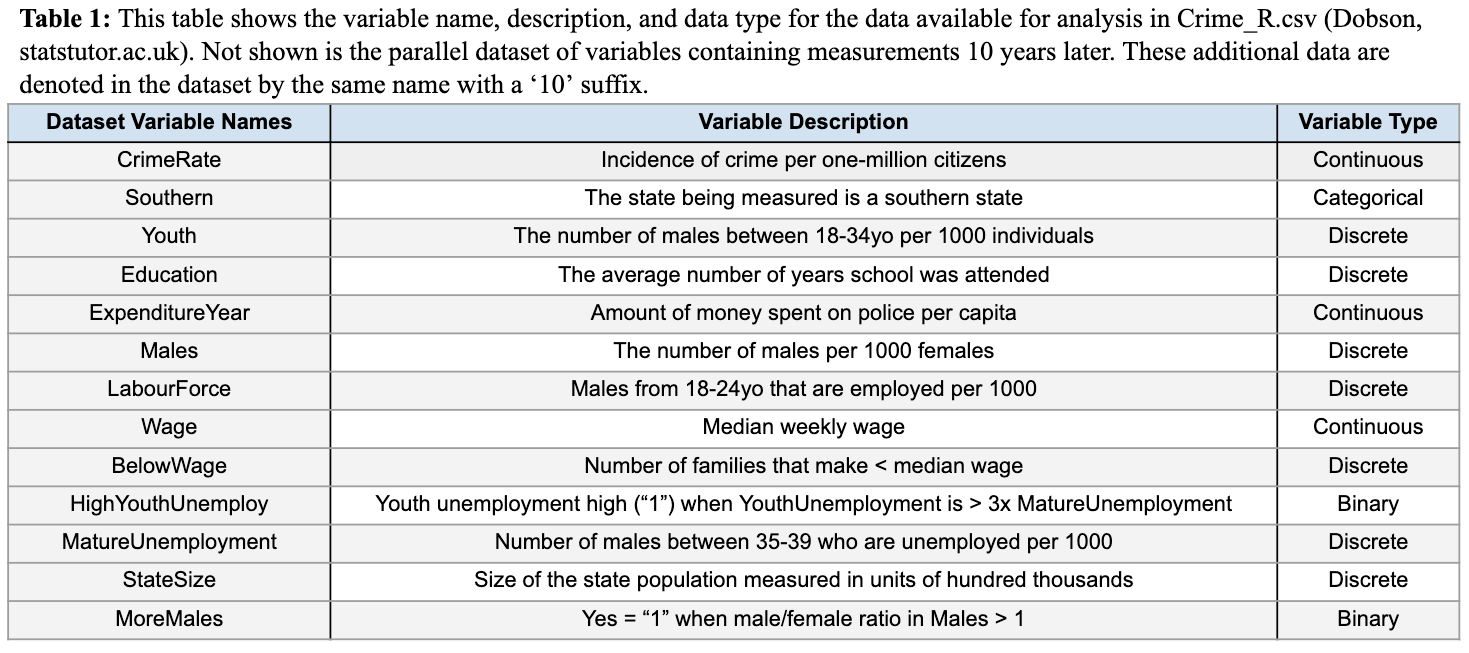
A Statistical Analysis of Crime Rate Correlation to Several Variables Measured Within 47 Unknown U.S. States

Samip Thapa & Ellis Torrance

**Proposed Dataset Description:**

The dataset we propose to analyze contains a total of 14 variables collected at a single time point and then ten-years-later from 47 different unknown states at an unknown date. It has been sourced from statstutor.ac.uk and where it was contributed by Katy Dobson from the University of Leeds. Specifically, this dataset looks at differences in crime rate per population of one-million citizens before and after a 10 year period and how it correlates to the number of male youth in the population, whether the state measured is a Southern state (states were denoted as Southern or Northern by an unknown metric), the average number of years of education the average citizen attains, the money spent on police within the region, the youth-male employment rate per 1000, the number of males per 1000 females, regions in which more males were present than females, the size of the state surveyed, the number of unemployed youth males per 1000, the number of unemployed mature males, whether there was high youth unemployment, the median weekly wage, and the number of families earning below the median wage per 1000. The variables CrimeRate, Expenditure, and Wage are continuous in nature whereas Southern, MoreMales, and HighYouthUnemployment are binary and the rest are discrete. The dataset in its entirety is summarized in **Table 1**. For this study, CrimeRate will be treated as the dependent variable and all other variables will be treated as independent.

**Project Proposal:**

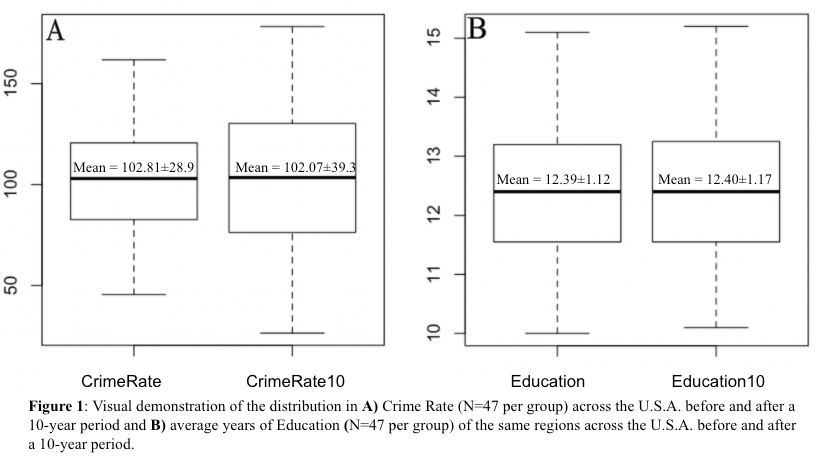
For our statistics project we propose to address the following questions: *i) Does the average crime rate change over a ten-year period and how does it relate to the change in education over ten-years? ii) Has Northern State Crime Rate changed significantly within the last ten years? iii) Has Southern State Crime Rate changed significantly within the last ten years? iv) What is the relationship between crime rate and police expenditure? v) Can the number of families below half wage predict crime rate? and vi) What are the best variables in the dataset for prediction of crime rate in a multiple regression?.*

