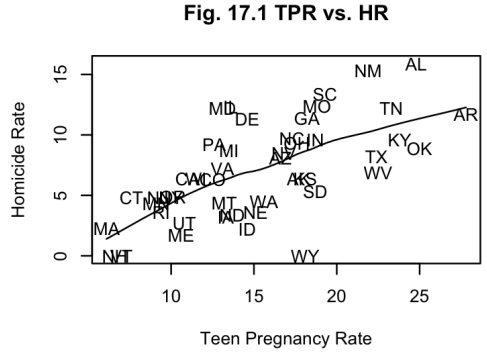


Table 1: *Descriptive Statistics*

In *Table 1*, are the descriptive statistics for all variables that are non-categorical, here, that is all variables. The standard deviation for HR, UR, IL and RR is disbursed relatively close to that variables mean, indicating that they all have moderate to low variability. The standard deviation is higher than the mean for IL, meaning this data deviates in the opposing direction.

	name	obs	max	min	mean	median	std	skew	kurt
1	HR	48	23.70	0.00	7.696	6.650	5.045	0.942	4.274
2	UR	48	7.30	2.60	4.773	4.750	1.248	0.197	2.151
3	IL	48	83461.00	46577.00	61737.896	59635.000	8599.662	0.613	2.779
4	BD	48	46.60	24.10	34.092	33.750	5.562	0.305	2.434
5	HSD	48	94.50	84.40	90.756	91.200	2.543	-0.458	2.499
6	OR	48	40.60	25.10	33.817	34.050	3.767	-0.114	2.328
7	RR	48	188.93	8.91	45.620	43.275	30.050	2.152	11.734
8	SR	48	32.30	7.10	17.102	16.000	5.710	0.702	3.733
9	TPR	48	27.90	6.10	15.802	14.850	5.732	0.343	2.326

The measure of kurtosis is positive and slightly higher or more peaked than normal distribution in all variables. This is higher around our dependent variable, and especially high in RR (indicating that this data is the most dispersed). The higher kurtosis seen indicates sharper peaks, with a higher chance of outliers.

The measure of skewness is asymmetric across all variables in this model. Both HSD and OR are skewed negatively, whereas all other variables were positive. UR and OR are the closest to being normally distributed (in opposing directions), whereas HR, RR and SR were the least (stretching further right from center than the other variables).

Table 2: *Correlation Matrix*

Via the correlation matrix in *Table 2*, it can be observed that the homicide rate is most positively correlated with the TPR (at 0.717) and most negatively correlated with HSD (at -0.625). The highest is positive correlations between independent variables are between OR and TPR (at 0.793) and between IL and BD (at 0.775). The highest negative correlations are between BD and TPR (at -0.823) (the strongest correlation here) and between BD and OR (at -0.773). These correlations are worth noting because although they are moving in opposing directions, there is still a relationship between these variables greater than others in this model.

	HR	UR	IL	BD	HSD	OR	RR	SR	TPR
HR	1.000	0.239	-0.538	-0.499	-0.625	0.574	0.376	-0.132	0.717
UR	0.239	1.000	0.320	0.127	-0.498	-0.333	0.448	-0.292	-0.077
IL	-0.538	0.320	1.000	0.775	0.247	-0.731	-0.027	-0.355	-0.711
BD	-0.499	0.127	0.775	1.000	0.395	-0.773	-0.008	-0.412	-0.823
HSD	-0.625	-0.498	0.247	0.395	1.000	-0.270	-0.410	0.301	-0.549
OR	0.574	-0.333	-0.731	-0.773	-0.270	1.000	-0.049	0.154	0.793
RR	0.376	0.448	-0.027	-0.008	-0.410	-0.049	1.000	-0.107	0.176
SR	-0.132	-0.292	-0.355	-0.412	0.301	0.154	-0.107	1.000	0.319
TPR	0.717	-0.077	-0.711	-0.823	-0.549	0.793	0.176	0.319	1.000

Due to high correlation coefficients, we can assume that there are signs of multicollinearity in this model, or a high correlation in our independent variables.

