

Integrative Experience Project Report

Statistics 525

What is the Effect of Race, Education, Police Expenditure and Probability of Imprisonment on Crime Rate?

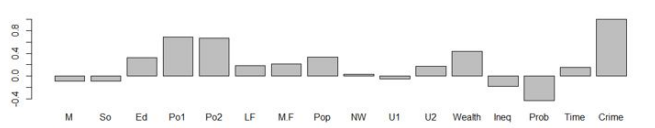
Introduction

The data set we have been working is derived from aggregated data from 47 states in 1960 and contains fifteen variables including population information, income, and wage disparities, racial information, education levels, work-force participation, number of offenses and time served. This rich data set provides us with a plethora of information that will help us to build our model and help answer our research question.

We would like to explore the effects of education, race, police expenditure and probability of imprisonment on crime rate in an area. We chose this research question because we are interested in how each of these variables affect crime which is a widespread problem. This project will give us insight and a better understanding of each variable and how it impacts crime.

Data

Predictors:	Pearson correlation between y and X(i):	Qualitative Variable
M	-0.0894724	Variables Used for our purposes
So	-0.09063696	
Ed	0.32283487	
Po1	0.68760446	
Po2	0.66671414	
LF	0.18886635	
M.F	0.21391426	
Pop	0.33747406	
NW	0.03259884	
U1	-0.05047792	
U2	0.17732065	
Wealth	0.44131995	
Ineq	-0.17902373	
Prob	-0.42742219	
Time	0.14986606	
Response(Y):	Where Y is "Crime" and X(i) changes	
Crime	1	



The data set contains the following columns:

Variable	Description
M	percentage of males aged 14-24 in total state population
So	indicator variable for a southern state
Ed	mean years of schooling of the population aged 25 years or over
Po1	per capita expenditure on police protection in 1960
Po2	per capita expenditure on police protection in 1959
LF	labour force participation rate of civilian urban males in the age-group 14-24
M.F	number of males per 100 females
Pop	state population in 1960 in hundred thousands
NW	percentage of nonwhites in the population
U1	unemployment rate of urban males 14-24
U2	unemployment rate of urban males 35-39
Wealth	wealth: median value of transferable assets or family income
Ineq	income inequality: percentage of families earning below half the median income
Prob	probability of imprisonment: ratio of number of commitments to number of offenses
Time	average time in months served by offenders in state prisons before their first release
Crime	crime rate: number of offenses per 100,000 population in 1960

Table 1: Results of Pearson's Correlation.

As mentioned above, this data set only includes data from 47 states. While looking at this data set we were a little concerned as to why only 47 states were included and how that might affect our results. As we did

not have any more information on how the data was gathered, we couldn't really tell which states were excluded and their effects on the analysis.

While choosing our predictor variables we referred to the Pearson Correlation Coefficient for all the X variables. This coefficient measures the strength of association between our chosen Y, crime rate, and all the other variables. To see which variables were correlated with the response variable, crime rate(Y), the Pearson Correlation between each variable and Y was calculated and the results are summarized in Table 1. From that table it can be noted that Po1, which is per capita expenditure on police protection, has the highest coefficient-it has a value closest to one. This means that there is some variation around the line of best fit. So, if a variable has a coefficient closer to 1 or -1, it is a good predictor for the response variable because it has less variation: a higher coefficient means that more of the variability we see in Y can be explained by its covariates(X1, X2 , X3 , X4). The next variable that would be a good predictor is Prob, which is the probability of imprisonment. From the table, you can see that crime rate and Prob have a negative correlation. Following that is Ed, which is the mean years of schooling. Finally, we also chose NW, which is the percentage of non-whites in the population. This is an exception from the rest because we, as a group, also wanted to look at race, as race and crime are known to be closely related. However, as its Pearson's Correlation coefficient was 0.0326, we realized we might have to transform it. With this information in mind, we started to look at scatterplots.