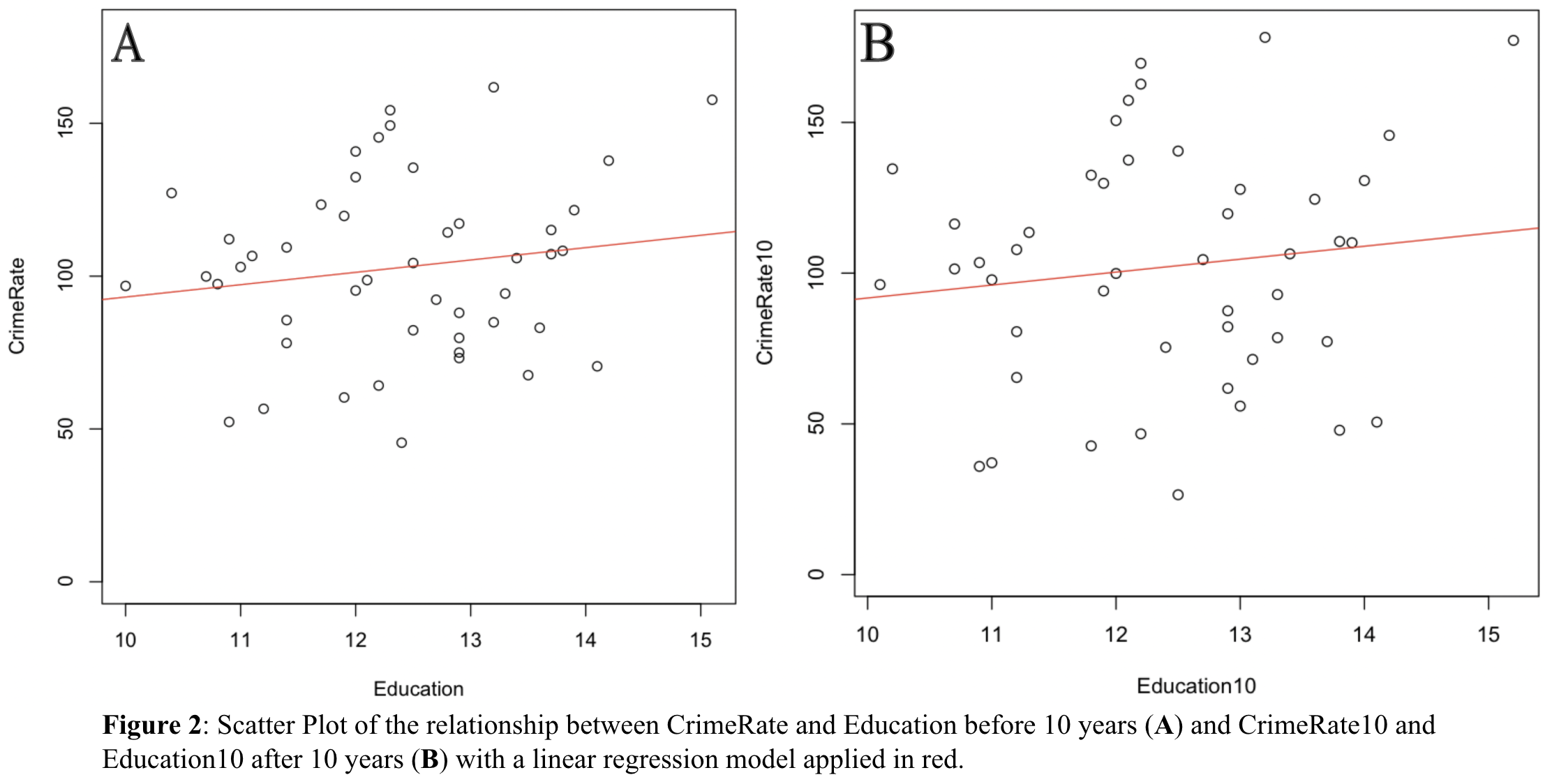
In answering these questions, we will employ a variety of statistical tests and methodologies. Firstly, all data to be used in this study will be visualized and compared via box plot or scatter plot scatter plot as necessary. i) To check whether there is a difference in crime rate before and after a 10 year period and how it relates to changes in education over a ten-year period, a paired t-test will be used to look at the change in education and crime rate and then a linear regression will be used to relate education to crime rate. A paired t-test is appropriate for this analysis as we will be comparing the same group before and after a ten-year period and the linear regression to model the relationship between the independent variable (education years) to the dependent variable (crime rate). ii & iii) To compare the relationship between Southern States and crime rate we will use a paired t-test to compare the average crime rate of Northern to Southern states before and after a 10 year period because we are comparing the same groups but in different time periods. iv) To check the relationship between crime rate and police expenditure, we will be using correlation techniques which are appropriate for answering the question as correlation helps to identify the strength among those two variables. v) To see whether the number of families below half wage predict crime rate, we can use simple linear regression. This analysis technique is appropriate to address this question because we can see the relationship between the two variables and also see if the independent variable (number of families) is a good predictor for crime rate. vi) To determine the best predictors for crime rate, we can use computer assisted variable selection to generate a multiple regression model for Crime Rate. This is appropriate as we are trying to find the best predictors for crime rate from multiple independent variables. If any of the above datasets are found to violate any assumptions of their proposed models, the violations will be listed and a new model will be chosen to better fit the dataset.

In summary, completion of these proposed studies will allow us to determine what social, socio-economic, and regional data best relate to crime rate and how these factors may change over a 10-year period.

**Preliminary Analysis:**

To determine whether there is a difference in crime rate before and after a 10 year period and how it relates to changes in education over a ten-year period the columns CrimeRate (incidence of crime occurrence per one-million citizens) were compared to CrimeRate10 (CrimeRate recorded 10 years after the initial dataset) via a boxplot in **Figure 1A** to determine whether the dataset would violate any assumptions of a two-sample t-test. The variable CrimeRate (N=47) has a mean of 102.81 (with a 95% confidence interval between 73.91 between 131.71) crime incidents per million persons and the variable CrimeRate10 (N=47) has a mean of 102.07 (with a 95% confidence interval between between 62.77 and 141.37). Visually, these two datasets appear to have some slight skew and a slight inequality in variance however these differences are relatively small and do not appear to violate any of assumptions of the paired t-test. Furthermore, according to the dataset description the two observations of CrimeRate before and after 10 years are independent of each other and contain nearly normally distributed values within. To determine whether there was a significant difference in crime rate before and after a ten-year a paired t-test was used. The test found that there was no significant difference in the means of reported crime rate between the two datasets (two-tailed p-value = 0.64, d.f. = 46).

The same process was performed for the variable Education and Education10 which is the average years of schooling attended in the region of crime rate sampling before and after a 10-year period. The two variables were visualized in a boxplot in **Figure 1B.** Education (N=47) was found to have a mean of 12.39 (with a 95% confidence interval between between 11.27 and 13.51) years of school attended with a mean of 12.40 (with a 95% confidence interval between between 11.23 and 13.57) years of school attended for Education10 (N=47) which is the average years of schooling attended for the demographic after a ten-year period. Visually and numerically these datasets appear to be nearly identical with nearly equal means, variance, and standard error. Both datasets appear to follow a normal distribution and according to the dataset description, the two samples are independent and contain normally distributed samples within. To determine whether there was a significant difference in average years of schooling attended before and after a ten-year period, a paired t-test was used. The test found that there was no significant difference reported in mean education between the two datasets (two-tailed p-value = 0.56, d.f. = 46).

