Daniel Stumpf
Econ 203 Semester Project
12/9/17

The effect of Geographical location on the Death Rate in the United States

1. Introduction

Man has always strived for knowledge, especially knowledge that would help them live longer. Today, one of the greatest troves of knowledge about survival is the National Vital Statistics Reports. These provide statistical data on a nation-wide scale, ranging in topics form causes of death, life expectancy, births, and more. But only so much can be extrapolated from data, meaning there are some gaps.

In this study I seek to analyze the main determinants of the death rate is the United States, measured by the hundred thousand persons. Specifically, I am interested in seeing the effects of geographical location on the death rate. To do so, I have gathered all my data on a state-by-state basis. I predict that the location will have a large effect, and Southern or Western states will have higher rates than those in the Midwest or North.

The United States already has large amounts of this data available to the public, most commonly available through the census. From the *National Vital Statistics Report: Deaths of 2013* I was able to gather data on all 50 states, with more than 22 different analyses of leading causes of deaths. I obtained the 3 factors along with the death rates, the dependent variable, from this report, and the other 5 from the actual 2013 Census Report. I ran a multiple regression analysis on these 8 variables against the death rate, as well as a simple linear regression test against regions prove the validity of my hypothesis.

While analyzing the original in its fullness, I came across many flaws in my data. Two of my factors were very highly correlated, which caused multicollinearity within the model and many of my factors were rendered invalid. There was also one major outlier that I had to remove from the full model. Upon further analysis in simple regression of each factor, it became evident that heteroscedasticity was prevalent in 4 of them, but this was easily remedied since all of these were too insignificant to be kept. All the histograms presented normality, and there was no time series data which meant no autocorrelation.

The results of my study prove different from what I expected. While the geographical location was significant in simple regression, in multiple regression it was rendered invalid. Other factors that were not significant to the death rate were the total population, total number of deaths, median household income, and if the state legalized marijuana. Factors that were significant included the violent crime rate, average life expectancy, and percentage of the population that was white. Despite these results, it became evident that there is a correlation between region and white percentage, violent crime rate, and life expectancy. Based on these results, it can be inferred that while region may not directly influence the death rates, the geographical location does influence its major factors.

2. Data Description

In this study my data was obtained from two different sources, the United States 2013 Census report as well as *National Viral Statistics Report: Deaths of 2013*. I was only able to use two sources because the data used in the NVSR was obtained from the census as well. Most of my data had to be obtained from different census reports, as they are not usually associated with death rates – such as the median household income and legalization of marijuana. I was also able to find this data categorized by state, making the full regression model easier to use. After categorizing everything by their respective states, outliers were already visible. I had to remove the data for the state of Alabama – it became a serious outlier in every scenario. The dependent variable, Death Rate (DR), was being compared to eight different variables, the total number of deaths (T#D), the total population (TP), the violent crime rates (VCR), average life expectancy (LE), if marijuana was legalized (MAR), the median household income (MHHI), white percentage of population (W%P), and the region located (REG).

Table 1 shows the descriptive statistics for the full model. This sample was originally made of 50 data points, but due to serious outliers the sample size decreased to 48. The average death rate was 725.8 per 100,000 people per state out of an average of 6,499,470 people per state. On average, people live to be 78.6 years old and the average state has 77% of their population to on identifying as white. *Table 2* is the correlation matrix between all the variables. It is evident that there is multicollinearity between Total Population per state and Total Deaths per state due to their value being |.95|, which exceeds the acceptable limit of |.8|**.** the descriptive statistics for the full model. This sample was originally made of 50 data points, but due to serious outliers the sample size decreased to 48. The average death rate was 725.8 per 100,000 people per state out of an average of 6,499,470 people per state. When looking at the plots of residuals vs. The predicted values, it is evident that there is heteroscedasticity. This was acceptable though because these variables were too insignificant to keep.

Figure one depicts the death rate vs each independent variable. It becomes evident that there are outliers here, as well as possible curvilinear relationships. Unfortunately, nothing arises from this due to multicollinearity.