Data Analysis

Figure 1 - Crime Rate Distribution

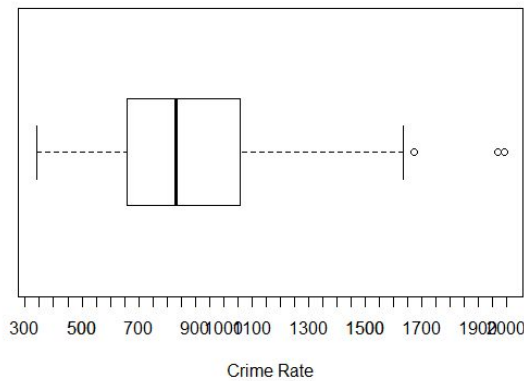


Figure 2 - Percentage of Non-White Distribution

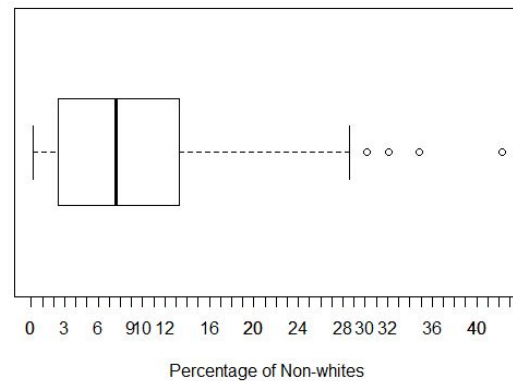


Figure 3 - Probability of Imprisonment Distribution



From the above box plots, we can observe the approximate minimum, maximum and medium values presented through the data provided as well as determine whether there are any outliers in our data which we need to worry about. Through the box plot of our response variable(Y), crime rate (Figure 1) we can observe the three outliers in this set of data can be taken from the 4th, 11th and 26th rows of the data set and can be determined as 1969, 1674 and 1993 respectively. From our first independent variable (X_1), contains no outliers. Figure 2 shows us the box plot of the percentage of non-whites (X_2) in the area with four outliers which correspond to the 1st, 16th, 23rd and 37th rows of the data set being, 30.1, 32.1, 42.3 and 34.9 respectively. The X_3 distribution shows us that per capita expenditure for police protection has no outliers. And finally, Figure 3 shows the distribution of the probability of imprisonment which gives us two outliers taken from rows 18, 22 and 42 of the dataset that correspond to the values 0.119804, 0.089502 and 0.088904, respectively.