We hypothesized that there would be a linear relationship between the response variable CrimeRate and the explanatory variable Education however, in visualizing the scatter plots in **Figure 2**, we can see that there does not appear to be a linear pattern relating the two datasets. Furthermore, in applying a regression model to the relationship of CrimeRate and Education and CrimeRate10 and Education10 we additionally observed a two-tailed p-value of 0.37 and 0.52 respectively indicating that there is no convincing evidence of a linear association between 𝜇CrimeRate and Education either before or after 10 years. To see whether a transformation might change the linearity of the data, we applied both a log and semi-log transformation of the variables but did not find a linear relationship with either transformation. Because the relationship between these two variables is not linear in nature, we may not apply a linear model to explain the relationship between these two variables. However, Education may still be a useful variable in a multiple regression model for CrimeRate. 

To determine whether there is a difference in the crime rate before and after 10 years in the Southern states, the difference of the column CrimeRate and CrimeRate10 of the Southern states was compared via a boxplot and a density plot to determine whether the dataset would violate any assumptions of a paired t-test (**Figure 3**). Before taking the difference of these columns, the variable CrimeRate of the Southern state has a mean of 100.7 crime incidents per million persons with a standard deviation of 22.6 and the variable CrimeRate10 of the Southern state has a mean of 100.5 with a standard deviation of 31.2. After taking the difference of the columns to prepare it for a paired t-test, the difference in the mean was 0.15625 with a standard deviation of 8.80121. Visually looking at the boxplot, we see that there might be two potential outliers. This may also happen to make the distribution skewed. The data is a continuous data. Furthermore, according to the dataset description the two samples are independent and contain random samples within. A paired t-test was carried out to assess the difference in crime rate before and after a 10-year period of the Southern states with the outlier and with the outlier removed. With the outlier, there is no statistical evidence that the mean difference between crime rate in southern states before and after a 10-year period is different than zero (p-value = 0.9443 from a paired t test). With the two outliers removed, there is also no statistical evidence that the mean difference between crime rate in Southern states before and after a 10-year period is different than zero (p-value = 0.1138 from a paired t test).

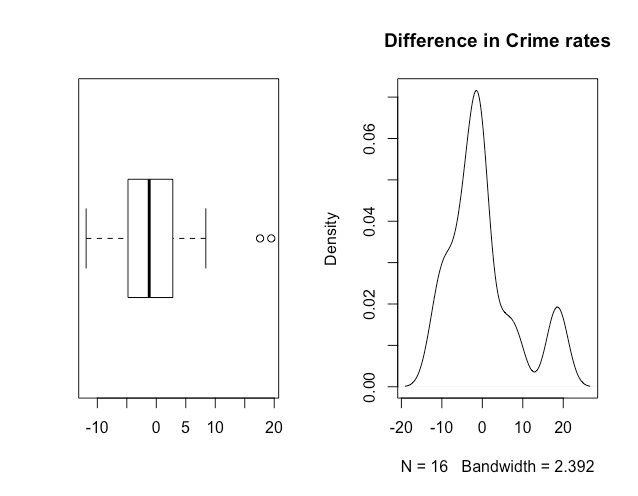


Fig:3. Difference in Crime rate before and after of the Southern state

Similar procedure was carried out to assess the mean difference in the crime rate before and after the 10-year period of the Northern state (**Figure 4**). The mean difference in the crime rate was found to be 1.03871 crime incidents per million persons with a standard deviation of 11.75593. Visually looking at the boxplot at the density plot, the data looks approximately normally distributed. The data is a continuous data. Furthermore, according to the dataset description the two samples are independent and contain random samples within. A paired t-test was carried out to assess the difference in crime rate before and after a 10-year period of the Northern states. There is no statistical evidence that the mean difference between crime rate in Northern states before and after a 10-year period is different than zero (p-value = 0.6263 from a paired t test).

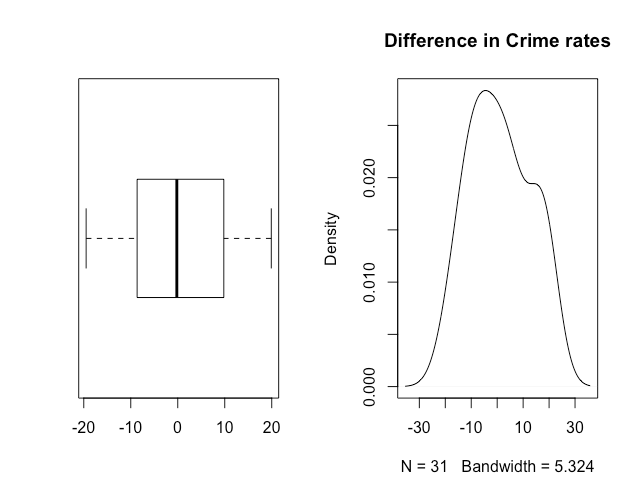


Fig:4. Difference in Crime rate before and after 10 years of the Northern state.

We hypothesized that there would be a linear relationship between the dependent variable CrimeRate and the independent variable police expenditure. When looking at the Spearman’s correlation coefficient between crime rate and the expenditure year 0, we found it to be 0.6493537. This shows a positive correlation between Crime rate and Expenditure for the base year 0. In checking whether we could apply a linear regression model, we saw that taking the log(Expenditure Year) provides better result when assessing the scatter plot, scatterplot with the fitted line, Residual vs Fitted plot, and the Normal Q-Q plot to assess linear regression model between crime rate and expenditure year 0. **Fig 5** represents various plots without the transformation and **Fig 6** represents various plots with log transformation on police expenditure for base year 0.

When a simple linear regression model was fitted with Crime rate as the dependent variable and Expenditure being the independent variable for the base year, a multiple was observed and an adjusted was observed. On the other hand, when a simple linear regression model was fitted with Crime rate as the dependent variable and log(Expenditure) being the independent variable for the base year, multiple was observed and an adjusted was observed. This shows a 3.92% increase in variation explained by the model in predicting crime rate around its mean. However, since this model explains only 44.38% of the variation in the crime rate variable around its mean, the explanatory variable expenditure alone may not be a good predictor for crime rate and a richer model with various other explanatory variables needs to be assessed.

