Is the Correlation between Suicide Rates and Other Socioeconomic Factors Still Strongly Prevalent?

A State Level Analysis

Tuan Anh Vu & Tung Truong
CS440 - Introduction to Artificial Intelligence
Spring 2020

Abstract

The overall purpose of the study was to determine if the risk factors such as unemployment, being a veteran, mental health, disability, poverty, insurance status which were prevalent in the past are still valid today. The hypothesis is that since suicide is still such a big topic today, all these independent variables are still siginificant and impact suicide rate. To confirm this hypothesis, a nonlinear regression was fitted to the data, and it showed that all risk factors in the past are still significant today. This led to the conclusion that the initiatives that have been implemented have not caused a significant impact in lowering suicide risk for at risk populations.

Introduction

Rationale and Objective

According to the National Institute of Health (NIH), suicide is the 10th leading cause of death in the United States. In 2017, there were 1,400,000 suicide attempts, 47,173 of these attempts were successful.[1] This is one of the reasons why suicide is a well studied subject. There are many scientific articles dating back to 1897 on the relationship between suicide and risk factors such as unemployment, mental health, etc.[2] The rationale behind this project is to see if the previous research inspired change such that the relationships between suicide and dependent factors like veteran status have decreased to a non-significant level. This paper contributes as a checkup to see if the factors implemented to help US citizens at risk of suicide are working.

Hypothesis

In order to understand the relationship between suicide and known risk factors such as unemployment, veteran status, mental health, disability status, and traumatic experience such as arrest rate, machine learning was used to analyze a set of time-series, state-level data to see if these still have any statistical significance to suicide. The hypothesis is that there still exist a correlation between suicide rate and all these factors.

[1] (https://afsp.org/about-suicide/suicide-statistics/).
[2] Durkheim 1897 “Suicide: A Study in Sociology”

Materials and Methods

This study consists of 50 states as well as the District of Columbia, and the observation period is 2011-2017. Table 1 lists and defines the variables used in this study:

Table 1. Variable definitions and sources

Variable	Definition	Source
Suicide Rate	Intentional self-harm (annual death per 100,000)	Center for Disease and Control Development
Unemployment Rate	The number unemployed[3] as a percent of the labor force.	U.S. Bureau of Labor Statistics
Veteran Percentage	Percent of the civilian population 18 years and over who are veterans	U.S. Census Bureau
Serious Mental Health Percentage	Percentage of population having a diagnosable mental, behavioral, or emotional disorder, other than a developmental or substance use disorder, based on the 4th edition of the *Diagnostic and Statistical Manual of Mental Disorders* (DSM-IV)	Substance Abuse and Mental Health Services Administration
Disability Percentage	Percentage of civilians in the noninstitutionalized population that has a hearing,vision,cognitive,ambulatory, self-care, and or independent living difficulty	U.S Census Bureau
Arrest Crude Rate	All crimes per their specific states laws (ranging from murder to curfew violations) excluding traffic violations (per 100,000)	Federal Bureau of Investigation
Percentage Uninsured	Percentage without health insurance coverage	U.S. Census Bureau
Poverty Rate	Percentage of households in poverty[4]	U.S. Census Bureau

[3] People are classified as unemployed if they do not have a job, have actively looked for work in the prior 4 weeks, and are currently available for work

[4] If a family's before-tax money income is less than the dollar value of their threshold, then that family and every individual in it are considered to be in poverty.

Data Collection

Data was collected over multiple years (2011-2017) in order to reduce the effect that time has as a variable, and the same logic was applied to location. Data was collected from the multiple states (and the District of Columbia) in order to reduce the effect of demographic factors. This creates a panel data set consisting of 357 observations per variable. This data set includes both time and state fixed effects that allows the data and its analysis to be attributed solely to the dependent variable .

Data Analysis Methodology

The official excel file was imported into R where the data was split into a training and test set, each with 75% and 25% of the data respectively. Then, each independent variable was plotted against the dependent variable to validate the simple linear regression assumption that the variables are linear in response. The trial and error method was then used to determine a non-linear variation of a covariate that is linear in response to the dependent variable.

