Linear Models

Shaun Allard, Collin Giguere, Xulun Wang, Arpan Giri

December 10, 2017

For Integrated Experience project this semester, we decided to explore and analyze a data set dealing with crime in the United States. Our particular data set was from 1960 and was gathered from the US Census Bureau, the FBI Uniform Crime Reports and the National Prison Statistics bulletin. It focused on crime data and socio-economic data from 47 US states, excluding Alaska, Hawaii and New Jersey. The data was originally used to study the relationship between punishment regimes and crime rates and resulted in a paper published by Isaac Ehlrich in 1973 titled “Participation in Illegitimate Activities: A Theoretical and Empirical Investigation”. The crimes considered for the crime rate were murder, rape, sexual offenses other than rape, assault, larceny, robbery, burglary and auto theft. Our analysis focused primarily on determining which variables we could use to predict crime rates.

Our Data

Our data consisted of 47 states and 17 variables relating to a particular socio-economic, demographic or crime factor. Each state was assigned a number from 1 to 47 and were referred to in that way. Our variables were as follows:

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: labor force participation rate of civilian urban males in the age-group 14-24

M.F: number of males per 100 females

Pop:$/quad$ 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: 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

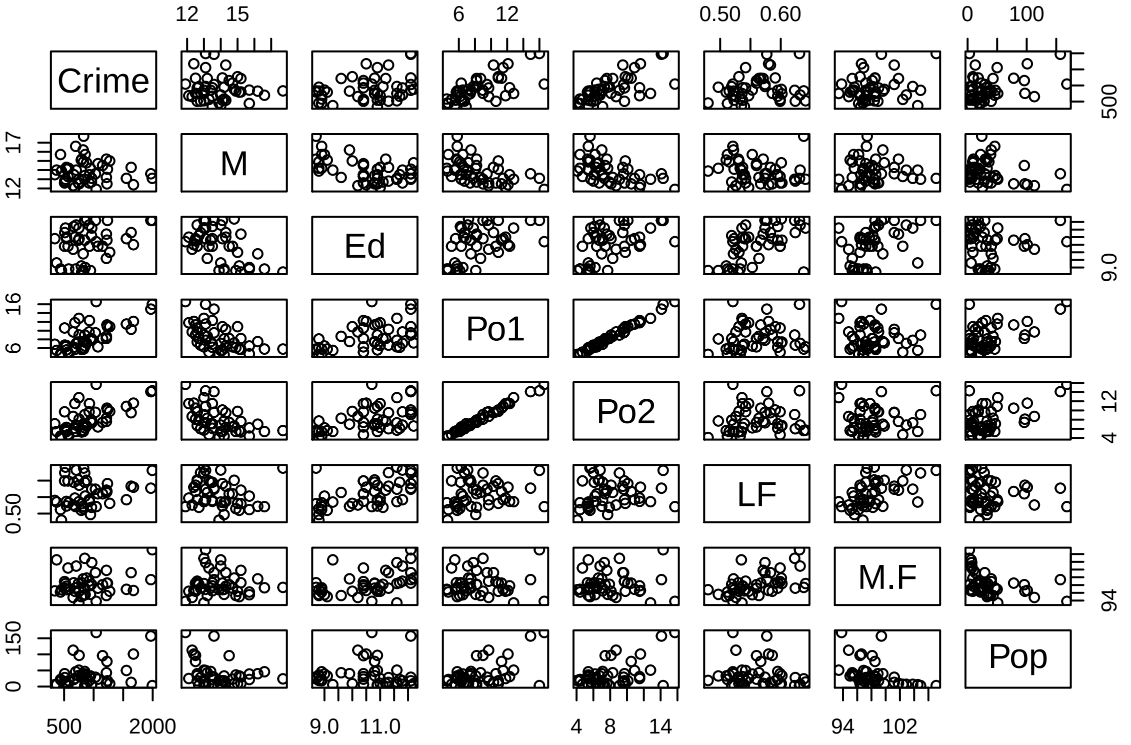
There were a couple of interesting things we noticed about our data. First of all, So referred to whether or not a state was considered Southern or not. For the most part, these states included the usual suspects, though we found it curious that both Delaware and Oregon were included among the southern states. Upon further investigation, Delaware was included because of it’s status as a “border state” during the Civil War. “Border states” were states that did not secede from the Union but continued to allow slavery. There was not a compelling reason to indicate why Oregon would be a southern state, though. We are inclined to believe it may have been an error.