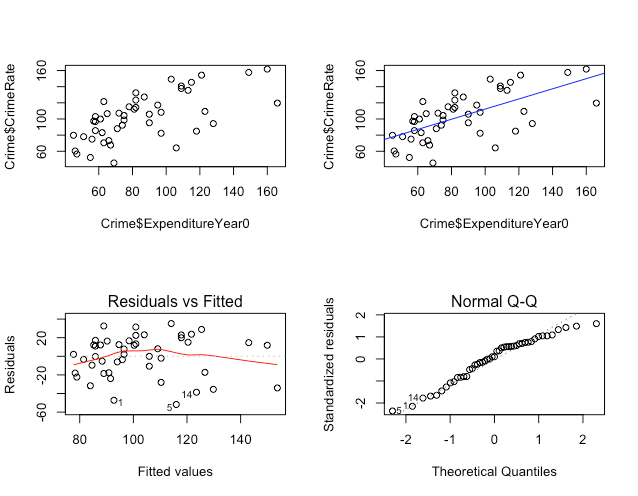


Fig:5. Scatterplot (A), Scatterplot with the fitted line (B), Residual vs Fitted plot (C), and the Normal Q-Q (D) plot to assess linear regression model between crime rate and expenditure year 0

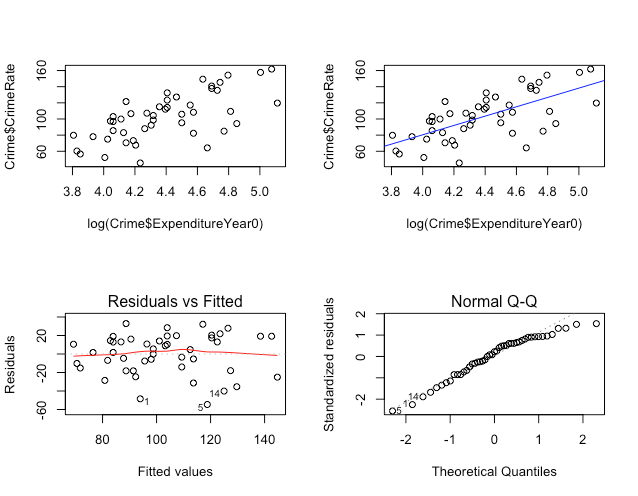


Fig:6. Scatterplot (A), Scatterplot with the fitted line (B), Residual vs Fitted plot (C), and the Normal Q-Q (D) plot to assess linear regression model between crime rate and log (expenditure year 0)

Similarly, for the expenditure year 10, the Spearman’s correlation coefficient between crime rate and the expenditure year 10, we found it to be 0.6502011. This shows a positive correlation between Crime rate and Expenditure for year 10. In checking whether we could apply a linear regression model, we saw that taking the log(Expenditure Year) provides better result when assessing the scatter plot, scatterplot with the fitted line, Residual vs Fitted plot, and the Normal Q-Q plot to assess linear regression model between crime rate and expenditure year 10. **Fig 7** represents various plots without the transformation and **Fig 8** represents various plots with log transformation on police expenditure for year 10.

When a simple linear regression model was fitted with Crime rate as the dependent variable and Expenditure being the independent variable for year 10, multiple was observed and an adjusted was observed. On the other hand, when a simple linear regression model was fitted with Crime rate as the dependent variable and log(Expenditure) being the independent variable for the base year, a multiple was observed and an adjusted was observed. This shows a 4.21% increase in variation explained by the model in predicting crime rate around its mean. However, since this model explains only 42.64% of the variation in the crime rate variable around its mean, the explanatory variable expenditure alone may not be a good predictor for crime rate and a richer model with various other explanatory variables needs to be assessed.

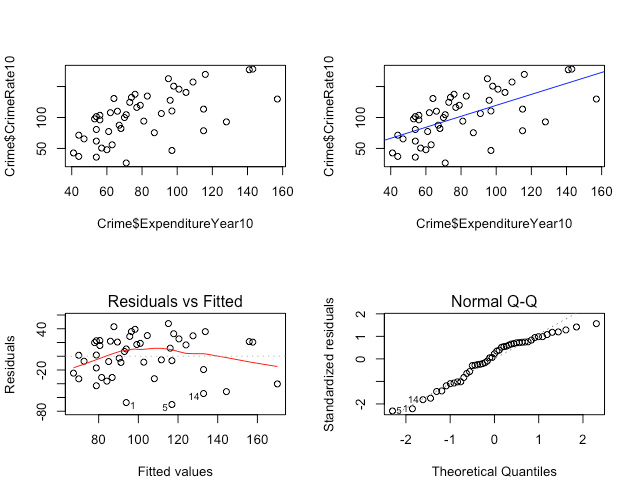


Fig:7. Scatterplot (A), Scatterplot with the fitted line (B), Residual vs Fitted plot (C), and the Normal Q-Q (D) plot to assess linear regression model between crime rate and expenditure year 10

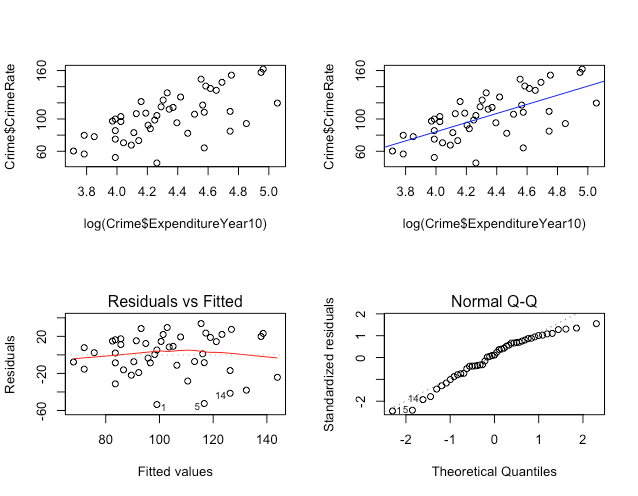


Fig:8. Scatterplot (A), Scatterplot with the fitted line (B), Residual vs Fitted plot (C), and the Normal Q-Q (D) plot to assess linear regression model between crime rate and log (expenditure year 10)

Next, we tried to assess whether we can fit a simple linear regression model to check whether the independent variable number of families below half wage is a good predictor for the dependent variable crime rate at the base year 0. When looking at the scatterplot with the fitted values (**Fig 9.B**), we see that most of the points do not lie close to the fitted line. Additionally, if we look at the residual vs the fitted values (**Fig 9.C**), we see that the residual tends to depart away from the 0 line. Moreover, the normal probability plot (**Fig 9.D**) also does not seem to be close to the straight line. Hence, the assumption of the linear model seems to be violated and using a linear regression model by using the variable below wage to predict crime rate may not be appropriate. Furthermore, = 0.028, which shows that the model explains only 2.8% of the variation in the crime rate variable around its mean. This shows that below wage is not a good predictor for crime rate using the simple linear regression.