Once all the variables passed or were modified to pass the linear in response assumption, a basic model was formulated as follow:

$$\begin{aligned} SuicideRates_{st} = Intercept_{st} + \beta_1 * AdjustedUnemploymentRate_{st} + \beta_2 * VeteranPercentage_{st} \\ + \beta_3 * CrudeArrest_{st} + \beta_4 * MentalIllnessRate_{st} + + \beta_5 * DisablePercentage_{st} \\ + \beta_6 * UninsuredRate_{st} + \beta_7 * Poverty_{st} + \beta_i * InteractionTerms_{st} + \varepsilon \end{aligned}$$

To arrive at the best model possible, the best subsets technique was used. Models using all possible combinations of the 7 variables were created, compared, and ranked by their adjusted R squared value. The model with the highest adjusted R squared value with all of its coefficients statistically significant was chosen.

To test the accuracy of this model, the 10-fold cross validation technique was implemented. The data was divided into 10 equal partitions where, reiteratively, each partition is predicted by the rest of the data. Then, the cross-validated error was calculated as the average of all the error terms of each partition. The Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC) were also deployed to test for the accuracy of the model. These are both penalized-based measurements that penalize the model if an additional variable is included. This is opposite to the adjusted R squared value, which generally increases whenever a variable is added to the model. Another way that the models were compared was by using the MAE, which stands for Mean Absolute Error which takes the mean of the absolute difference of the predicted and actual. The reason why the MAE value was used over the root mean squared error (RMSE) is because the MAE measurement is less subjected to outliers than the RMSE[5].

After having chosen the best model, interaction terms were incorporated in order to see how the independent variables affected one another, and if the relationship between the two independent variables affected suicide rate. A trial-and-error method was performed by adding

---

5 http://www.sthda.com/english/articles/38-regression-model-validation/158-regression-model-accuracy-metrics-r-square-aic-bic-cp-and-more/

interaction terms individually to the best model yielded from the best subsets and 10-fold cross validation. If an interaction term improves the model adjusted R squared while holding all coefficients statistically significant, then adding that term into the model will improve its accuracy. Else, the interaction term is invalid.

After finalizing all variables and their coefficients, the remaining tests to validate the assumptions for a linear regression model were carried out. This is an intensive error term analysis that consisted of testing for the independence of the error terms, whether the error terms follow a normal distribution, and for homoscedasticity. To reduce heteroscedasticity in the model, different models were built, namely, robust standard error regression and generalized least squares regression. Both methods involve adding a weight to each observation based on its residual variance from the found model.

Once all assumptions are tested, the model was used to predict the test data. The accuracy of the predictions were then measured by the mean, the standard deviation, and the mean absolute error of the residuals. The minimum and maximum suicide rates were also calculated for both the training data and the test data to ensure that the test observations are within range of the model.