We also noticed that in the original data set that was used by Ehrlich's paper the crime data, Prob data and Time data was far more precise than what we were given. For example, crime was separated into the particular offense rather than being grouped together into one all encompassing crime variable. This would prove to be a hindrance in our search for a unifying model to explain criminal behavior.

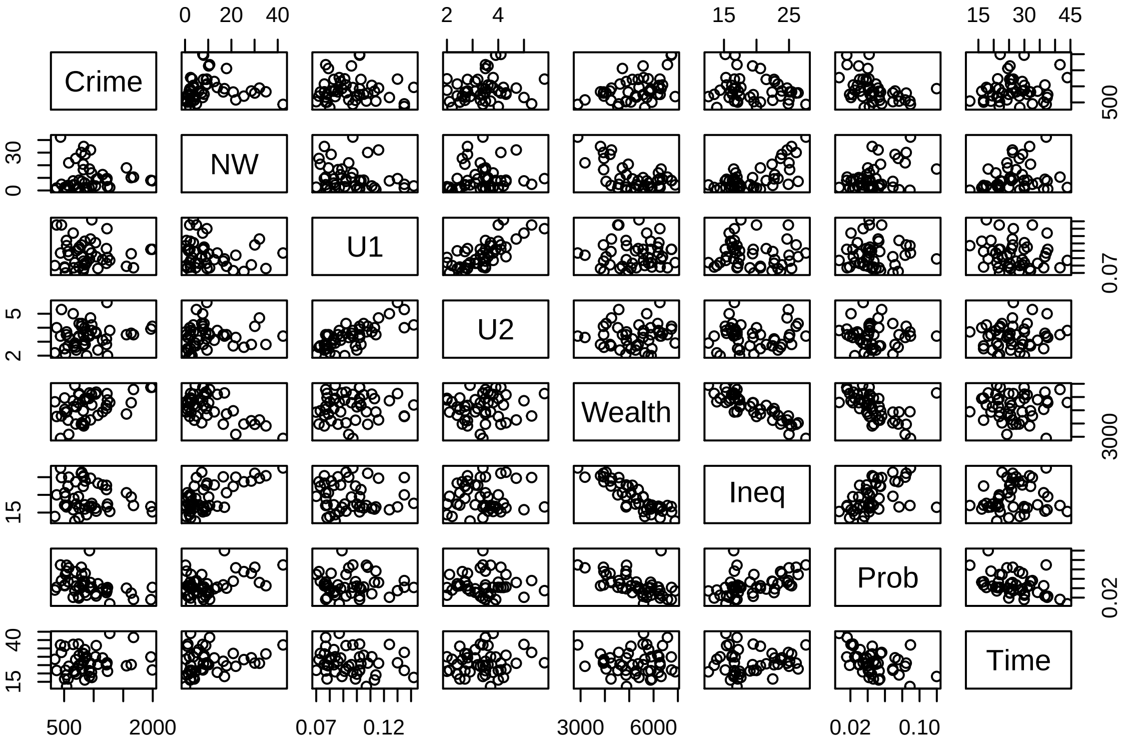
Our Process

After loading our data into R along with the appropriate R packages, our first step was to create a scatter plot matrix to see if any relationships became immediately apparent to us to help focus our exploration of the data set.

```
suppressWarnings(library(tidyr))
library(plyr)
suppressWarnings(suppressMessages(library(dplyr)))
library(ggplot2)
library(readxl)
library(readr)
suppressWarnings(suppressMessages(library(car)))
crime <- read_xls("/Users/mac/Desktop/School Stuff/STAT525/Group Project/crime.xls",
    col_types = c("numeric", "text", "numeric",
        "numeric", "numeric", "numeric",
        "numeric", "numeric", "numeric",
        "numeric", "numeric", "numeric",
        "numeric", "numeric", "numeric",
        "numeric"))
attach(crime)
pairs(crime[,c(16,1,3:8)])
```