3. Regression Analysis

In order to analyze the determinants of the death rate I estimated the linear regression with this initial model:

$$
\mathbf{DR_i} = \alpha + \beta_1 T\#D + \beta_2 TP + \beta_3 VCR + \beta_4 LE + \beta_5 E{\wedge}2 + \beta_6 AR + \beta_7 W\%P + \beta_8 MHHI + + \beta_9 REG + \beta_{10} REG{\wedge}2 + \varepsilon
$$

As mentioned earlier, the region is most likely an important determinant of the death rate, so it would possibly have a curvilinear relationship. This also applies to life expectancy, since it is very common for people of an older age to die than of a younger person. Since there were quite a few variables, I did not believe that these determinants would have a curvilinear relationship to a higher degree than squared. Despite total number of deaths and total population seeming to have strong correlations with the death rates, I believed that it would only be a linear relation. This is because when doing simple algebra, you can easily calculate a death rate using these two variables, one that would probably be similar to the one used for this study.

When thinking about societal norms and what influences death, some of the biggest associations are how old you are, what kind of family you come from, are you safe, ethnicity, etc. All these factors are greatly influenced by where you live. Regions of the south are “more dangerous” than areas in the north, and there are many dangerous factors that characterize the west. Also, area’s that are more densely populated, such as metropolises, will have higher crime rates and higher populations since there are more people. Due to this, it is respectably inferred that geographical location would have a great impact on the death rates.

The results for the initial multiple regression tests are displayed in *table 3*. This initial model had a great overall fit, with an r-squared value of .9331 and an F statistic of 71.429. This meant that the overall model was valid, but upon inspection some of the individual factor’s P value rendered them invalid at a 5% rate of error. The determinants that did not appear to be significant at first were T#D, TP, MAR, W%P, MHHI, and REG due to their P values being greater than 5%. For W%P, this was remedied through removing the significant outliers, reducing the p-value from .3490 to .0208. When inspecting the correlation table (*table* 2) it was evident that there was multicollinearity between T#D and TP, so T#D had to be removed.

Besides this evidence of multicollinearity and removed outliers *(figure 2),* the initial model passed the remaining test. *Figure 3* shows the histogram of standardized residuals there was no non-normality’s. There was also no time series data, meaning there was no autocorrelation prevalent. In the overall model, there was no heteroscedasticity. All this being accounted for, I was able to move on to making a reduced model.

To begin the reduction process, I eliminated T#D to rid of multicollinearity. As I eliminated variables, the other’s p-values shifted accordingly, but none shifted enough to be significant. After removing all the insignificant values, I went through each of them to check for curvilinear relationships. As show in figure, the death rate vs. median household income looked promising but even when straightening the graph, it was still insignificant in the model. This being the only possible curvilinear relationship, there was nothing else to do for the reduced model other than do a partial F test to determine if it should be used. Looking at *table 4*, the partial F test results were 3.18E-10, which easily passes this test, meaning the reduced model should be used. This model's r-squared value is .965, which is more accurate than the initial model, given that the initial model already a great fit. My final model ended up being:

$$\mathbf{DR_i = \alpha + \beta_3 VCR + \beta_4 LE + \beta_7 W\%P + \varepsilon}$$

4. Empirical Results

The coefficients in *Table 4* suggest that if the average violent crime rate decreases by 100 cases, the death rate will increase by 3%, or roughly by 27 out of every 100,000 people, keeping everything else constant. Considering that the average life expectancy dropped to 35 years old – in the case of a plague of other lethal killer – with all others held constant, the death rate would be increase by over 400% being over 3,000 deaths per 100,000.

Higher percentages of the population being white correlate to there being a lower death rate, but it is not that influential in the overall equation. The most influential of these statistics is the average life expectancy, with a coefficient of -54, which means that if the average life expectancy decreases, the death rate will increase.

I am surprised however that geographical region did not play a larger role in the death rate formula. However, these three determinates are all heavily influenced by the region of the United States they are in. Violent crime rate is alarming too, being a negative relationship – as crime rates decrease the death rate will increase. This could be possible due to the fact that there are many different forms of crime, and when applied to the real world more crimes like murder or arson could be occurring.