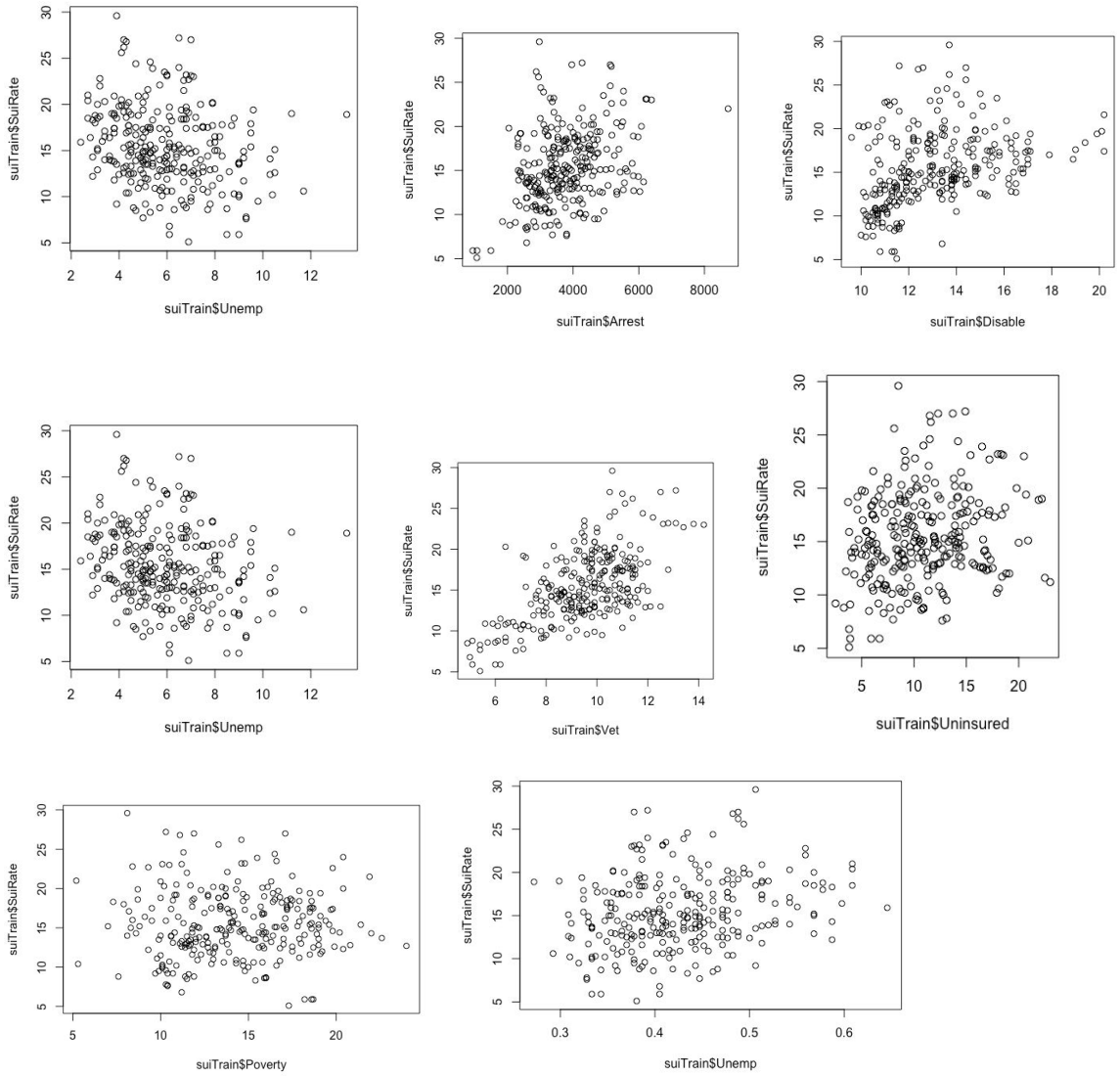
Table 2 lists the techniques used and their corresponding functions and packages in R:

Table 2. Techniques, functions, and packages used

Technique	Functions	Packages
Data splitting	createDataPartition()	caret
Linear Response Test	plot(dependentVar,independentVar)	graphics
Best-subset	regsubsets(...,method=”exhaustive”)	leaps
Cross-Validation	trainControl(...,method=”cv”)	caret
Error terms independence	plot(error_terms) bptest(best_model)	graphics lmtest
Error term normal distribution	hist(error_terms) qqnorm(error_terms) qqline(error_terms)	graphic stats
Test data prediction	predict(model, testdata) mean(), sd(), mae() min(), max()	stats
Other accuracy test	AIC(), BIC()	stats
Creating model	lm(dependentVar ~ independentVar, data)	stats
Robust standard error regression	lmrob()	robustbase
Generalized least squares	log(), exp() lm(...,weight = 1/sqrt(var))	stats

Results

Plotting each covariate with the dependent variable (crude suicide rate), the following graphs were produced:



Graphs. Unemployment rate, crude arrest rate, disable percentage, mental illness percentage, veteran percentage, percentage uninsured, poverty rate, and adjusted unemployment rate (left-to-right, top-to-bottom)

Out of the 7 covariate, only unemployment rate failed the linear response test as there was no linear relationship between unemployment rate and suicide rate in the first plot. Thus, through trial and error, the variable for unemployment rate was replaced with $1/sqrt(Unemp)$ which plot is shown as the last graph, indicating a linear relationship with suicide rate. For the rest of the paper the variable adjusted unemployment rate will refer to 1/sqrt(Unemp).