Regression Equation

Equation for Sample Regression Line.

			+	+	+	+	+	+	+	+
Eqn. 4		Homicide Rate =	f(UR,	IL,	BD,	HSD,	OR,	RR,	SR,	TPR)
	t-stat		(-1.05)	(2.94)***	(-2.94)***	(1.63)**	(0.9)	(0.34)	(1.01)	(-2.7)***	(3.86)***
	p-value		(0.3)	(0.01)	(0.1)	(0.11)	(0.37)	(0.74)	(0.32)	(0.01)	(0.00)
	r (corr)			()	()	()	()	()	()	()	()

n = 48 r-sq. = 0.773 F = 16.62*** F-Prob = SE =

Confidence Intervals.

*	Significant at the	0%	level of significance (90% Sure, or "are below") (1.28)
**	Significant at the	5%	level of significance (95% Sure, or "are below") (1.65)
***	Significant at the	1%	level of significance (99% Sure, or "are below") (2.33)

Results of an F-test for the entire model.

	Ho:	bur = bil = bbd = bhsd = bor = bsr = brr = btpr = 0	(Null Hypothesis)
* 1%	Ha:	at least 1 bi not equal to 0 (16.62 > 4.99)	(Alternate Hypothesis)

The above F-test includes the null (Ho) and alternative hypothesis (Ha) for the entire model.

```
Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)  -2.981e+01  2.848e+01  -1.047 0.301754
unempRate     1.381e+00  4.706e-01   2.935 0.005562 **
income       -2.358e-04  8.021e-05  -2.940 0.005494 **
bachDegree    2.685e-01  1.648e-01   1.629 0.111388
hsDegree      2.935e-01  3.249e-01   0.904 0.371749
obesityRate   8.504e-02  2.537e-01   0.335 0.739268
robberyRate   1.568e-02  1.547e-02   1.013 0.317074
suicideRate  -3.202e-01  1.185e-01  -2.702 0.010155 *
teenPregRate  7.315e-01  1.898e-01   3.855 0.000421 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.637 on 39 degrees of freedom
Multiple R-squared:  0.7732,    Adjusted R-squared:  0.7267
F-statistic: 16.62 on 8 and 39 DF,  p-value: 2.164e-10
```

Table 3: *Regression Statistics*

To further interpret the regression statistics in *Table 3*, the F-Statistic (equal to 16.62) indicates that the regression model is statistically significant, as it is above the cutoff (>4.99) for this measure. **This inclines us to determine that the null hypothesis should be rejected.**

Rather, the higher value insists that the model explains a significant amount of the variation in the dependent variable with the predictor variables, than without.

The results for our p-value show that the variables with the association that is the most significant are TPR (at 0.000), UR (at 0.006), IL (at 0.005) and SR (at0.01), with TPR being the strongest. **These variables all having high significance, inclines us to reject the null hypothesis at the 99% confidence level** (all confidence levels were above 99% confidence level interval).

Regression Results, continued…

The results for the t-Statistic indicate are inline with the levels of significance expressed in these p-values. When looking at the results for the coefficient of determination, it can be concluded that that statistical model predicts an association between the dependent and independent variables that is strong. **There is 77.3% of variation in the dependent variable can be explained by variations in the independent variables.**

There is a lot of room for uncertainty in this analysis which could be due to initial assumptions of multicollinearity; too much correlation between our independent variables. **The greater standard error value seen in the intercept (HR) determines that the dependent variable is being affected by other independent variables, ones not included in this model.**

To discuss the statistically significant regression coefficients, the dependent variable (HR) is most impacted by the UR (at 1.381), TPR (at 0.732) and the SR (at -0.320). **That is to say that for every unit increase in the independent variable UR, there is a 1.381 unit increase in the dependent variable (HR) (similar logic can be applied to TPR), and that there is a 0.320 unit decrease in HR for every unit increase in the SR, holding all other variables constant.**