5. Summary and Discussion

This study explored the determinants in the average death rates in the United States on a state-by-state basis. I was interested in testing whether this empirical data and findings would predict the actual scenario and how it can be applied. Contrary to my original assertion, the geographical location one resides in does not play as big of a factor in determining the death rate, however life expectancy does. While life expectancy did not have a curvilinear relationship, it still has a great relationship such that – holding all else constant – for every year the life expectancy falls by a single year, the death rate will increase by 54 people.

Of course, this study was not without flaws. Ideally, if I would have gathered all the data myself, this would minimize the chance for translation error. This is unrealistic though, since this study was conducted on such a large scale. Another flaw was with my determinants. I narrowly selected factors that that I thought would have some correlation or would be significant in calculating the death rate. This meant I had a bias opinion and initially ruled out certain data that could have lead to having a more fit model. The largest flaw in my data though is probably the sample size, being only 48 data points. This severely bottlenecks the results, and having more data could have lead to a much stronger and more fit equation. I could have included territories of the US as well as other countries.

Figure 1: Graphs of Death Rate vs. Independent Variables

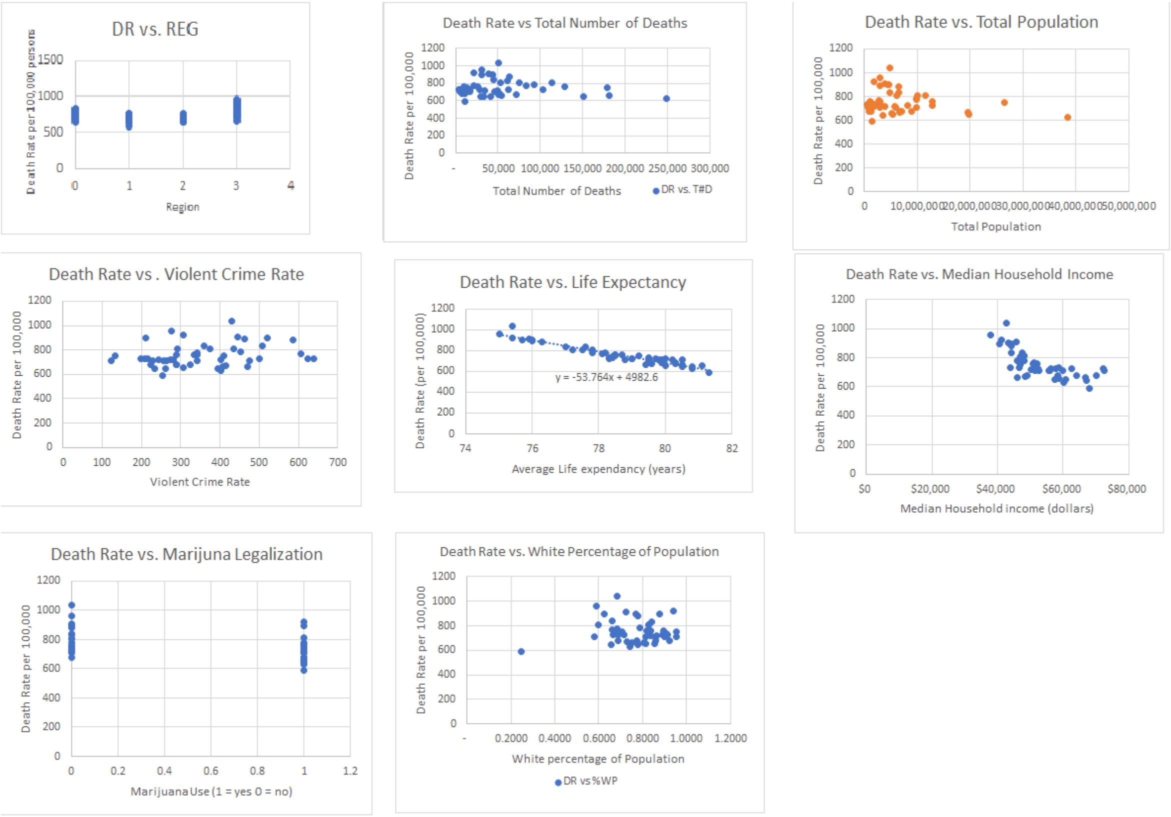


Figure 2: Initial data plot of standardized residuals

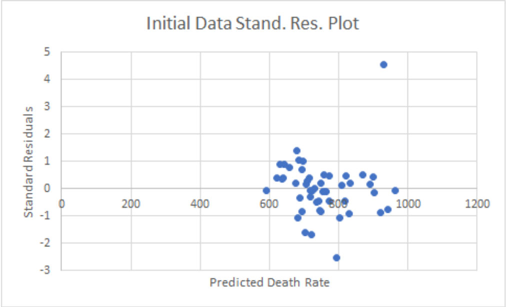


Figure 4: Initial data histogram of standardized residuals

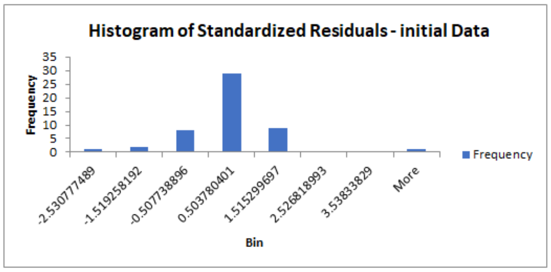


Figure 5: Reduced data plot of standardized residuals

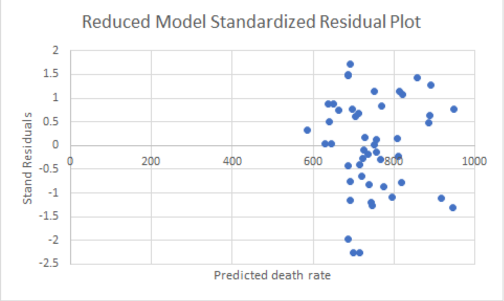


Figure 6: Reduced data histogram of standardized residuals

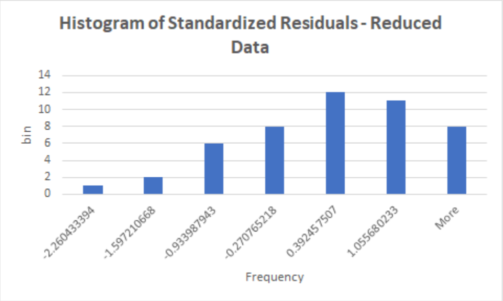
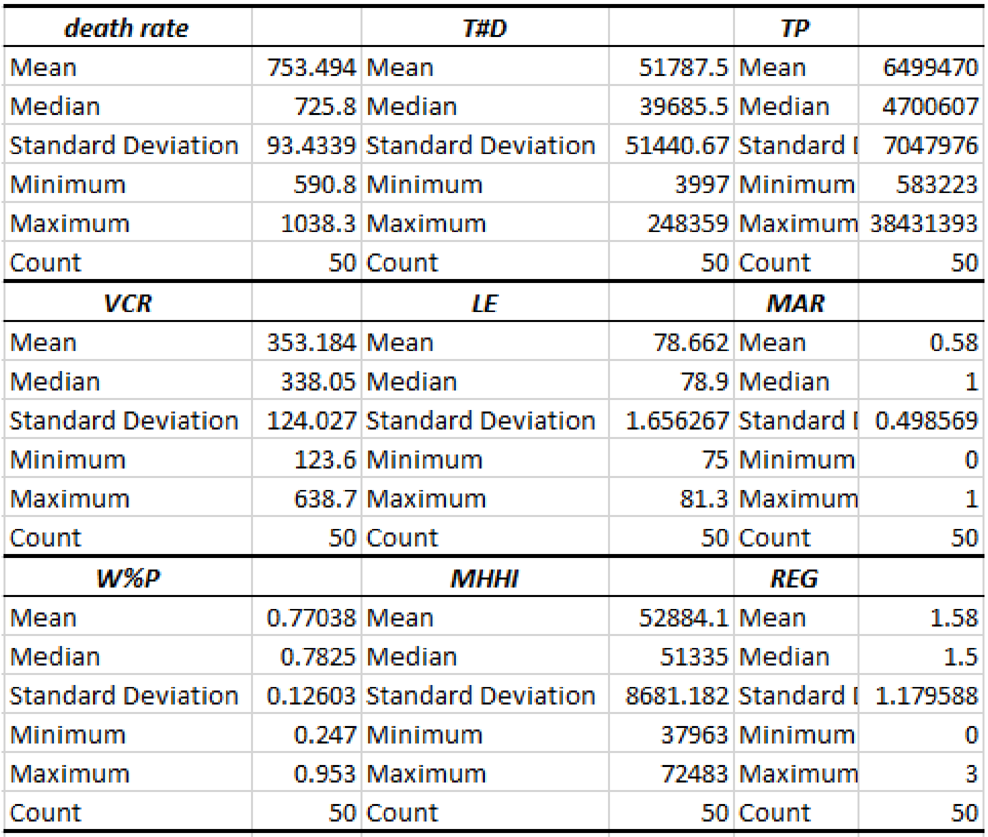
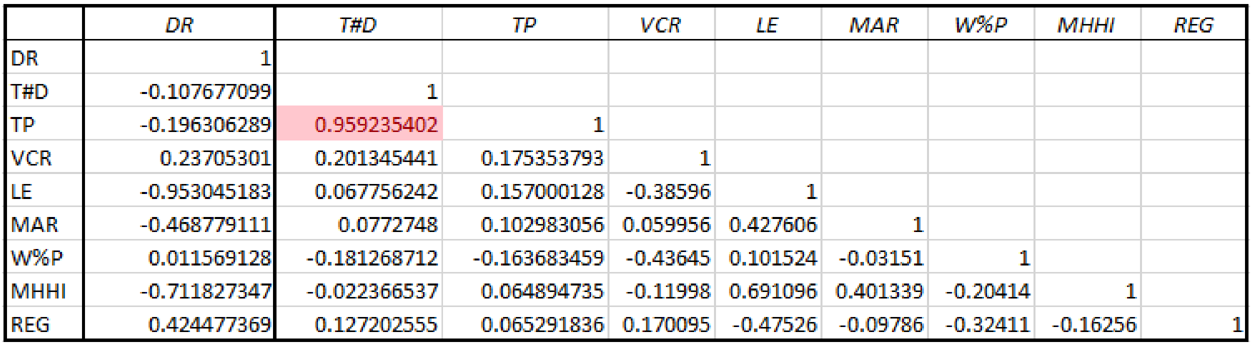