Figure 4 - Crime Rate and Education Years

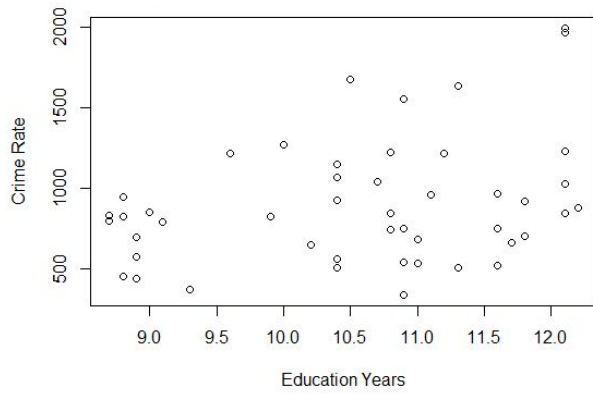


Figure 5 - Crime Rate and Percentage of Non-Whites

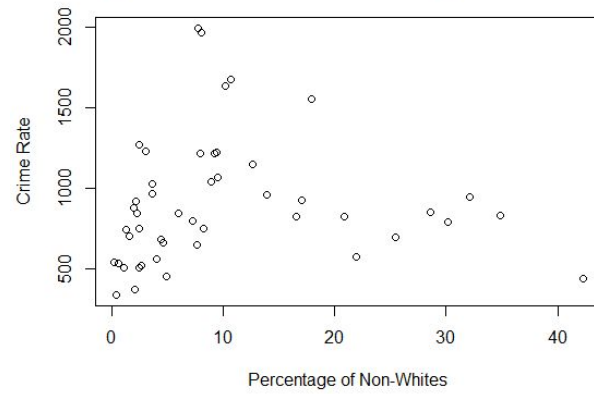


Figure 6 - Crime Rate and Percentage of Non-Whites

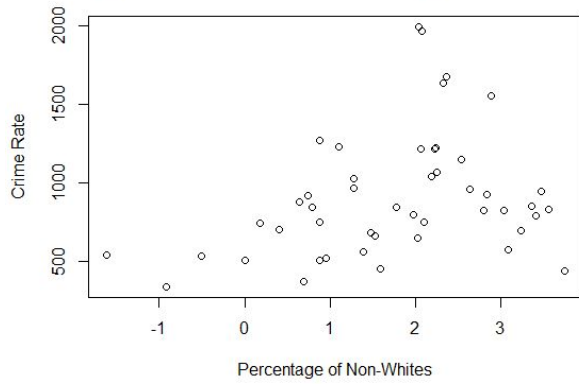


Figure 8 - Crime Rate and Probability of Imprisonment

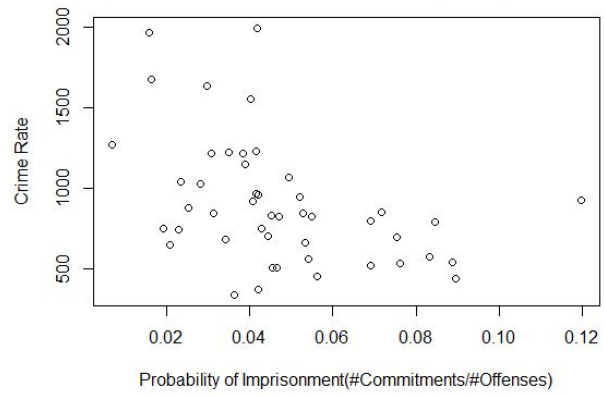
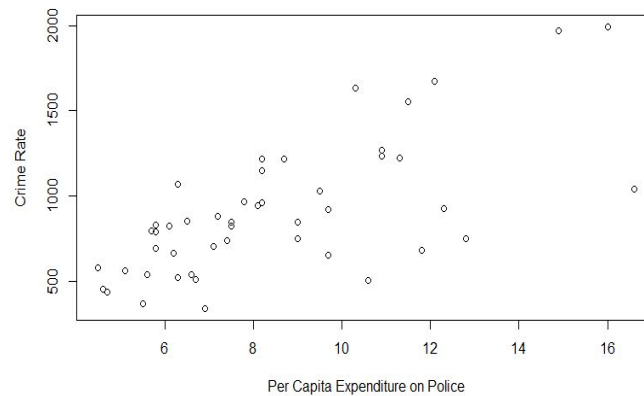


Figure 7 - Crime Rate and Per Capita Expenditure on Police Protection in 1960



The scatterplots presented above show us the linear relationships between our dependent variable, crime rate (Y) and our independent variables, years of education (X_1), percentage of non-whites (X_2), per capita expenditure on police protection (X_3) and the probability of imprisonment (X_4). When comparing these scatter plots we are able to see that the plot expressing the relationship between crime rate and education has a linear regression with a lesser slope which indicates the relatively small effect education has on crime rate. The data provided also tell us that even though the effect may be small there is an upward slope which shows us that crime rate increases with education. When considering the population of non-whites our initial scatterplot did not show a linear relationship, therefore we transformed the X variable into natural log form and were able to obtain a positive linear relationship between the crime rate and the log of the population of non-whites which shows that the crime rate increases along with the population of non-whites in the community. Through figure 7 we observe a linear relationship between the crime rate and the per capita expenditure on police protection showing us that with the increase in crime rate there is also an increase in expenditure towards police protection. The relationship depicted in figure 8, between the crime rate and the probability of imprisonment shows us a relatively linear negative relationship which may be slightly skewed due to the nature of its outliers, showing us that the crime rate decreases with the increase of the probability of imprisonment.