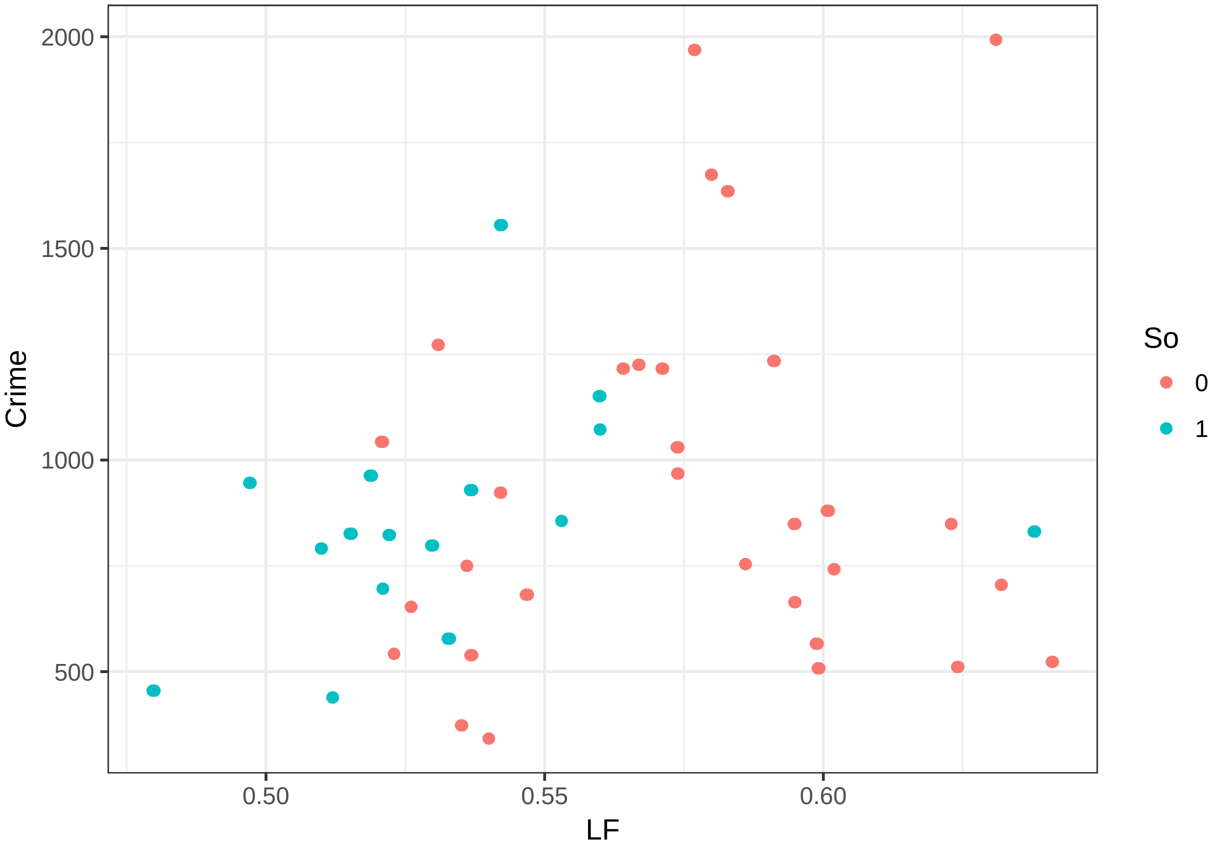
```r
pairs(crime[,c(16,9:15)])
```



One thing we do notice immediately is that Po1 and Po2, police expenditures in 1959 and 1960, do show strong relationships with crime. At first this is a promising lead, but upon further consideration we realize that police spending is probably caused by high crime rates. For this reason, we decide that Po1 and Po2 are no longer in consideration as predictors of crime. Beyond Po1 and Po2, there aren't any other striking relationships that are immediately obvious. Nevertheless, we decide to focus in on any variable that shows even the slightest correlation with crime. We end up settling on labor force participation by young men (LF), percentage of young men in the population (M), the state population (Pop), wealth (Wealth), the percentage of non-whites in the population (NW), education level of adults over 25 (Ed), unemployment level for men 14-24 and 35-39 (U1 & U2), income inequality (Ineq), the probability that an arrest will lead to incarceration (Prob), and the average time spent in prison before an inmate's first release (Time).
Once we narrowed down our data set to a subset of variables, we decided to take a closer look at each of our variables individually vs our response variable, Crime.