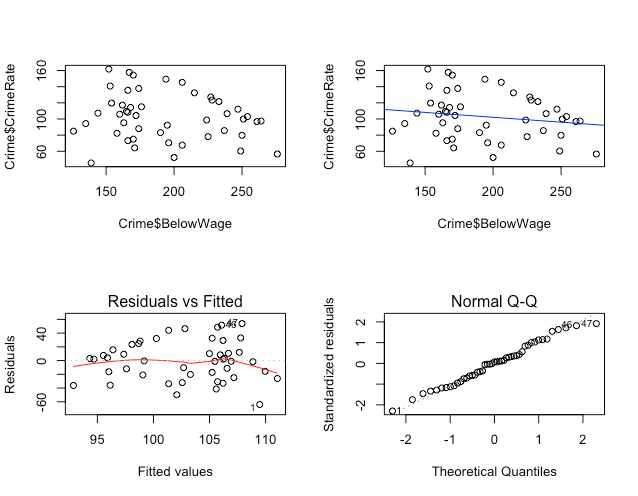


Fig:9. Scatterplot (A), Scatterplot with the fitted line (B), Residual vs Fitted plot (C), and the Normal Q-Q (D) plot to assess linear regression model between crime rate and Below Wage at base year

The results were similar for year 10 in using the simple linear regression model to predict the crime rate. When looking at the scatterplot with the fitted values (**Fig 10.B**), we see that most of the points do not lie close to the fitted line. Additionally, if we look at the residual vs the fitted values (**Fig 10.C**), we see that the residual tends to depart away from the 0 line. Moreover, the normal probability plot (**Fig 10.D**) also does not seem to be close to the straight line. Hence, the assumption of the linear model seems to be violated and using a linear regression model by using the variable below wage to predict crime rate may not be appropriate. Furthermore, = 0.004351, which shows that the model explains only 0.4% of the variation in the crime rate variable around its mean. This shows that below wage is not a good predictor for crime rate using the simple linear regression.

Finally, to see whether a transformation might change the linearity of the data, we applied both a log and semi-log transformation of the variables but did not find a linear relationship with either transformation.

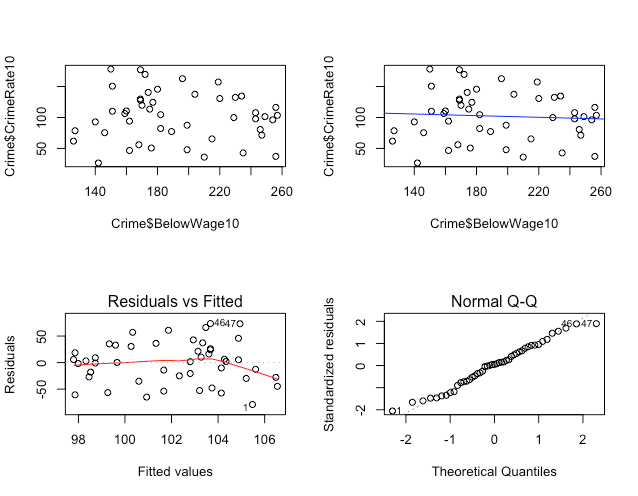
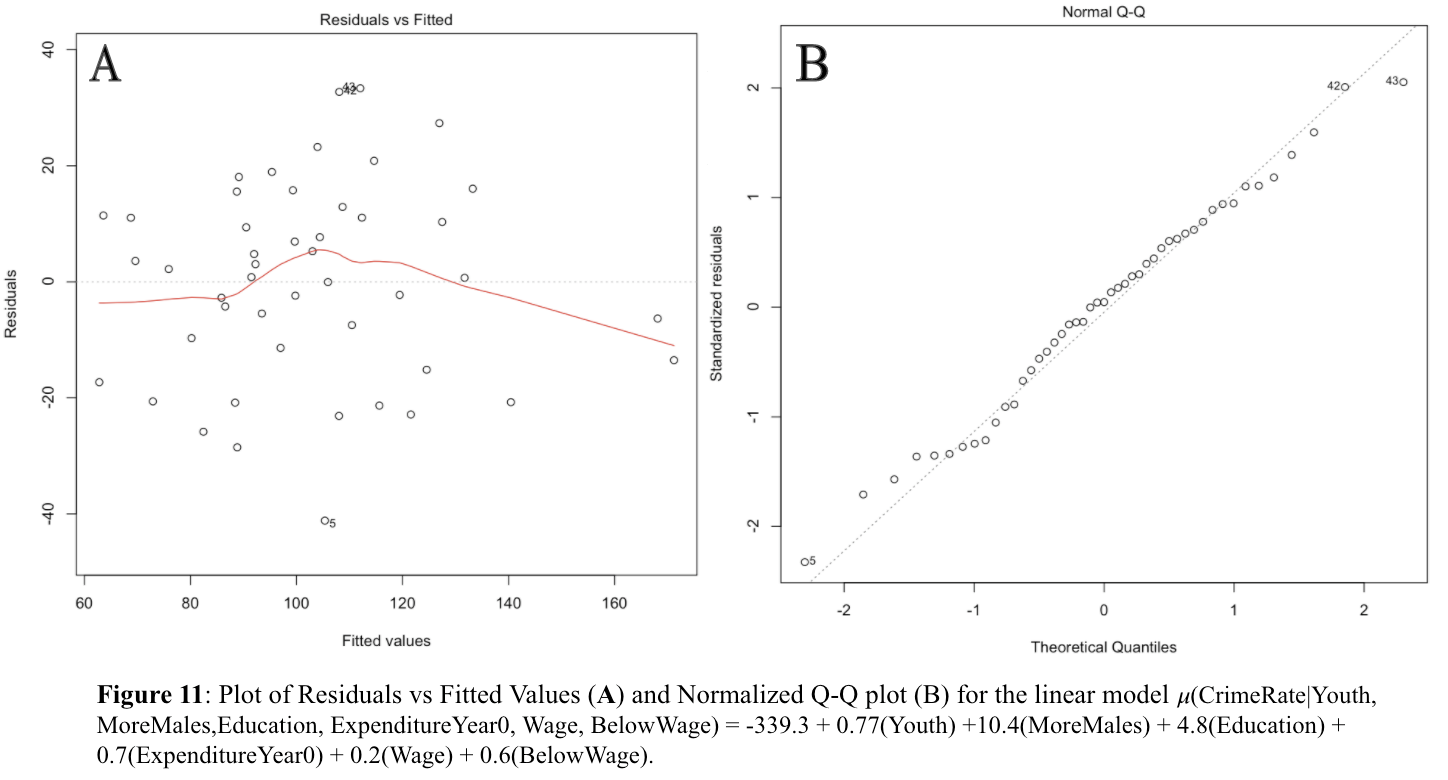
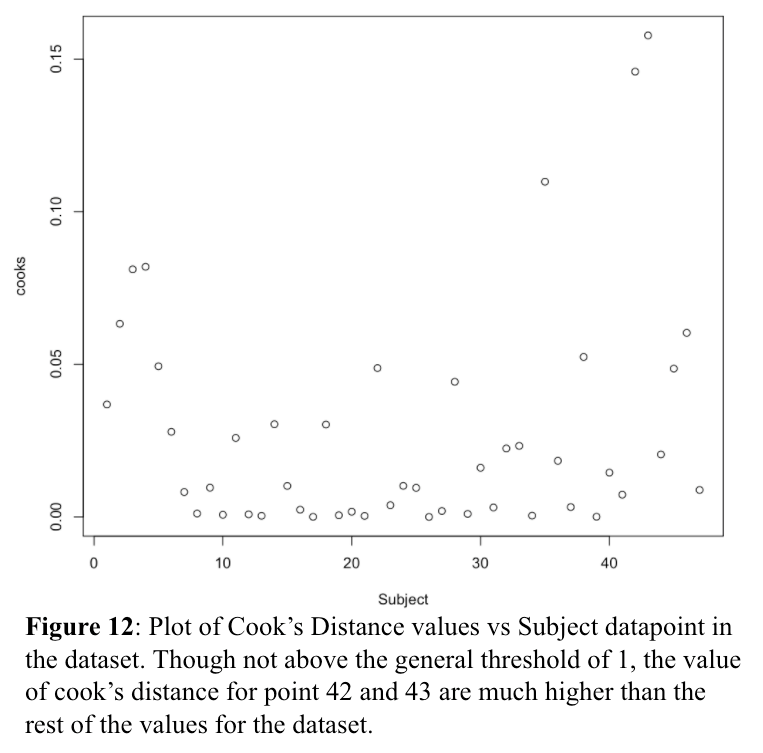