Table 1: Descriptive Statistics



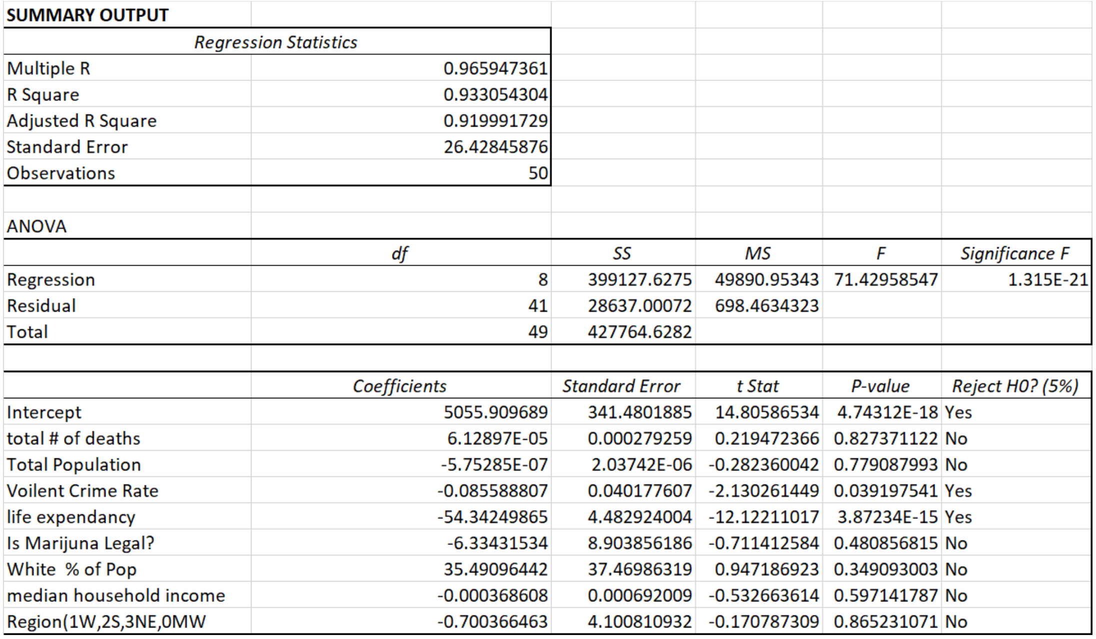
death rate		*T#D*		*TP*	
Mean	753.494	Mean	51787.5	Mean	6499470
Median	725.8	Median	39685.5	Median	4700607
Standard Deviation	93.4339	Standard Deviation	51440.67	Standard I	7047976
Minimum	590.8	Minimum	3997	Minimum	583223
Maximum	1038.3	Maximum	248359	Maximum	38431393
Count	50	Count	50	Count	50
VCR		*LE*		*MAR*	
Mean	353.184	Mean	78.662	Mean	0.58
Median	338.05	Median	78.9	Median	1
Standard Deviation	124.027	Standard Deviation	1.656267	Standard I	0.498569
Minimum	123.6	Minimum	75	Minimum	0
Maximum	638.7	Maximum	81.3	Maximum	1
Count	50	Count	50	Count	50
W%P		*MHHI*		*REG*	
Mean	0.77038	Mean	52884.1	Mean	1.58
Median	0.7825	Median	51335	Median	1.5
Standard Deviation	0.12603	Standard Deviation	8681.182	Standard I	1.179588
Minimum	0.247	Minimum	37963	Minimum	0
Maximum	0.953	Maximum	72483	Maximum	3
Count	50	Count	50	Count	50

Table 2: Correlation Statistics



	DR	*T#D*	*TP*	*VCR*	*LE*	*MAR*	*W%P*	*MHHI*	*REG*
DR	1								
T#D	-0.107677099	1							
TP	-0.196306289	0.959235402	1						
VCR	0.23705301	0.201345441	0.175353793	1					
LE	-0.953045183	0.067756242	0.157000128	-0.38596	1				
MAR	-0.468779111	0.0772748	0.102983056	0.059956	0.427606	1			
W%P	0.011569128	-0.181268712	-0.163683459	-0.43645	0.101524	-0.03151	1		
MHHI	-0.711827347	-0.022366537	0.064894735	-0.11998	0.691096	0.401339	-0.20414	1	
REG	0.424477369	0.127202555	0.065291836	0.170095	-0.47526	-0.09786	-0.32411	-0.16256	1