Figure 9-Percentage of Non-Whites and Education in Years

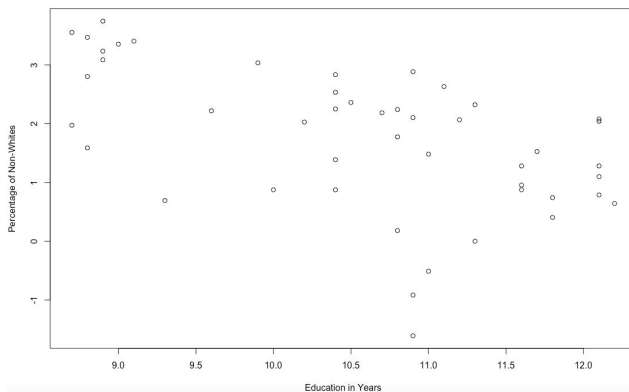


Figure 10 - Percapita expenditure on police protection and Education in Years

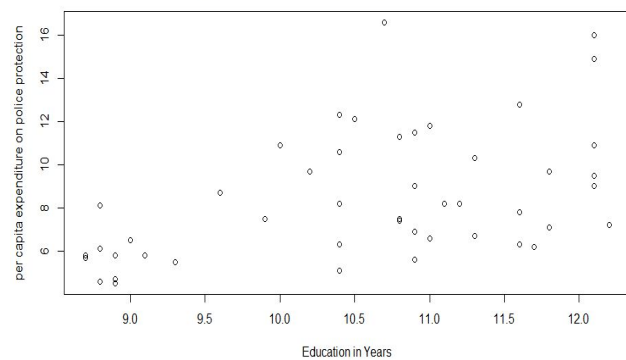
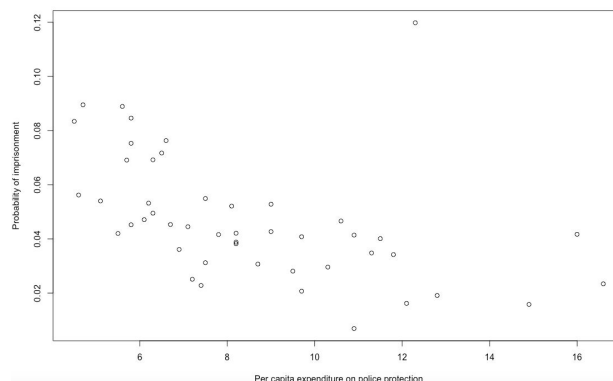


Figure 11-Probability of imprisonment and Per capita expenditure on police protection



When considering the relationships amongst the independent variables we are able to identify a very slight negative but no clear relationship between the log percentage of non-whites and the years of education (Figure 9) . We are also unable to observe a clear relationship between the years of education and the probability of imprisonment, the log percentage of non- whites and the probability of imprisonment or the per capita expenditure on police protection and the log percentage of non-whites. Although, we can observe a weak positive correlation between the years of education and the per capita expenditure on police protection (Figure 10) and a relatively inverse relationship between the probability of imprisonment and the per capita expenditure on police protection(Figure 11).

Figure 12 - Crime Rate and Education Years

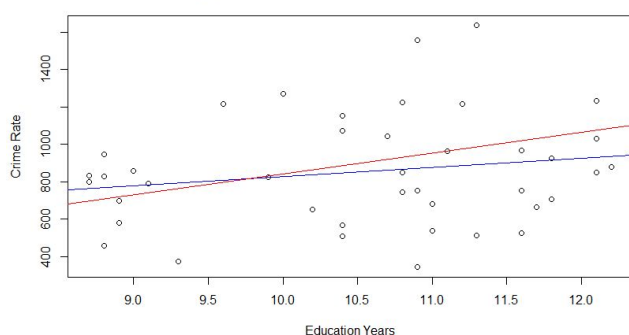


Figure 13 - Crime Rate and Percentage of Non-Whites(Transformed)