Crime vs LF

```r
ggplot(data = crime, aes(x = LF, y = Crime, col = So)) +
  geom_point() +
  geom_jitter(height = 0.2) +
  theme_bw()
```



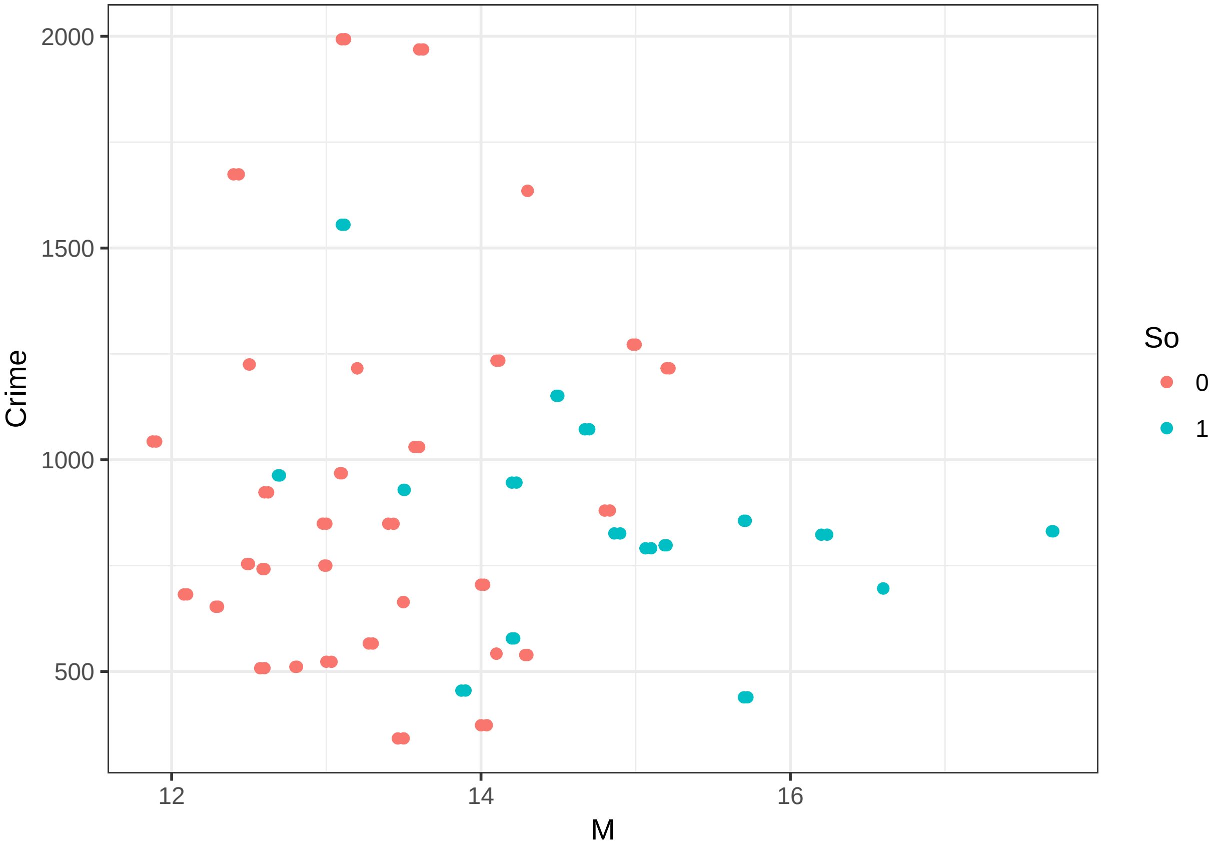
```
summary(lm(crime$Crime ~ crime$LF))
```

```
##
## Call:
## lm(formula = crime$Crime ~ crime$LF)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -526.3 -285.1  -50.7  189.8 1035.3
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -109.3      788.2  -0.139    0.890
## crime$LF      1807.6     1401.0   1.290    0.204
##
## Residual standard error: 384 on 45 degrees of freedom
## Multiple R-squared:  0.03567,    Adjusted R-squared:  0.01424
## F-statistic: 1.665 on 1 and 45 DF,  p-value: 0.2036
```

This plot does not reveal any apparent relationships between crime and labor force participation. The summary of the linear model makes this fact even more clear. There is little statistical significance to the slight positive relationship that does exist as evidenced by the low t-value and high corresponding p-value. LF will not be a candidate for our final model.