Table 3 lists the best model using each techniques and their accuracy/effectiveness measurements:

Table 3. Optimum models obtained from each regression and comparison method

Model	Formula	Adjusted R Sq	F Statistic	MAE	AIC	BIC
bestsubset_model	Unemp + Vet + Arrest + Mental + Disable + Uninsured	0.6163	73.02	2.0466	1309.11	1337.901
linear_model	Mental + Uninsured	0.313	62.27	2.8549	1462.49	1476.882
bestkfold_model	Unemp + Vet + Arrest + Mental + Disable + Uninsured	0.6191	97.43	1.9831	1723.39	1754.407
bestoverall_model	Unemp + Vet + Arrest + Mental + Disable + Uninsured + dxu	0.6361	68.18	1.9528	1285.80	1328.180

It can be seen from the table that the K-fold validation technique outputed the same model as the best subset technique. This confirmed that those set of independent variables best explains the dependent variable.

Once the combination of covariates is finalized, stata was used to test for the effect of adding an interaction term to the model. Figure 1 shows the output exported from data. Out of the 15 possible interaction terms[6], only by adding the interaction term between disable percentage and uninsured percentage that the coefficients of all other covariates remain statistically significant. This interaction term, recorded as dxu, was then added into the most recent best model where an increase in adjusted R squared was observed. Thus, the interaction term dxu was added into the model.

[6] The names of the interaction terms are the first letter of the 2 variables, separated by an x. It is also worthy to note that all the u that stands for the first variable represents unemployment while all the u that stands for the second variable represents uninsured rate. For example, uxa is the interaction term between unemployment and arrest while axu is the interaction term between arrest and uninsured rate.

	suirate	suirate	suirate	suirate	suirate	suirate	suirate	suirate	suirate	suirate	suirate	suirate	suirate	suirate	suirate	suirate
unemp	23.32***	19.00**	-19.68	-21.3	40.53**	13.71*	24.20***	23.53***	23.47***	24.19***	23.24***	22.33***	23.53***	22.95***	22.96***	24.01***
vet	1.022***	1.026***	1.093***	1.028***	1.787**	1.022***	0.579*	0.682	0.123	0.569*	1.006***	0.904***	1.024***	1.014***	1.034***	1.011***
arrest	0.000508**	0.0000194	0.000501**	0.000577***	0.000496**	0.000527***	-0.000633	0.000527**	0.000570***	0.000540***	0.00108	0.00309**	0.000387	0.000507**	0.000513**	0.000497**
mental	2.257***	2.259***	-2.053	2.256***	2.166***	2.297***	2.275***	1.533	2.242***	2.191***	2.737**	2.208***	2.249***	4.689***	2.749***	2.065***
disable	0.14	0.144	0.180*	-1.366**	0.148	0.124	0.171*	0.136	-0.582	0.165*	0.14	0.922**	0.142	1.098*	0.133	0.978***
uninsured	0.188***	0.192***	0.179***	0.192***	0.184***	-0.204	0.196***	0.186***	0.174***	-0.192	0.189***	0.186***	0.145	0.175***	0.388	1.157***
uxa		0.00114														
uxm			9.682**													
uxd				3.611***												
uxv					-1.835											
uxu						0.972										
vxa							0.000121									
vxm								0.0793								
vxd									0.0751							
vxu										0.0406*						
axm											-0.000127					
axd												-0.000193**				
axu													0.0000116			
mxd														-0.198		
mxu															-0.0469	
dxu																-0.0749***
_cons	-20.03***	-18.30***	-2.044	-1.732	-26.85***	-16.07***	-16.90***	-17.03**	-11.53*	-16.46***	-22.01***	-28.65***	-19.71***	-31.35***	-22.01***	-30.11***
N	357	357	357	357	357	357	357	357	357	357	357	357	357	357	357	357
="* p<0.05	** p<0.01	*** p<0.001"														