Residuals Analysis

Fig. 18 Histogram of Fitted Residual

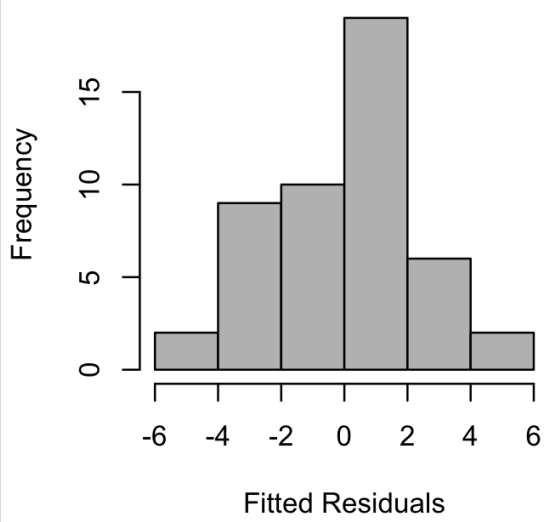
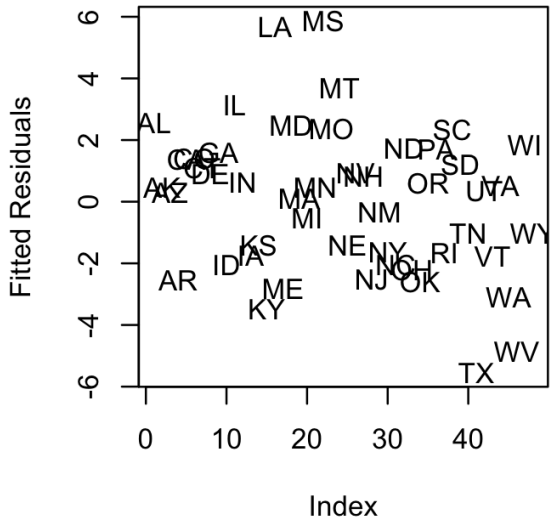


Fig. 19 Plot of Fitted Residuals



The results from the above fitted residuals were are approximately normally distributed in *Fig. 18* , though there is a lack of clear positive and negative pattern or visible linear trendline in *Fig 19.* **The accuracy of this model is debatable.**

Conclusions

For this model, the null hypothesis is rejected, and the alternate hypothesis is accepted. The research presented here can be considered successful, as 77.3% of variation in the dependent variable can be explained by the independent variables. It can be determined true that the dependent variable is better supported by the independent variables, than if it were not. **This model was not entirely predictive, as the variables here do not influence one-another enough, but it was somewhat predictive when looking at TPR, UR, IL, RR, and SR.**

This model can be improved on by increasing the number of observations, as there were missing states and missing values under some states that were included in this input data. Missing or "NA" values were omitted from this model. **More observations would increase the validity of any model, but it can be assumed that this would not change our overall results in this analysis.**

Should there be a need to break down the moments of closer associations, breaking this model up by time periods may add to the findings that this model is capable of. **The highly correlated independent variables or multicollinearity seen in this model can be adjusted by switching those variables out and taking the first difference for those variables to reanalyze the trends in those variables.**

Public Policy Implications

Public policy does not entirely pertain to driving a relationship between these variables but could be influential in the context of analyzing the social and economic factors that play into changes in regional homicide rate, as well as in research fields including law, criminology or even sociology. This model or an expanded version of this model could be beneficial to professionals or investors who require more insight on the variables that influence the homicide rate or general social issue levels per region of America.

