

# Statistical Modeling Study of Crime Rate Changes due to Socioeconomic Factors

Barbara T. Sanders

*Los Alamos High School, Houston, TX*

## RESEARCH QUESTION

The subject of research we are going to pursue for regression analysis is based on finding the relationship between crime rate in the fifty states of United States of America and various other variables such as social, economic and demographic factors. Therefore the basic research question is as following:

“To what extent is crime rate of the year 2015 in different states of United States of America affected by changes in various social, economic and personal factors?”

As indicated in the question the data is based on the year 2015, 2014 and 2013 for all states.

Crime rate is a global prevalent issue where even the developed states have been effected by the increase in crime rates. We aim to study the model of a developed country such as of USA to determine which variables greatly contribute towards the increase in crime rate. Accordingly, we can determine what corrective and preventive measures can be taken to curb the increase in crime rate.

## 1. DATA COLLECTION

We used secondary sources to collect the data for crime rate, literacy rate, poverty level, unemployment, median household income, percentage of female population, percentage of male population, race and gender. The data was collected for three years (2013-2015) to increase the reliability and decrease the variability of the data. Reliable and authentic websites were used to obtain the data. The following are the list of variables that will be used in the analysis:

## 2. RESPONSE VARIABLE

1) Crime rate: The data available has categorized various types of crime committed in the states. To utilize the information we will take aggregate of all these different crime categories and express as the rate per 100

## 3. PREDICTOR VARIABLES

### Quantitative variables

1. Income level: The data of median income of the households in each state is available. We will convert the median income into percentage by expressing each states income as the proportion of total income of all the states.
2. Unemployment rate: The data available is expressed as a percentage of total number of people unemployed as proportion of the labor force. Since the data is already expressed as the percentage form it will be directly used in the analysis.
3. Poverty level: The data available is based on the percentage of population of each states following under federal poverty level as defined by Federal Government. The poverty threshold for 2015 was \$19,078. For our analysis, we will only consider one of the four brackets provided by data; the percentage of households in the states that fall below the poverty threshold (under 100% of poverty threshold)
4. Education level is defined in percentage in terms of high school graduates, bachelor’s degree and advanced degree. For our purpose of analysis we will use only the high school graduation rate.

### Qualitative variables

1. Gender: The data available segregates percentage of male and female in each state.
2. Race: The available data segregates the percentage of different races residing in each states. The classification is based on Whites, Blacks, Hispanics, Asians and others.

## 4. MODEL 1

First we ran the complete regression model and obtained the following results:

```

Residuals:
  Min       1Q   Median       3Q      Max
-2.3869 -0.7361 -0.0878  0.6112  4.3784

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  7.121e+01  6.239e+01  1.141  0.25564
POVERTY      4.810e-01  4.445e-01  1.082  0.28107
Unemployed   -9.402e-02  7.051e-02  -1.334  0.18448
Median       -1.398e-05  1.815e-05  -0.770  0.44247
LiteracyRate -2.383e+00  8.836e+00  -0.270  0.78774
PovLit       -5.918e-01  5.254e-01  -1.126  0.26185
GenderFM     -2.213e-02  2.535e-02  -0.873  0.38402
White        -7.885e-02  1.575e-02  -5.006  1.61e-06 ***
Black        -5.304e-03  1.704e-02  -0.311  0.75603
Hispanic     -1.039e-02  1.367e-02  -0.760  0.44821
Asian        -8.536e-02  3.089e-02  -2.763  0.00648 **
DummyYear2014 3.986e-02  2.621e-01  0.152  0.87933
DummyYear2015 -4.207e-01  3.433e-01  -1.226  0.22237
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.169 on 143 degrees of freedom
Multiple R-squared:  0.5341,    Adjusted R-squared:  0.495
F-statistic: 13.66 on 12 and 143 DF,  p-value: < 2.2e-16

```

### Fitted Equation of the model

$$\begin{aligned}
Y = & 71.21 + 0.481(X1) - 0.09402(X2) - 0.0001398(X3) \\
& - 2.383(X4) - 0.5918(X5) - 0.02213(X6) - 0.07885(X7) \\
& - 0.005304(X8) - 0.01039(X9) - 0.08536(X10) + 0.03986(X11) \\
& - 0.4207(X12)
\end{aligned}$$

Variables:

X1= POVERTY

X2= Unemployed

X3= Median

X4=LiteracyRate

X5=PovLit

X6=GenderFM

X7= White

X8= Black

X9= Hispanic

X10= Asian

X11=DummyYear2014

X12=DummyYear2015

53.41% variability in Y is explained by the model. We grouped the independent variables in various subsets and showed their effect on dependent variable through regression model. We removed an insignificant variable each time before rerunning the regression model and arriving at a higher value for the adjusted R2, the goodness of fit of the model.

## 5. MODEL 2

```

Residuals:
  Min       1Q   Median       3Q      Max
-4.8075 -0.9680 -0.0684  0.9017  5.8619

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.769e+01  2.918e+00  6.063  1.02e-08 ***
POVERTY      9.997e-02  4.576e-02  2.185  0.0305 *
Unemployed   -1.386e-01  6.637e-02  -2.088  0.0385 *
LiteracyRate -1.713e+01  2.545e+00  -6.729  3.33e-10 ***
Median       3.617e-05  1.868e-05  1.937  0.0547 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.394 on 151 degrees of freedom
Multiple R-squared:  0.3007,    Adjusted R-squared:  0.2822
F-statistic: 16.23 on 4 and 151 DF,  p-value: 4.443e-11

```

We used the economic variables along with literacy rate to assess the effect on the crime rate. According to the results obtained literacy rate has the greatest effect on crime rate. For each unit increase in literacy rate the crime rate is decreased by 17.13 units. The effect of economic conditions is lower than that of the literacy rate. The unemployment can't be accurately interpreted since the data shows that for each unit increase in unemployment rate, crime rate decreases by 0.01386 units. 30.07% in the variability of Y can be explained by the model.