Figure 1. Interaction term analysis using stata

The model that had the best statistics for all the linear models was the model labeled linear_model. This was chosen when comparing all combinations of models that passed the linear assumptions. It can be seen that out of the 4 models, it had the worst regression, F statistic, MAE, and it was the third worst for AIC and BIC. However, out of all 4 models, it could be considered the most “valid linear model” since it passed all the assumptions.

The best model overall was the model named bestoverall_model. It was the model that contained the non-linear variable, *AdjustedUnemploytmentRate,* and the interaction term, *dxu*. This model had the highest R squared value while having the lowest AIC as well as BIC. This shows that even with the penalty imposed due to having 7 variables, it was the best model to explain suicide rate. It even had the best prediction rate due to the MAE being the lowest out of the four models. The subsequent analysis was done with this model.

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -3.009e+01  3.187e+00  -9.441  < 2e-16 ***
Unemp        2.287e+01  2.867e+00   7.978 4.66e-14 ***
Vet          1.087e+00  1.123e-01   9.678  < 2e-16 ***
Arrest       5.030e-04  1.751e-04   2.874 0.004391 **
Mental       1.904e+00  3.347e-01   5.690 3.39e-08 ***
Disable      1.036e+00  2.273e-01   4.557 7.96e-06 ***
Uninsured    1.114e+00  2.517e-01   4.428 1.40e-05 ***
dxu         -7.378e-02  1.887e-02  -3.911 0.000117 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.618 on 262 degrees of freedom
Multiple R-squared:  0.6456,    Adjusted R-squared:  0.6361
F-statistic: 68.18 on 7 and 262 DF,  p-value: < 2.2e-16
```

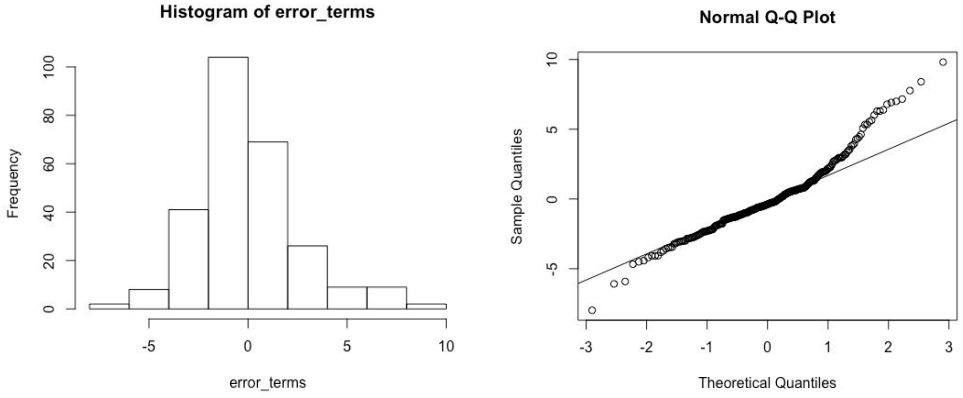
Figure 2. Best overall model summary

$$SuicideRates_{st} = -30.09 + 22.87 * AdjustedUnemploymentRate_{st} + 1.087 * VeteranPercentage_{st} + 0.000503 * CrudeArrest_{st} + 1.904 * MentalIllnessRate_{st} + 1.036 * DisablePercentage_{st} + 1.114 * UninsuredRate_{st} - 0.07378 * dxu_{st} + \varepsilon_{st}$$

The coefficients can be interpreted as follows:

- At the 0.1% significance level, the intercept coefficient indicates that if all other covariates are set to 0, the crude suicide rate will be -30 per 100,000 people[7]
- At the 0.1% significance level, for every 1% increase in adjusted unemployment rate, the crude suicide rate is expected to increase by 22.87
- At the 0.1% significance level, for every 1% increase in veteran percentage, crude suicide rate is expected to increase by 10.87
- At the 1% significance level, for every unit increase in arrest number, crude suicide rate is expected to increase by 0.000503
- At the 0.1% significance level, for every 1% increase in mental illness percentage, crude suicide rate is expected to increase by 1.90
- At the 0.1% significance level, for every 1% increase in disable percentage, crude suicide rate is expected to increase by 1.04
- At the 0.1% significance level, for every 1% increase in percentage uninsured, crude suicide rate is expected to increase by 1.11
- At the 0.1% significance level, there exists an interaction effect between disable percentage and percentage uninsured to crude suicide rates

Error analysis was then carried out on the best overall model with the results indicating independence and normal distribution of the error terms:



Graphs. Error terms histogram and Q-Q Plot

The best overall model was then used to predict the test dataset, which is 25% of the original dataset split at random. The accuracy of the prediction was measured by calculating

[7] Even though the number is unrealistic, having 0% unemployment, 0% veteran rate, etc are also only hypothetical and not realistic. This is also due to having omitted variable bias in the model.

mean absolute error (MAE), error mean and standard deviation, and the range of the observations in the training dataset versus in the test dataset. The following are the results from the error terms and prediction analysis:

$$MAE = 2.8549$$