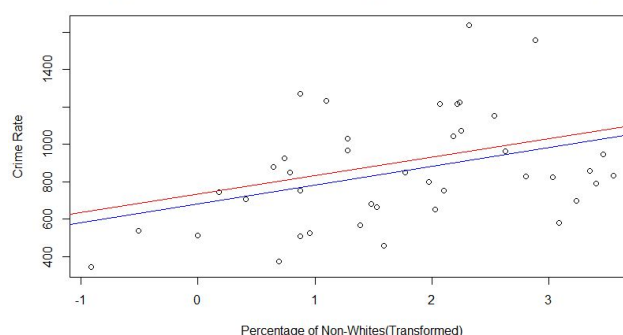


Figure 14 - Crime Rate and Per Capita Expenditure on Police Protection in 1960

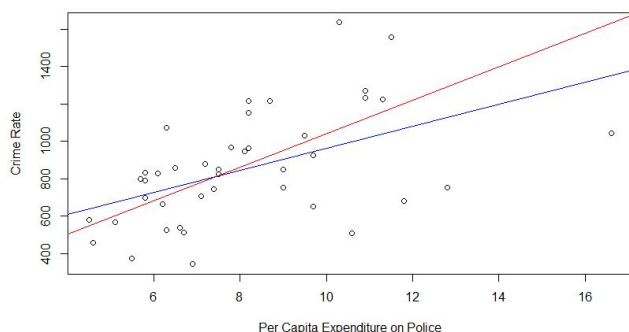


Figure 15 - Crime Rate and Probability of Imprisonment

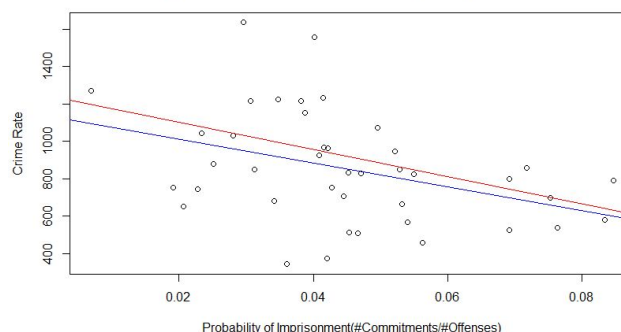


Figure 12 above visualizes the linear regression between our first predictor variable X_1 (mean years of schooling for the population aged 25 and above/education) and the response variable Y , crime rate with the altered data frame that omitted outliers. The linear regression line in blue is for the altered data frame and the red line is for the original regression between Y and X_1 with no changes. The obtained linear regression equation with the new data frame(blue) is $Y=48.94 \cdot X_1 - 338.32$. This is a positive relationship: higher crime rates are associated with higher mean years of schooling. The slope of the linear

regression on the original data frame(red) is observed to be more than twice as large as the slope of the linear regression on the altered data frame.

Figure 13 above visualizes the linear regression between our second predictor variable X_2 (percentage of non-whites in the population) and the response variable Y , crime rate with the altered data frame that omitted outliers. The obtained linear regression equation with the new data frame(blue) is $Y=101.24*X+679.63$. From previously constructed plots, we can see that the original data of X_2 needed to be transformed since the relation was not linearized. The transformed Y vs X_2 regression line with the original data frame is highlighted in red. This is also a positive relation: higher crime rates are associated with a higher log of percentage of non-whites in the population.

Figure 14 visualizes the linear regression between our third predictor variable X_3 (per capita expenditure on police protection) and the response variable Y , crime rate with the altered data frame that omitted outliers. The obtained linear regression equation with the new data frame(blue) is $Y=58.95*X_3+373.40$. We see from the plots that there is a relationship between the two variables and that the slope of the linear regression on the original data frame(red) is larger than the slope of the linear regression on the altered data frame. This is a positive relationship, thus showing that a higher crime rate is also associated with a higher per capita expenditure on police protection.

Figure 15 visualizes the linear regression between our fourth predictor variable X_4 (probability of imprisonment) and the response variable Y , crime rate with the altered data frame that omitted outliers. The obtained linear regression equation with the new data frame(blue) is $Y=-6363.9*X_4+1139.3$. There is a negative linear relationship between the two variables: as crime rate increases, the probability of imprisonment decreases. Furthermore, the previously obtained slope of the linear regression on the original data frame(red) is more negative than the slope of the linear regression on the altered data frame.

From these analyses, we can see that overall the slope of the red regression line is either more positive or more negative than the blue line. This could indicate that the outliers may have skewed the data/regression line largely, pulling the data points to it. Thus, the observed linear regression line for the altered data could actually be a better representation of the true relation. However, it is key to note that we actually do not know if these outliers should actually be omitted from the data, and, since we have no information on how the data was collected, we simply aim to provide further analysis of the relationship between the variables by doing so.

Linear Regression Assumptions

Figure 16-Multiple Regression(Original Data Frame) Residual vs Fitted Values Plot

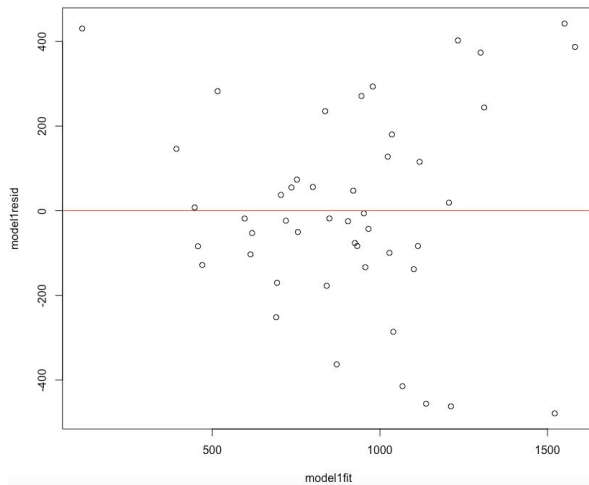
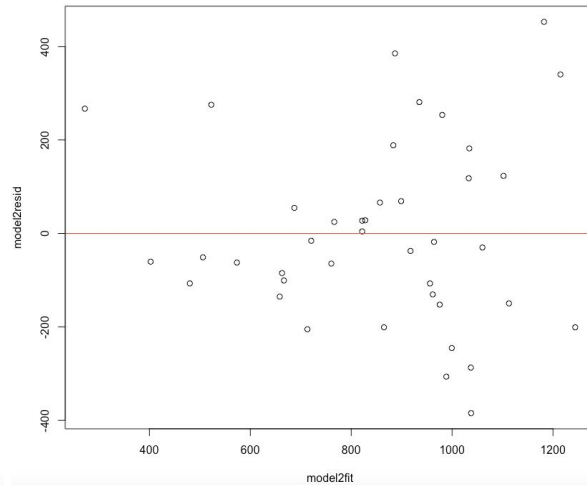