Appendix I: *Bibliography*

[CITED DATA SOURCE FOR SOCIAL VARIABLES GOES HERE]

Appendix II: *Input Data*

Via Microsoft Excel file.

year	index	stateShort	state	homcideRate	unempRate	income	bachDegree	hsDegree	obesityRate	robberyRate	suicideRate	teenPregRate
2021	1	AL	Alabama	15.9	3.4	50,059	27.4	87.9	39.90	39.6	15.8	24.8
2021	2	AK	Alaska	6.4	6.4	65,662	32.8	93.3	33.50	15.32	30.8	17.7
2021	3	AZ	Arizona	8.1	5.1	56,420	32.4	89	31.30	30.52	19.5	16.6
2021	4	AR	Arkansas	11.7	4.1	51,636	25.3	88.7	38.70	44.05	20.6	27.8
2021	5	CA	California	6.4	7.3	76,991	36.2	84.4	27.60	80.22	10.1	11
2021	6	CO	Colorado	6.3	5.4	71,923	44.4	92.4	25.10	75.63	22.8	12.5
2021	7	CT	Connecticut	4.8	6.3	80,691	42.1	91.1	30.40	53.36	10	7.6
2021	8	DE	Delaware	11.3	5.5	58,889	35.6	91.4	33.90	56.31	13.6	14.6
2021	9	GA	Georgia	11.4	3.9	56,184	34.6	89	33.90	48.11	15.3	18.2
2021	10	ID	Idaho	2.2	3.5	54,148	30.7	91.3	31.60	8.91	20.5	14.6
2021	11	IL	Illinois	12.3	6.1	67,278	37.1	90.2	34.20	42.25	11.1	13.6
2021	12	IN	Indiana	9.6	3.9	56,934	28.9	90.6	36.30	59.74	16.4	18.7
2021	13	IA	Iowa	3.2	3.8	58,049	30.5	93.3	36.40	24.1	17.5	13.3
2021	14	KS	Kansas	6.4	3.3	58,569	35.4	91.9	36	18.42	19.4	18.1
2021	15	KY	Kentucky	9.6	4.4	51,561	27	88	40.30	49.23	17.9	23.8
2021	16	LA	Louisiana	21.3	5.6	54,531	26.4	86.7	38.60	43.05	14.8	25.7
2021	17	ME	Maine	1.7	4.6	58,687	36	94.5	31.90	11.69	19.5	10.6
2021	18	MD	Maryland	12.2	5.3	69,052	42.5	91.1	34.30	60.03	9.7	13.1
2021	19	MA	Massachusetts	2.3	5.5	83,461	46.6	91.1	27.40	37.59	8	6.1
2021	20	MI	Michigan	8.7	5.8	56,601	31.7	92	34.40	40.84	14.3	13.5
2021	21	MN	Minnesota	4.3	3.8	66,846	38.9	94.1	32.40	71.32	13.9	9.1
2021	22	MS	Mississippi	23.7	5.5	46,577	24.8	86.5	39.10	27.47	16.2	27.9