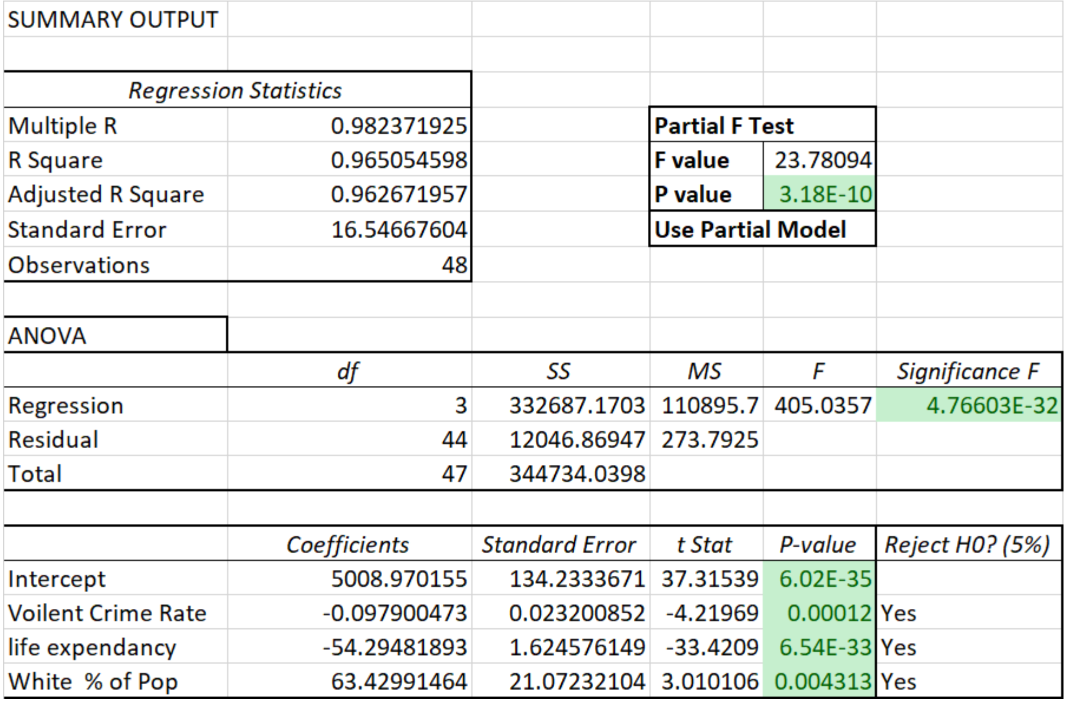
Table 3: Initial Multiple Regression Analysis



SUMMARY OUTPUT					
Regression Statistics					
Multiple R	0.965947361				
R Square	0.933054304				
Adjusted R Square	0.919991729				
Standard Error	26.42845876				
Observations	50				
ANOVA					
	df	*SS*	*MS*	*F*	*Significance F*
Regression	8	399127.6275	49890.95343	71.42958547	1.315E-21
Residual	41	28637.00072	698.4634323		
Total	49	427764.6282			
	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Reject H0? (5%)*
Intercept	5055.909689	341.4801885	14.80586534	4.74312E-18	Yes
total # of deaths	6.12897E-05	0.000279259	0.219472366	0.827371122	No
Total Population	-5.75285E-07	2.03742E-06	-0.282360042	0.779087993	No
Voilent Crime Rate	-0.085588807	0.040177607	-2.130261449	0.039197541	Yes
life expendancy	-54.34249865	4.482924004	-12.12211017	3.87234E-15	Yes
Is Marijuna Legal?	-6.33431534	8.903856186	-0.711412584	0.480856815	No
White % of Pop	35.49096442	37.46986319	0.947186923	0.349093003	No
median household income	-0.000368608	0.000692009	-0.532663614	0.597141787	No
Region(1W,2S,3NE,0MW	-0.700366463	4.100810932	-0.170787309	0.865231071	No

Note: Own calculations done based on data from US 2013 Census report

Table 4: Final Multiple Regression Analysis



SUMMARY OUTPUT					
Regression Statistics					
Multiple R	0.982371925		**Partial F Test**		
R Square	0.965054598		**F value**	23.78094	
Adjusted R Square	0.962671957		**P value**	3.18E-10	
Standard Error	16.54667604		**Use Partial Model**		
Observations	48				
ANOVA					
	df	*SS*	*MS*	*F*	*Significance F*
Regression	3	332687.1703	110895.7	405.0357	4.76603E-32
Residual	44	12046.86947	273.7925		
Total	47	344734.0398			
	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Reject H0? (5%)*
Intercept	5008.970155	134.2333671	37.31539	6.02E-35	
Voilent Crime Rate	-0.097900473	0.023200852	-4.21969	0.00012	Yes
life expendancy	-54.29481893	1.624576149	-33.4209	6.54E-33	Yes
White % of Pop	63.42991464	21.07232104	3.010106	0.004313	Yes

Note: Own calculations done based on data from US 2013 Census report