Crime vs M

```
ggplot(data = crime, aes(x = M, y = Crime, col = So)) +
  geom_point() +
  geom_jitter(height = 0.2) +
  theme_bw()
```



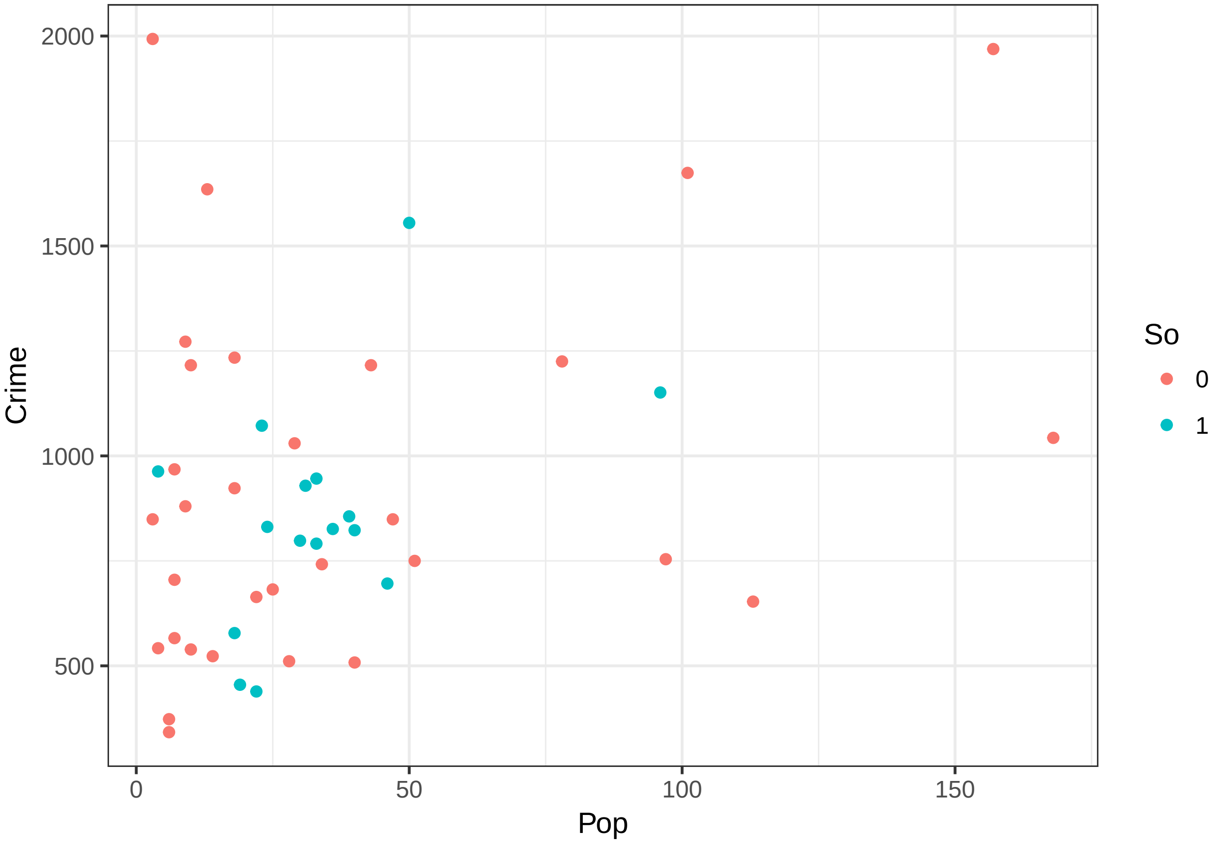
```
summary(lm(crime$Crime ~ crime$M))
```

```
##
## Call:
## lm(formula = crime$Crime ~ crime$M)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -572.93 -283.22  -50.38  153.97 1067.06
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1286.65     635.72   2.024   0.0489 *
## crime$M       -27.53      45.69  -0.603   0.5498
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 389.5 on 45 degrees of freedom
## Multiple R-squared:  0.008005,   Adjusted R-squared:  -0.01404
## F-statistic: 0.3631 on 1 and 45 DF,  p-value: 0.5498
```

We have another variable that does not appear to have a relationship with crime. The plot looks like a null plot with no apparent relationship between the percentage of young men in the state and the crime rate. The summary of a simple linear model confirms the lack of relationship observed in the plot. We have a very low R^2 and t-value paired with a high p-value. M will not be a candidate for our final model.