2021	23	MO	Missouri	12.4	4.1	56,073	31.7	91.6	37.30	67.22	18.7	18.8
2021	24	MT	Montana	4.4	3.4	58,344	34.8	94.4	31.80	26.64	32	13.2
2021	25	NE	Nebraska	3.6	2.7	62,682	34.4	92.2	35.90	18.09	15	15.1
2021	26	NV	Nevada	8.5	6.9	61,024	27.6	87.2	31.30	74.11	21.5	16.8
2021	27	NH	New Hampshire	0	3.4	72,214	40.2	94.4	30.60	14.52	15.1	6.6
2021	28	NJ	New Jersey	4.8	6.6	76,079	43.1	91	28.20	52.42	7.1	9.2
2021	29	NM	New Mexico	15.3	7.1	51,141	30.1	87.5	34.60	188.93	25	21.9
2021	30	NY	New York	4.8	7	75,948	39.9	88	29.10	43.5	7.9	10
2021	31	NC	North Carolina	9.7	4.9	56,705	34.9	89.7	36	61.58	13.2	17.3
2021	32	ND	North Dakota	3.4	3.1	65,895	31.7	93.6	35.20	24.02	20.8	13.7
2021	33	OH	Ohio	9.3	5.1	57,026	30.7	91.7	37.80	73.31	14.6	17.6
2021	34	OK	Oklahoma	8.9	4	55,165	27.9	88.7	39.40	44.44	22.1	25
2021	35	OR	Oregon	4.9	5.2	61,646	36.3	91.9	30.40	63.64	19.5	10.1
2021	36	PA	Pennsylvania	9.2	6	64,042	34.5	91.9	33.30	18.1	13.9	12.6
2021	37	RI	Rhode Island	3.6	5.5	63,663	36.5	89.1	30.10	25.31	10.3	9.4
2021	38	SC	South Carolina	13.4	3.9	52,828	31.7	89.6	36.10	50.74	15.2	19.3
2021	39	SD	South Dakota	5.3	2.6	65,421	31.7	93.1	38.40	24.44	23.2	18.7
2021	40	TN	Tennessee	12.2	4.5	56,970	30.5	89.7	35	72.57	17	23.3
2021	41	TX	Texas	8.2	5.6	60,548	33.1	85.4	36.10	78.66	14.2	22.4
2021	42	UT	Utah	2.7	2.7	57,042	36.8	93.2	30.90	34.38	20.1	10.8
2021	43	VT	Vermont	0	3.7	61,214	44.4	94.5	29	10.14	20.3	7
2021	44	VA	Virginia	7.2	3.9	66,838	41.8	91.4	34.20	33.84	13.2	13.1
2021	45	WA	Washington	4.5	5.2	74,188	39	92.3	28.80	75.18	15.3	15.6
2021	46	WV	West Virginia	6.9	5.1	49,071	24.1	88.8	40.60	15.37	20.6	22.5
2021	47	WI	Wisconsin	6.4	3.8	60,381	32.5	93.3	33.90	49.88	15.1	11.5
2021	48	WY	Wyoming	0	4.6	70,522	30.2	93.6	32	10.08	22.2	18.1