Figure 17-Multiple Regression(Outliers Omitted) Residual vs Fitted Values Plot



For the linear regressions that were performed(from both data frames), the residual plots, residuals against fitted values plots, and normal probability plots were each constructed to check the assumptions of linear regression and to check the dataset. In order to use linear regression on a dataset, the assumptions of linearity, homoscedasticity, independence and normality must be met. The data must show a linear pattern and each data point must be independent of another data point. We assumed that the data was independent and the constructed scatter plots after transformations displayed evidence of linear relations. The residual plots for each of the independent variables also showed no apparent pattern and seemed to be distributed evenly. Finally, the assumption for normality was checked using normal probability plots of the residual values which showed a clear linear relationship thus, indicating that the data and its errors were normally distributed.

The multiple regression residual against fitted values plot with the original data frame(Figure 16) showed fairly even distribution and constant error variances near the center of the plot but the error variances seemed to increase as the data points near the extreme values on the x-axis were further from the line $y=0$. The multiple regression residual against fitted values plot with the new data frame(Figure 17) could be classified as evenly distributed but could also display slight heteroscedasticity as some data points on the positive extreme of the x-axis have increased error variances.

Multiple Regression Model and Hypothesis Testing

Throughout this project, we ran two main regression models. Both of our multiple regression models were of the form $Y = \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_n X_n + \epsilon_i$ where Y is our dependent variable, X_i is our independent variable and b_i is our estimate of beta 1.

The first regression model did not exclude outliers while the second one did. Before executing the first regression model, a correlation matrix was made to see the relationship between the variables. In our correlation matrix, it was seen that Y , which is crime rate, had positive moderate correlations with the independent variables education(X_1), log of percentage of non-whites(X_2), and police expenditure(X_3) but had a negative correlation with probability of imprisonment(X_4). This gives us a basis as to how our linear regression function will look. It is worth noting that education has a moderate negative correlation with the log of percentage of non whites, telling us that there may be some racial inequality in play.

The first regression model's equation is as follows:

$$E\{Y\} = -794.67 + 104.9X_1 + 157.09X_2 + 55.95X_3 - 3348.16X_4$$

Where Y is crime rate, X_1 is education, X_2 is the log of percentage of non-whites, X_3 is per capita expenditure on police protection and X_4 is the probability of imprisonment. The regression model shows us that a one unit increase in X_1 , X_2 and X_3 causes crime rate to increase by 104.9, 157.09 and 55.95 respectively, given that the other variables are held constant. These variables are statistically significant at the 95% confidence level. However, a one unit increase in probability of imprisonment decreases crime rate by 3348.16, holding other factors constant. This variable has the biggest effect on crime rate due to the magnitude of its coefficient but it should be noted that it is only statistically significant at the 90% confidence level. This tells us that even though it has a larger magnitude than the other variables, it is not as significant as the others. The R-squared for this model is 0.6289 which means that the model explains 62.89% of the variability of the response data. This is a moderately high value for R-squared which means our model is a good fit for our variables.