## MODEL 3

```

Residuals:
  Min       1Q   Median       3Q      Max
-3.0379 -0.8110 -0.0144  0.8207  3.8729

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -11.04899   8.09380  -1.365  0.17426
GenderM      0.42430   0.15928  2.664  0.00857 **
white       -0.06038   0.01383  -4.366  2.35e-05 ***
Black       0.05014   0.01702  2.946  0.00374 **
Hispanic    0.01543   0.01280  1.206  0.22982
Asian      -0.06843   0.02937  -2.330  0.02115 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.275 on 150 degrees of freedom
Multiple R-squared:  0.4187,    Adjusted R-squared:  0.3994
F-statistic: 21.61 on 5 and 150 DF,  p-value: 2.97e-16

```

We used the demographic variables to assess their effect on the crime rate. It was observed that the percentage of male population has a higher effect on the crime rate. For every one unit increase in % of male population, 0.42430 unit increase in crime rate is observed. White and Asian population has a negative correlation with crime rate. An increase in White and Asian population leads to decrease in crime rate. 41.87% in the variability of Y can be explained by the model.

## CHECKING THE USEFULNESS OF THE MODEL (BY USING MODEL NUMBER 1)

Ho : X1=X2=X3=X4=X5=X6=X7=X8=X9=X10=X11=X12=0  
Ha Atleast one of the independent variable is not zero

F-TEST

$$R^2 = 0.5341$$

$$K = 12$$

$$N = 156$$

$$F \text{ STATS} = (R^2/K)/(1-R^2)/(N-(K+1))$$

$$F \text{ STATS} = (0.5341/12)/(1-0.5341)/(156-13) = 13.66$$

$$F(12,143) = 1.82$$

$$F \text{ STATS} > F(\text{ALPHA})$$

Therefore we reject  $H_0$  since the F stats lies in the rejection region. Thus we accept the  $H_a$  as at least one of the independent variable is none zero

## 6. LIMITATION OF THE ANALYSIS

1. The data for each and every state is not available due to which we have some inconsistency in our analysis. This reduces the reliability in our analysis however to counter this we took the data for three years.
2. Relationship between some variable is not found to be as per what was expected. The relation between gender and crime rate do not yield informative result. Similarly there is inverse relation between the unemployment and crime rate which questions the reliability of the data.
3. Sources of data although reliable may include biases in sampling. The data collected is obtained from secondary source thus the third party may have collected it from few selected sectors in a state.
4. The sample may not be true representative of the population of the state hence may not be reliable. This is a drawback of data and inconsistency.

## 7. CONCLUSION

From the above data we can analyze that almost all the independent variables were able to explain the dependent variable crime rate. However, we will fail to find a positive relation between the unemployment rate and crime rate. Similarly, the analysis of the relation between Gender and Crime rate did not yield any significant information and we conclude that gender play little role in explaining variability in Crime rate.

However, we were able to find inverse relation between median income and crime rate as well as between literacy and crime rate. This shows that states with better income level and education faced with lower crime rate. Similarly poverty had positive relation with crime rate that signifies that crime often results from inaccessibilities to basic necessities.

Our complete model one has  $R^2$  of 53.41 which is the highest among all the other. This signifies that by reducing the number of variables in other models  $R^2$  decreased (in some it remained same as well) which shows that complete model was able to explain the variability well then reduced models.

## REFERENCES

1. Crime Rate: <https://ucr.fbi.gov/crime-in-the-u.s/2015/crime-in-the-u.s.-2015/tables/table-4>
2. Income level: <http://kff.org/other/state-indicator/median-annual-income/?currentTimeframe=0&selectedDistributions=median-annual-household-income&selectedRows=%7B%22nested%22:%7B%22all%22:%7B%7D%7D-%22wrapups%22:%7B%22united-states%22:%7B%7D%7D%7D&sortModel=%7B%22colId%22:%22Location%22,%22sort%22:%22asc%22%7D>
3. Unemployment rate: <http://kff.org/other/state-indicator/unemployment-rate/?currentTimeframe=0&sortModel=%7B%22colId%22:%22Location%22,%22sort%22:%22asc%22%7D>
4. Education level: <http://www.census.gov/hhes/socdemo/education/data/cps/2015/tables.html>
5. Poverty Level: <http://kff.org/other/state-indicator/distribution-by-fpl/?currentTimeframe=0&sortModel=%7B%22colId%22:%22Location%22,%22sort%22:%22asc%22%7D>
6. Race: <http://kff.org/other/state-indicator/distribution-by-raceethnicity/?currentTimeframe=0&sortModel=%7B%22colId%22:%22Location%22,%22sort%22:%22asc%22%7D>
7. Gender: <http://kff.org/other/state-indicator/distribution-by-gender/?currentTimeframe=0&sortModel=%7B%22colId%22:%22Location%22,%22sort%22:%22asc%22%7D>