Fig:10. Scatterplot (A), Scatterplot with the fitted line (B), Residual vs Fitted plot (C), and the Normal Q-Q plot (D) to assess linear regression model between crime rate and Below Wage at year 10

Next, we sought to address what the best prediction model for the dependent variable CrimeRate was using all 13 variables within the dataset as predictors. Using computer assisted variable selection, all 13 predictors were assessed to determine best subsets of predictors for CrimeRate using Mallow’s Cp as a measure of comparison between the models. The lowest Cp score was 3.29 for the model 𝜇(CrimeRate|Youth, MoreMales,Education, ExpenditureYear0, Wage, BelowWage) = -339.3 + 0.77(Youth) +10.4(MoreMales) + 4.8(Education) + 0.7(ExpenditureYear0) + 0.2(Wage) + 0.6(BelowWage) which has a multiple R2 value of 0.65 indicating that 65% percent of the variation of CrimeRate around its mean is explained by this reduced model. Next, we observed diagnostic plots of the reduced model. In (**Figure 11A)**, we see a slight horn-shaped pattern in the center of the graph as the residuals depart from zero in the graph of residuals vs. fitted values which may indicate that the dataset violates the assumption of linearity for the model - indicating the model should be used with caution. The graph also shows 3 potential outliers at points 5, 42, and 43. A Normal Q-Q plot was then used to assess normality of the data (**Figure 11B**). The linearity of the plot seems to indicate the majority of the data is normally distributed, however the outliers may be contributing some skew as seen at the tails. 

We then applied the influence statistic, Cook’s Distance, to determine whether the points identified as outliers were heavily influencing the slope. Though not above the general threshold of 1, the value of cook’s distance for point 42 and 43 are much higher than the rest and so we expect removing these values to have an influence on the model (**Figure 12**). The cook’s distance value for point 5 is low and so we do not expect its removal to have an influence on the model. Removal of these points was found to affect the regression coefficients and their p-values and so they were excluded from the analysis. The new suggested model for CrimeRate prediction with outliers (point 42 and 43) removed is 𝜇(CrimeRate|Youth, MoreMales,Education, ExpenditureYear0, Wage, BelowWage) = -410.1 + 0.73(Youth) +12.4(MoreMales) + 5.5(Education) + 0.5(ExpenditureYear0) + 0.3(Wage) + 0.7(BelowWage) with a multiple R-squared value of 0.70 indicating that 70% percent of the variation of CrimeRate around its mean is explained by this reduced model - a 5% improvement in model performance pre-outlier removal.

**Discussion**

For this analysis, we proposed to answer the following questions about Crime Rate in the U.S. using a dataset that contained a total of 14 variables collected at a single time point and then ten-years-later from 47 unknown states at an unknown date (**Table 1**): *i) Does the average crime rate change over a ten-year period and how does it relate to the change in education over ten-years? ii) Has Northern State Crime Rate changed significantly within the last ten years? iii) Has Southern State Crime Rate changed significantly within the last ten years? iv) What is the relationship between crime rate and police expenditure? v) Can the number of families below half wage predict crime rate? and vi) What are the best variables in the dataset for prediction of crime rate in a multiple regression?.*

To determine whether the average crime-rate and years of education changed over a ten-year period, the two datasets were assessed visually with a boxplot (**Figure 1**) and then with a paired t-test. Neither dataset was found to have significant difference in the compared means indicating that there was no significant change in Education or Crime Rate over the ten-year period in the dataset tested. Additionally, we hypothesized that there would be a linear relationship between the response variable CrimeRate and the explanatory variable Education however, we found no evidence of a linear relationship even after transformations were applied.

To determine whether there is a difference in the crime rate before and after 10 years in the Southern states and Northern States (“Southern” and “Northern” were denoted by an unknown metric in this dataset), the difference of the column CrimeRate and CrimeRate10 of the Southern states and Northern States was compared via a boxplot and a density plot (**Figure 3 & 4**) and then with a paired t-test. In both cases, it was found that there was no significant evidence that the mean difference between crime rate, in either Northern or Southern states before and after a 10-year period is different than zero - even after removal of potential outliers.

We next sought to determine the relationship between crime rate and police expenditure for the base year 0 and ten-years later. We found that there was a positive correlation between the crime Rate and police expenditure in both years of measurement with a Spearman’s correlation coefficient of >0.6 for both datasets. In both cases, the log transformation on the explanatory variable of police expenditure was found to improve the fit of the linear regression model on the dataset. However, the explanatory variable expenditure alone was not good enough to predict crime rate and that a richer model with various other explanatory variables were needed to be accounted for predicting crime rate.

Similarly, we sought to address whether the number of families below half-wage was a good predictor of crime-rate in the base-year 0 in a simple linear regression. We found that this relationship most likely violated the assumptions of linearity for the model and thus that below wage is not a good predictor for crime rate using the simple linear regression either with or without transformations applied. In assessment of the same variables for year 10, we found the same result - below wage is not a good predictor for crime rate using the simple linear regression due to violations of linearity.

Next, we sought to address what the best prediction model for the dependent variable crime rate, in year zero, was using all 13 variables within the dataset as predictors. Using computer assisted variable selection, the best performing model, based on the comparison metric of Mallow’s Cp,was found to use the variables Youth, MoreMales, Education, ExpenditureYear0, Wage, BelowWage to predict crime rate in the dataset with the highest accuracy. Once outliers were detected and removed from this dataset, the suggested model for prediction of crime rate within this dataset was found to be 𝜇(CrimeRate|Youth, MoreMales,Education, ExpenditureYear0, Wage, BelowWage) = -410.1 + 0.73(Youth) +12.4(MoreMales) + 5.5(Education) + 0.5(ExpenditureYear0) + 0.3(Wage) + 0.7(BelowWage).

In summary, our analysis found that there was no evidence of a change in the average crime rate or education over a 10-year period and that there was no evidence of a linear relationship between crime rate and education. There was also found to be no evidence of a change in crime rate in Northern or Southern states after a 10-year period and the number of families below half-wage was not found to be a good predictor of crime rate using linear regression. Crime rate and police expenditure, however, were found to be linearly related and positively correlated in year zero and year 10. But, police expenditure alone was not a good predictor for crime rate. Finally, our analysis concluded that the best variables for prediction of crime rate within this dataset were the variables Youth, MoreMales, Education, ExpenditureYear0, Wage, BelowWage in a multiple regression model.