```
> mean(suiTest$SuiRate - predict(best_model, suiTest))
[1] -0.04836894
> sd(suiTest$SuiRate - predict(best_model, suiTest))
[1] 2.568907
>
> # Determine observation range between training and test set
> min(suiTrain$SuiRate)
[1] 5.1
> max(suiTrain$SuiRate)
[1] 29.6
>
> min(suiTest$SuiRate)
[1] 6
> max(suiTest$SuiRate)
[1] 29.7
>
> mean(suiTest$SuiRate - predict(linear_model, suiTest))
[1] -0.2263172
> sd(suiTest$SuiRate - predict(linear_model, suiTest))
[1] 2.988327
>
```

Figure 3. Error terms and prediction analysis of the best overall model

Observing that the MAE number is relatively low, the mean of the prediction error is close to 0, and that the observations are in range, it is safe to conclude that the model used is relatively accurate.

Discussion and Limitations:

Discussion

The results using the best_model lead us to conclude that the factors in the past aside from poverty are still statistically significant and affect suicide rate today. This also could mean that that the initiatives that have been implemented have not caused a significant impact in lowering suicide risk for at risk populations. These results should be used to hold the government accountable and demand a further increase initiatives to help at risk populations lower their risk of suicide.

With roughly 40% of the model unaccounted for, there is a lot of room for further research to be conducted to further reduce the effect of omitted variable bias and the effect of time and location as a variable. At the time this paper was written, such specific data for all the covariates were not available.

Limitations

Suicide is a complex problem that deals with more than the 6 variables listed in our model. There exists omitted-variable bias that is caused by not including relevant variables in the model. An example is that the LGBTQ+ community are at higher risk of suicide, but this was not tested nor would this explain why the group is at higher risk. Gender and age information were not available for all used variables. These are also very important factors that should be taken into account because different genders and age groups have different responses to the independent factors. For example, there exists biological differences related to mental health and also gender inequality in the workforce.

This study is also limited to only 7 years worth of observations over 50 states (and the District of Columbia). If provided with a panel data that spans over a longer period of time, the model could better explain the correlation between suicide rates and the covariates by improving the time-fixed effect of the model. The information for mobility between states and other demographic data that could affect the model's covariates was also not readily available. If the study was provided with data at an individual level, the model could greatly improve by accounting for heteroscedastic caused from demographic factors.

Conclusion