Appendix III: *R-Script*

```
# REGRESSION ANALYSIS...
# DEPENDENT: Homicide Rate
# INDEPENDENT: Unemployment, Income, Bachelor's Degree,
High School Degree, Obesity, Robbery, Suicide, Teen Pregnancy

# IMPORT LIBRARIES & FILE...
library(YRmisc)
library(readxl)
SocialData_1_ <- read_excel("Documents/SocialData (1).xlsx",
                    sheet = "data1")
sddf<-
SocialData_1_[,c("stateShort","homcideRate","unempRate","inc
ome","bachDegree","hsDegree","obesityRate","robberyRate","su
icideRate","teenPregRate")]
thesis<-sddf
View(thesis)

# GRAPHICAL ANALYSIS BEGINS HERE...
# FIGURES 1 - 9: Histograms
par(mfrow=c(3,3))
hist(thesis$homcideRate, col="red", xlab="Homicide Rate",
ylab="Frequency", main="Fig. 1 Hist of Homicide Rate")
hist(thesis$unempRate, col="orange", xlab="Unemployment
Rate", ylab="Frequency", main="Fig. 2 Hist Unemployment
Rate")
hist(thesis$income, col="yellow", xlab="Income",
ylab="Frequency", main="Fig. 3 Hist of Income")
hist(thesis$bachDegree, col="green", xlab="Bachelor's Degree",
ylab="Frequency", main="Fig. 4 Hist of Bachelor's Degree")
hist(thesis$hsDegree, col="blue", xlab="High School Degree",
ylab="Frequency", main="Fig. 5 Hist of High School Degree")
hist(thesis$obesityRate, col="purple", xlab="Obesity Rate",
ylab="Frequency", main="Fig. 6 Hist of Obesity Rate")
hist(thesis$robberyRate, col="pink", xlab="Robbery Rate",
ylab="Frequency", main="Fig. 7 Hist of Robbery Rate")
hist(thesis$suicideRate, col="cyan", xlab="Suicide Rate",
ylab="Frequency", main="Fig. 8 Hist of Suicide Rate")
hist(thesis$teenPregRate, col="brown", xlab="Teen Pregnancy
Rate", ylab="Frequency", main="Fig. 9 Hist of Teen Pregnancy
Rate")

# FIGURES 10 - 17: Scatterplots
par(mfrow=c(3,3))
scatter.smooth(thesis$unempRate, thesis$homcideRate,
xlab="Unemployment Rate", ylab="Homicide Rate", main="Fig.
10 Unemployment Rate vs. Homicide Rate", type="n")
  text(thesis$unempRate, thesis$homcideRate,
as.character(thesis$stateShort), cex=1)
scatter.smooth(thesis$income, thesis$homcideRate,
xlab="Income", ylab="Homicide Rate", main="Fig. 11 Income
vs. Homicide Rate", type="n")
  text(thesis$income, thesis$homcideRate,
as.character(thesis$stateShort), cex=1)
scatter.smooth(thesis$bachDegree, thesis$homcideRate,
xlab="Bachelor's Degree", ylab="Homicide Rate", main="Fig.
12 Bachelor's Degree vs. Homicide Rate", type="n")
  text(thesis$bachDegree, thesis$homcideRate,
as.character(thesis$stateShort), cex=1)
scatter.smooth(thesis$hsDegree, thesis$homcideRate,
xlab="High School Degree", ylab="Homicide Rate", main="Fig.
13 High School Degree vs. Homicide Rate", type="n")
  text(thesis$hsDegree, thesis$homcideRate,
as.character(thesis$stateShort), cex=1)
scatter.smooth(thesis$obesityRate, thesis$homcideRate,
xlab="Obesity Rate", ylab="Homicide Rate", main="Fig. 14
Obesity Rate vs. Homicide Rate", type="n")
  text(thesis$obesityRate, thesis$homcideRate,
as.character(thesis$stateShort), cex=1)
scatter.smooth(thesis$robberyRate, thesis$homcideRate,
xlab="Robbery Rate", ylab="Homicide Rate", main="Fig. 15
Robbery Rate vs. Homicide Rate", type="n")
  text(thesis$robberyRate, thesis$homcideRate,
as.character(thesis$stateShort), cex=1)
scatter.smooth(thesis$suicideRate, thesis$homcideRate,
xlab="Suicide Rate", ylab="Homicide Rate", main="Fig. 16
Suicide Rate vs. Homicide Rate", type="n")
  text(thesis$suicideRate, thesis$homcideRate,
as.character(thesis$stateShort), cex=1)
scatter.smooth(thesis$teenPregRate, thesis$homcideRate,
xlab="Teen Pregnancy Rate", ylab="Homicide Rate",
main="Fig. 17 Teen Pregnancy Rate vs. Homicide Rate",
type="n")
  text(thesis$teenPregRate, thesis$homcideRate,
as.character(thesis$stateShort), cex=1)

# FIGURES 10.1 - 17.1: Scatter plots - States with Outliers
# Drop states (outliers).
thesis_drop_outliers <- thesis