The second regression has the same coefficients and data as the first model except that outliers were removed from each variable. The correlation matrix shows us the same information as the first correlation matrix in terms of direction of magnitude (positive or negative) of variables. However, there is a shift in the magnitude of all the variables. The change in the magnitude of variables is attributed to the removal of outliers as these outliers changed the mean, standard deviation, mean squared error etc., of the variables they were affecting. The second regression model's equation is as follows:

$$E\{Y\} = -576.72 + 111.73X_1 + 195.82X_2 + 19.435X_3 - 5353.79X_4$$

The new regression model has some significant changes when compared to the first regression model. The intercept increased from -794.67 to -576.72 which was expected as it is a new regression function. The coefficients of education and the log of percentage of non-whites have changed but not significantly. The two biggest changes seen are the changes of the coefficients of per capita expenditure (X_3) and probability of imprisonment (X_4). Per capita expenditure on police protection decreased from 55.95 to 19.435. An interesting observation is that per capita expenditure which was originally statistically significant at the 99% confidence level is no longer statistically significant at any level in this second model, including the 90% confidence level. This tells us that the removal of outliers had a significant impact on per capita expenditure, causing it to be no longer statistically significant. Another important change is in the coefficient for probability of imprisonment, which significantly decreased from -3348.16 to -5353.79. However, the most notable part is that this coefficient increased from the 90% confidence level to the 95% confidence level, making it in fact, more statistically significant. The R-squared of the second model is 0.5683, which tells us that this model explains 56.83% of the variability of the response data.

We tested to see if we should still include education in our regression model. Our null hypothesis was that the coefficient of education is 0 and our alternate hypothesis is that the coefficient of education is not equal to 0. We ran this test at the 95% confidence level. Using an ANOVA table, the F value for our test was 2.8399 and the obtained F quantile value with 4 and (n-5) degrees of freedom for alpha level 0.05 was 2.633532. Since our value is greater than the obtained value, we reject the null hypothesis and accept our alternate hypothesis that the coefficient of education is not 0. We should include it in our model.

Conclusion

In conclusion, after considering the Pearson Correlations, which determine the strength of the correlation between our Y variable (Crime rate) and each of the X variables given in the data, we were able to narrow down which variables we would explore in this report. Through this, we found that the most suitable predictors in our analysis would be Po1(per capita expenditure on police protection), Prob(the probability of imprisonment) which gives us a negative correlation, Ed(mean years of schooling and finally, NW(percentage of non-whites).

The first regression model's equation is as follows:

$$E\{Y\} = -794.67 + 104.9X_1 + 157.09X_2 + 55.95X_3 - 3348.16X_4$$

The second regression model's equation is as follows:

$$E\{Y\} = -576.72 + 111.73X_1 + 195.82X_2 + 19.435X_3 - 5353.79X_4$$

The first regression model did not exclude outliers, the second regression model did. After analyzing the predictor variable X_1 , it can be noted that there is a small, positive linear relationship between X_1 and Y , implying that higher crime rates are associated with more mean years of education. The data for predictor variable X_2 and its relationship with Y , reveals that higher crime rates are also associated with a higher percentage of non-whites in the population. The data for the third predictor variable X_3 and the response variable Y , shows that there is a positive relationship between the two variables, thus, indicating again that a higher crime rate is also associated with a higher per capita expenditure on police protection. Both regression models show us that a one unit increase in X_1 , X_2 and X_3 causes crime rate to increase by (104.9, 111.73), (157.09, 195.82) and (55.95, 19.435) respectively, given that the other variables are held constant. The data for predictor variable X_4 and the response variable Y indicates that there is a negative linear relationship between the two. Thus, as crime rate increases, the probability of imprisonment decreases. However, a one unit increase in probability of imprisonment decreases crime rate by (-3348.16, -5353.79) holding other factors constant.

The multiple regression residual against fitted values plots with the original data frame (Figure 16) and new data frame (Figure 17) both showed an even distribution and increases in their error variance as data points neared the extreme values on the x-axis.

However, because of previously identified gaps in information, we cannot conclusively say that Y and its relationship with the predictor variables X_1 , X_2 , X_3 , X_4 confirms or signifies these associations. For example, the analyses for three of the four predictor variables involved the process of omitting outliers. And when comparing the obtained regression model, most cases showed that omitting outliers makes noteworthy changes when compared to the untouched regression model. Since we do not know if these outliers should actually be omitted from the dataset, as we have no information on how the data was collected, further research relating to the importance of these outliers should be conducted.

Furthermore, our study is also limited to the relationships between the predictor variables and crime rate, for only 47 states of the USA in 1960. Any sort of suggestion of policy reform, or its importance for practice and subsequent research, must be performed critically and hesitantly to account for the 60 year gap between the data set and reality today.