**Dataset:**

Katy Dobson, “Crime Rate Data” <https://www.sheffield.ac.uk/mash/statistics/datasets>, University of Leeds (Date Unknown)

**R code:**

**Torrance Contibution:**

#import data

crime\_table <- read.table("Crime\_R.csv", sep=",", h = T)

attach(crime\_table)

library(plyr)

# difference in crime rate before and after a 10 year period

N = length(CrimeRate)

mean = mean(CrimeRate)

sd = sd(CrimeRate)

se = sd/sqrt(N)

cat("CrimeRate" , "sd", sd , "se" , se , "mean", mean)

N = length(CrimeRate10)

mean = mean(CrimeRate10)

sd = sd(CrimeRate10)

se = sd/sqrt(N)

cat("CrimeRate10" , "sd", sd , "se" , se , "mean", mean)

boxplot(CrimeRate, CrimeRate10)

t.test(CrimeRate, CrimeRate10, paired=T, conf.level=0.95, data=crime\_table)

# difference in education before and after a 10 year period

N = length(Education)

mean = mean(Education)

sd = sd(Education)

se = sd/sqrt(N)

cat("Education" , "sd", sd , "se" , se , "mean", mean)

N = length(Education10)

mean = mean(Education10)

sd = sd(Education10)

se = sd/sqrt(N)

cat("Education10" , "sd", sd , "se" , se , "mean", mean)

boxplot(Education, Education10)

t.test(Education, Education10, paired=T, conf.level=0.95, data=crime\_table)

#Determine whether the relationship between CrimeRate and Education can be explained with a linear model

lm\_CrimeEdu <- lm(CrimeRate~Education, data=crime\_table)

summary(lm\_CrimeEdu) #determine whether a linear model applies

lm\_CrimeEdu10 <- lm(CrimeRate10~Education10, data=crime\_table)

summary(lm\_CrimeEdu10)

{plot(CrimeRate~Education, ylim=c(0,180), data = crime\_table) + abline(lm\_CrimeEdu, col="red")}

{plot(CrimeRate10~Education10, ylim=c(0,180), data = crime\_table) + abline(lm\_CrimeEdu10, col="red")}

lmlog\_CrimeEdu <- lm(log(CrimeRate)~log(Education), data=crime\_table)

summary(lmlog\_CrimeEdu) #determine whether a linear model applies

lmlog\_CrimeEdu10 <- lm(log(CrimeRate10)~log(Education10), data=crime\_table)

summary(lmlog\_CrimeEdu10)

{plot(log(CrimeRate10)~log(Education10), ylim=c(2,6), data = crime\_table) + abline(lmlog\_CrimeEdu, col="red")}

{plot(log(CrimeRate10)~log(Education10), ylim=c(2,6), data = crime\_table) + abline(lmlog\_CrimeEdu10, col="red")}

#exploratory extras

pairs(~logit(CrimeRate)+(Youth)+(Males)+(LabourForce)+(BelowWage)+(StateSize), upper.panel = panel.smooth, data=crime\_table)

#Use multiple linear regression to determine the best predictors for crime rate

library(leaps)

library(MASS)

lm\_full=lm(CrimeRate~Youth+Southern+MoreMales+HighYouthUnemploy+Education+ExpenditureYear0+LabourForce+Males+StateSize+YouthUnemployment+MatureUnemployment+Wage+BelowWage, data = crime\_table)

fit.best=regsubsets(CrimeRate~Youth+Southern+MoreMales+HighYouthUnemploy+Education+ExpenditureYear0+LabourForce+Males+StateSize+YouthUnemployment+MatureUnemployment+Wage+BelowWage, data = crime\_table, nbest = 13)

sum.fit.best <- summary(fit.best)

sum.fit.best

sum.fit.best$cp

sum.fit.best$bic

min(sum.fit.best$bic)

min(sum.fit.best$cp)

lm\_reduced = lm(CrimeRate~Youth+MoreMales+Education+ExpenditureYear0+Wage+BelowWage, data = crime\_table)

plot(lm\_reduced)

crime\_table$Subject <- seq.int(nrow(crime\_table))

crime\_table$cooks <- cooks.distance(lm\_reduced)

crime\_table2 <- crime\_table[ !(crime\_table$Subject %in% c(42)), ]

crime\_table3 <- crime\_table2[ !(crime\_table2$Subject %in% c(43)), ]

lm\_reduced2 = lm(CrimeRate~Youth+MoreMales+Education+ExpenditureYear0+Wage+BelowWage, data = crime\_table3)

**Thapa Contribution:**

require(mosaic)

library(readxl)

Crime = read\_excel("Desktop/Crime\_R.xlsx")

#ii & iii) To compare the relationship between Southern States and crime rate we will use

#a paired t-test to compare the average crime rate of Northern to Southern states before and after

#a 10 year period because we are comparing the same groups but in different time periods. (Samip)

# Crime rate for Southern state before and after 10 years

S.Crime = Crime$CrimeRate[Crime$Southern==1]

S.Crime10 = Crime$CrimeRate10[Crime$Southern==1]

df = data.frame(S.Crime,S.Crime10)

summary(df)

sd(S.Crime)

sd(S.Crime10)

df = transform(df, DIFF = S.Crime - S.Crime10)

favstats(~DIFF, data = df)

par(mfrow=c(1,2))

boxplot(df$DIFF, horizontal = TRUE)

d = density(df$DIFF)

plot(d, main="Difference in Crime rates")

t.test(df$S.Crime, df$S.Crime10, paired = TRUE)

#Removing one outlier

df3 = df[-c(1), ]

boxplot(df3$DIFF, horizontal = TRUE)

d = density(df3$DIFF)

plot(d, main="Difference in Crime rates")

t.test(df3$S.Crime, df3$S.Crime10, paired = TRUE)

#Removing two outliers

df2 = df[-c(1,2), ]

boxplot(df2$DIFF, horizontal = TRUE)

densityplot(df2$DIFF)

t.test(df2$S.Crime, df2$S.Crime10, paired = TRUE)

#Crime rate for Northern state before and after 10 years

N.Crime = Crime$CrimeRate[Crime$Southern==0]

N.Crime10 = Crime$CrimeRate10[Crime$Southern==0]

df = data.frame(N.Crime,N.Crime10)

summary(df)

sd(N.Crime)

sd(N.Crime10)

df = transform(df, DIFF = N.Crime - N.Crime10)

favstats(~DIFF, data = df)

boxplot(df$DIFF, horizontal = TRUE)

d = density(df$DIFF)

plot(d, main="Difference in Crime rates")

t.test(df$N.Crime, df$N.Crime10, paired = TRUE)

par(mfrow=c(1,1))