drop_MS <- "MS"
drop_LA <- "LA"
thesis_drop_outliers <- thesis[thesis$stateShort != drop_MS &
thesis$stateShort != drop_LA,]
View(thesis_drop_outliers)

par(mfrow=c(3,3))
scatter.smooth(thesis_drop_outliers$unempRate,
thesis_drop_outliers$homcideRate, xlab="Unemployment Rate",
ylab="Homicide Rate", main="Fig. 10.1 UR vs. HR", type="n")
  text(thesis_drop_outliers$unempRate,
thesis_drop_outliers$homcideRate,
as.character(thesis_drop_outliers$stateShort), cex=1)
scatter.smooth(thesis_drop_outliers$income,
thesis_drop_outliers$homcideRate, xlab="Income",
ylab="Homicide Rate", main="Fig. 11.1 IL vs. HR", type="n")
  text(thesis_drop_outliers$income,
thesis_drop_outliers$homcideRate,
as.character(thesis_drop_outliers$stateShort), cex=1)
scatter.smooth(thesis_drop_outliers$bachDegree,
thesis_drop_outliers$homcideRate, xlab="Bachelor's Degree",
ylab="Homicide Rate", main="Fig. 12.1 BD vs. HR", type="n")
  text(thesis_drop_outliers$bachDegree,
thesis_drop_outliers$homcideRate,
as.character(thesis_drop_outliers$stateShort), cex=1)
scatter.smooth(thesis_drop_outliers$hsDegree,
thesis_drop_outliers$homcideRate, xlab="High School Degree",
ylab="Homicide Rate", main="Fig. 13.1 HSD vs. HR", type="n")
  text(thesis_drop_outliers$hsDegree,
thesis_drop_outliers$homcideRate,
as.character(thesis_drop_outliers$stateShort), cex=1)
scatter.smooth(thesis_drop_outliers$obesityRate,
thesis_drop_outliers$homcideRate, xlab="Obesity Rate",
ylab="Homicide Rate", main="Fig. 14.1 OR vs. HR", type="n")
  text(thesis_drop_outliers$obesityRate,
thesis_drop_outliers$homcideRate,
as.character(thesis_drop_outliers$stateShort), cex=1)
scatter.smooth(thesis_drop_outliers$robberyRate,
thesis_drop_outliers$homcideRate, xlab="Robbery Rate",
ylab="Homicide Rate", main="Fig. 15.1 RR vs. HR", type="n")
  text(thesis_drop_outliers$robberyRate,
thesis_drop_outliers$homcideRate,
as.character(thesis_drop_outliers$stateShort), cex=1)
scatter.smooth(thesis_drop_outliers$suicideRate,
thesis_drop_outliers$homcideRate, xlab="Suicide Rate",
ylab="Homicide Rate", main="Fig. 16.1 SR vs. HR", type="n")
  text(thesis_drop_outliers$suicideRate,
thesis_drop_outliers$homcideRate,
as.character(thesis_drop_outliers$stateShort), cex=1)
scatter.smooth(thesis_drop_outliers$teenPregRate,
thesis_drop_outliers$homcideRate, xlab="Teen Pregnancy Rate",
ylab="Homicide Rate", main="Fig. 17.1 TPR vs. HR", type="n")
  text(thesis_drop_outliers$teenPregRate,
thesis_drop_outliers$homcideRate,
as.character(thesis_drop_outliers$stateShort), cex=1)

# STATISTICAL ANALYSIS BEGINS HERE...
# TABLE 1: DESCRIPTIVE STATISTICS
des_stats<-
ds.summ(thesis[,c("homcideRate","unempRate","income","bach
Degree","hsDegree","obesityRate","robberyRate","suicideRate","
teenPregRate")],3)[,-c(7,8)]
View(des_stats)

# TABLE 2: CORRELATION MATRIX
cor_matrix<-
round(cor(thesis[,c("homcideRate","unempRate","income","bach
Degree","hsDegree","obesityRate","robberyRate","suicideRate","
teenPregRate")]),3)
View(cor_matrix)

# TABLE 3: REGRESSION ANALYSIS
thesis1<-na.omit(thesis)
fit<-
lm(homcideRate~unempRate+income+bachDegree+hsDegree+o
besityRate+robberyRate+suicideRate+teenPregRate,na.action=n
a.omit,data=thesis1)
summary(fit)

# FIGURES 18 & 19: Fitted Residuals
#Vector; how much the computer missed each time...
thesis1$homcideRate
fit$fitted.values
fit$residuals
par(mfrow=c(2,2))
hist(fit$residuals,col="grey",xlab="Fitted
Residuals",ylab="Frequency",main="Fig. 18 Histogram of Fitted
Residuals")
plot(fit$residuals,xlab="Index",ylab="Fitted
Residuals",main="Fig. 19 Plot of Fitted Residuals",type="n")
text(fit$residuals,as.character(thesis1$stateShort),cex=1)
```

Copyright Notice

Thank you!

Questions & Correspondence

Brianna Palmisano
brianna.palmisano21@my.stjohns.edu