Based on data gathered from all 50 states in America and the District of Columbia between 2011 and 2017, the best model was determined to contain unemployment rate, veteran percentage, disable percentage, mental illness percentage, crude arrest rate, and percentage uninsured. The best model roughly explains 60% of the trend for suicide, but there can still be further research. This study shows that there needs more work to be done on helping at risk communities than just the current initiatives. Let these findings be a call to action, so that these risks may be eliminated or reduced to a minimum in the future.

Bibliography

"Local Area Unemployment Statistics." U.S. Bureau of Labor Statistics, https://www.bls.gov/lau/home.htm

"Poverty Data Tables." U.S. Census Bureau, https://www.census.gov/topics/income-poverty/poverty/data/tables.html

"Reports and Data Tables." Subtance Abuse and Mental Health Services Administration, https://www.samhsa.gov/data/all-reports

"Underlying Causes of Death, 1999-2018." Centers for Disease Control and Prevention, https://wonder.cdc.gov/ucd-icd10.html

"Crime in the United States." Federal Bureau of Investigation, Criminal Justice Information Services Division https://ucr.fbi.gov/crime-in-the-u.s/2018/crime-in-the-u.s.-2018/topic-pages/persons-arrested

"Disability Characteristics." U.S. Census Bureau, https://factfinder.census.gov/faces/tableservices/jsf/pages/productview.xhtml?pid=ACS_17_5YR_S1810&prodType=table

"Percent of the Civilian Population 18 Years and Over who are Veterans." U.S. Census Bureau, https://factfinder.census.gov/faces/tableservices/jsf/pages/productview.xhtml?pid=ACS_17_5YR_GCT2101.US01PR&prodType=table

"Health Insurance." U.S. Census Bureau https://www.census.gov/topics/health/health-insurance.html

	suirate	suirate	suirate	suirate	suirate	suirate	suirate	suirate
unemp	23.32***	19.00**	-19.68	-21.3	40.53**	13.71*	24.20***	23.53***
vet	1.022***	1.026***	1.093***	1.028***	1.787**	1.022***	0.579*	0.682
arrest	0.000508**	0.0000194	0.000501**	0.000577***	0.000496**	0.000527***	-0.000633	0.000527**
mental	2.257***	2.259***	-2.053	2.256***	2.166***	2.297***	2.275***	1.533
disable	0.14	0.144	0.180*	-1.366**	0.148	0.124	0.171*	0.136
uninsured	0.188***	0.192***	0.179***	0.192***	0.184***	-0.204	0.196***	0.186***
uxa		0.00114						
uxm			9.682**					
uxd				3.611***				
uxv					-1.835			
uxu						0.972		
vxa							0.000121	
vxm								0.0793
vxd								
vxu								
axm								
axd								
axu								
mxd								
mxu								
dxu								
_cons	-20.03***	-18.30***	-2.044	-1.732	-26.85***	-16.07***	-16.90***	-17.03**
N	357	357	357	357	357	357	357	357
="* p<0.05	** p<0.01	*** p<0.001"						

	suirate	suirate	suirate	suirate	suirate	suirate	suirate	suirate
unemp	23.47***	24.19***	23.24***	22.33***	23.53***	22.95***	22.96***	24.01***
vet	0.123	0.569*	1.006***	0.904***	1.024***	1.014***	1.034***	1.011***
arrest	0.000570***	0.000540***	0.00108	0.00309**	0.000387	0.000507**	0.000513**	0.000497**
mental	2.242***	2.191***	2.737**	2.208***	2.249***	4.689***	2.749***	2.065***
disable	-0.582	0.165*	0.14	0.922**	0.142	1.098*	0.133	0.978***
uninsured	0.174***	-0.192	0.189***	0.186***	0.145	0.175***	0.388	1.157***
uxa								
uxm								
uxd								
uxv								
uxu								
vxa								
vxm	0.0751							
vxd		0.0406*						
vxu			-0.000127					
axm				-0.000193**				
axd					0.0000116			
axu						-0.198		
mxd							-0.0469	
mxu								-0.0749***
dxu	-11.53*	-16.46***	-22.01***	-28.65***	-19.71***	-31.35***	-22.01***	-30.11***
_cons	357	357	357	357	357	357	357	357
N								