#iv) To check the relationship between crime rate and police expenditure,

#we will be using correlation techniques which are appropriate for answering the question

#as correlation helps to identify the strength among those two variables. (Samip)

#For year 0

par(mfrow=c(2,2))

plot(Crime$CrimeRate~Crime$ExpenditureYear0)

cor.test(Crime$ExpenditureYear0, Crime$CrimeRate, method=c("spearman"))

cor(Crime$ExpenditureYear0, Crime$CrimeRate, method=c("spearman"))

lm1 = lm(Crime$CrimeRate~Crime$ExpenditureYear0)

summary(lm1)

plot(Crime$CrimeRate~Crime$ExpenditureYear0)

abline(lm1, col = "blue")

plot(lm1, which = 1)

plot(lm1, which = 2)

#Log of Expenditure

lm2 = lm(Crime$CrimeRate~log(Crime$ExpenditureYear0))

summary(lm2)

plot(Crime$CrimeRate~log(Crime$ExpenditureYear0))

plot(Crime$CrimeRate~log(Crime$ExpenditureYear0))

abline(lm2, col = "blue")

plot(lm2, which = 1)

plot(lm2, which = 2)

#Log of crime rate

lm3 = lm(log(Crime$CrimeRate)~Crime$ExpenditureYear0)

summary(lm3)

plot(log(Crime$CrimeRate)~Crime$ExpenditureYear0)

plot(log(Crime$CrimeRate)~Crime$ExpenditureYear0)

abline(lm3, col = "blue")

plot(lm3, which = 1)

plot(lm3, which = 2)

#Log of both expenditure and crime rate

lm3 = lm(log(Crime$CrimeRate)~log(Crime$ExpenditureYear0))

summary(lm3)

plot(log(Crime$CrimeRate)~log(Crime$ExpenditureYear0))

plot(log(Crime$CrimeRate)~log(Crime$ExpenditureYear0))

abline(lm3, col = "blue")

plot(lm3, which = 1)

plot(lm3, which = 2)

#For year 10

plot(Crime$CrimeRate10~Crime$ExpenditureYear10)

cor.test(Crime$ExpenditureYear10, Crime$CrimeRate10, method=c("spearman"))

cor(Crime$ExpenditureYear10, Crime$CrimeRate10, method=c("spearman"))

lm2 = lm(Crime$CrimeRate10~Crime$ExpenditureYear10)

summary(lm2)

plot(Crime$CrimeRate10~Crime$ExpenditureYear10)

abline(lm2, col = "blue")

plot(lm2, which = 1)

plot(lm2, which = 2)

#Log of Expenditure

plot(Crime$CrimeRate~log(Crime$ExpenditureYear10))

lm4 = lm(Crime$CrimeRate~log(Crime$ExpenditureYear10))

summary(lm4)

plot(Crime$CrimeRate~log(Crime$ExpenditureYear10))

abline(lm4, col = "blue")

plot(lm4, which = 1)

plot(lm4, which = 2)

#Log of crime rate

lm3 = lm(log(Crime$CrimeRate)~Crime$ExpenditureYear10)

summary(lm3)

plot(log(Crime$CrimeRate)~Crime$ExpenditureYear10)

plot(log(Crime$CrimeRate)~Crime$ExpenditureYear10)

abline(lm3, col = "blue")

plot(lm3, which = 1)

plot(lm3, which = 2)

#Log of both expenditure and crime rate

plot(log(Crime$CrimeRate)~log(Crime$ExpenditureYear10))

lm4 = lm(log(Crime$CrimeRate)~log(Crime$ExpenditureYear10))

summary(lm4)

plot(log(Crime$CrimeRate)~log(Crime$ExpenditureYear10))

abline(lm4, col = "blue")

plot(lm4, which = 1)

plot(lm4, which = 2)

#v) To see whether the number of families below half wage predict crime rate,

#we can use simple linear regression. This analysis technique is appropriate to

#address this question because we can see the relationship between the two variables

#and also see if the independent variable (number of families) is a good predictor

#for crime rate. (Samip)

#For year 0

par(mfrow=c(2,2))

plot(Crime$CrimeRate~Crime$BelowWage)

lm3 = lm(Crime$CrimeRate~Crime$BelowWage)

summary(lm3)

plot(Crime$CrimeRate~Crime$BelowWage)

abline(lm3, col = "blue")

plot(lm3, which = 1)

plot(lm3, which = 2)

#Log of below wage

plot(Crime$CrimeRate~log(Crime$BelowWage))

lm3 = lm(Crime$CrimeRate~log(Crime$BelowWage))

summary(lm3)

plot(Crime$CrimeRate~log(Crime$BelowWage))

abline(lm3, col = "blue")

plot(lm3, which = 1)

plot(lm3, which = 2)

#Log of crime rate

plot(log(Crime$CrimeRate)~Crime$BelowWage)

lm3 = lm(log(Crime$CrimeRate)~Crime$BelowWage)

summary(lm3)

plot(log(Crime$CrimeRate)~Crime$BelowWage)

abline(lm3, col = "blue")

plot(lm3, which = 1)

plot(lm3, which = 2)

#Log of both

plot(log(Crime$CrimeRate)~log(Crime$BelowWage))

lm3 = lm(log(Crime$CrimeRate)~log(Crime$BelowWage))

summary(lm3)

plot(log(Crime$CrimeRate)~log(Crime$BelowWage))

abline(lm3, col = "blue")

plot(lm3, which = 1)

plot(lm3, which = 2)

#For year 10

plot(Crime$CrimeRate10~Crime$BelowWage10)

lm4 = lm(Crime$CrimeRate10~Crime$BelowWage10)

summary(lm4)

plot(Crime$CrimeRate10~Crime$BelowWage10)

abline(lm4, col = "blue")

plot(lm4, which = 1)

plot(lm4, which = 2)

#Log of below wage

plot(Crime$CrimeRate10~log(Crime$BelowWage10))

lm4 = lm(Crime$CrimeRate10~log(Crime$BelowWage10))

summary(lm4)

plot(Crime$CrimeRate10~log(Crime$BelowWage10))

abline(lm4, col = "blue")

plot(lm4, which = 1)

plot(lm4, which = 2)

#Log of crime rate

plot(log(Crime$CrimeRate10)~Crime$BelowWage10)

lm4 = lm(log(Crime$CrimeRate10)~Crime$BelowWage10)

summary(lm4)

plot(log(Crime$CrimeRate10)~Crime$BelowWage10)

abline(lm4, col = "blue")

plot(lm4, which = 1)

plot(lm4, which = 2)

#Log of both

plot(log(Crime$CrimeRate10)~log(Crime$BelowWage10))

lm4 = lm(log(Crime$CrimeRate10)~log(Crime$BelowWage10))

summary(lm4)

plot(log(Crime$CrimeRate10)~log(Crime$BelowWage10))

abline(lm4, col = "blue")

plot(lm4, which = 1)

plot(lm4, which = 2)