Crime vs Population

```
ggplot(data = crime, aes(x = Pop, y = Crime, col = So)) +
  geom_point() +
  theme_bw()
```



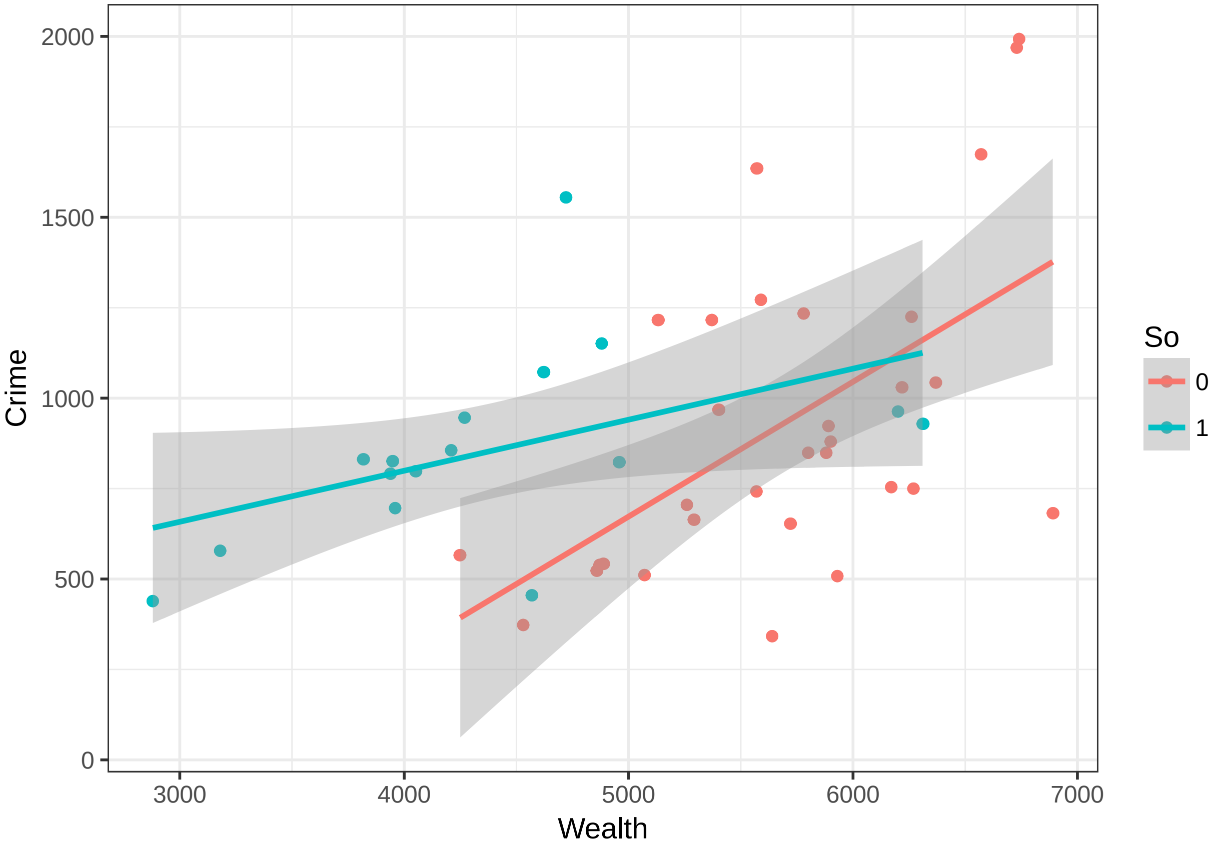
```
summary(lm(crime$Crime ~ crime$Pop))
```

```
##
## Call:
## lm(formula = crime$Crime ~ crime$Pop)
##
## Residuals:
##    Min      1Q  Median      3Q     Max
## -514.0 -257.3   -84.4   167.1  1203.2
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  779.548     74.885  10.410 1.45e-13 ***
## crime$Pop      3.428      1.426   2.405   0.0204 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 368.1 on 45 degrees of freedom
## Multiple R-squared:  0.1139, Adjusted R-squared:  0.0942
## F-statistic: 5.784 on 1 and 45 DF,  p-value: 0.02035
```

Again, no relationship is immediately apparent. As you can see from the plot, there are both high and low populations with very high crime rates as well as both high and low populations with very low crime rates. If we try and just focus on the initial cluster of data points, though, a positive relationship starts to take shape. We realize given the range of population we may want to consider a transformation on the variable. The summary of the linear model justifies this approach. While the t-value isn't extremely high, it is the best we have observed so far. We decide that we will keep this variable as a possible candidate for our final model upon further inspection.

Crime vs Wealth

```
ggplot(data = crime, aes(x = Wealth, y = Crime, col = So)) +
  geom_point() +
  geom_jitter(height = 0.2) +
  geom_smooth(method = lm) +
  theme_bw()
```



```
summary(lm(crime$Crime ~ crime$Wealth))
```

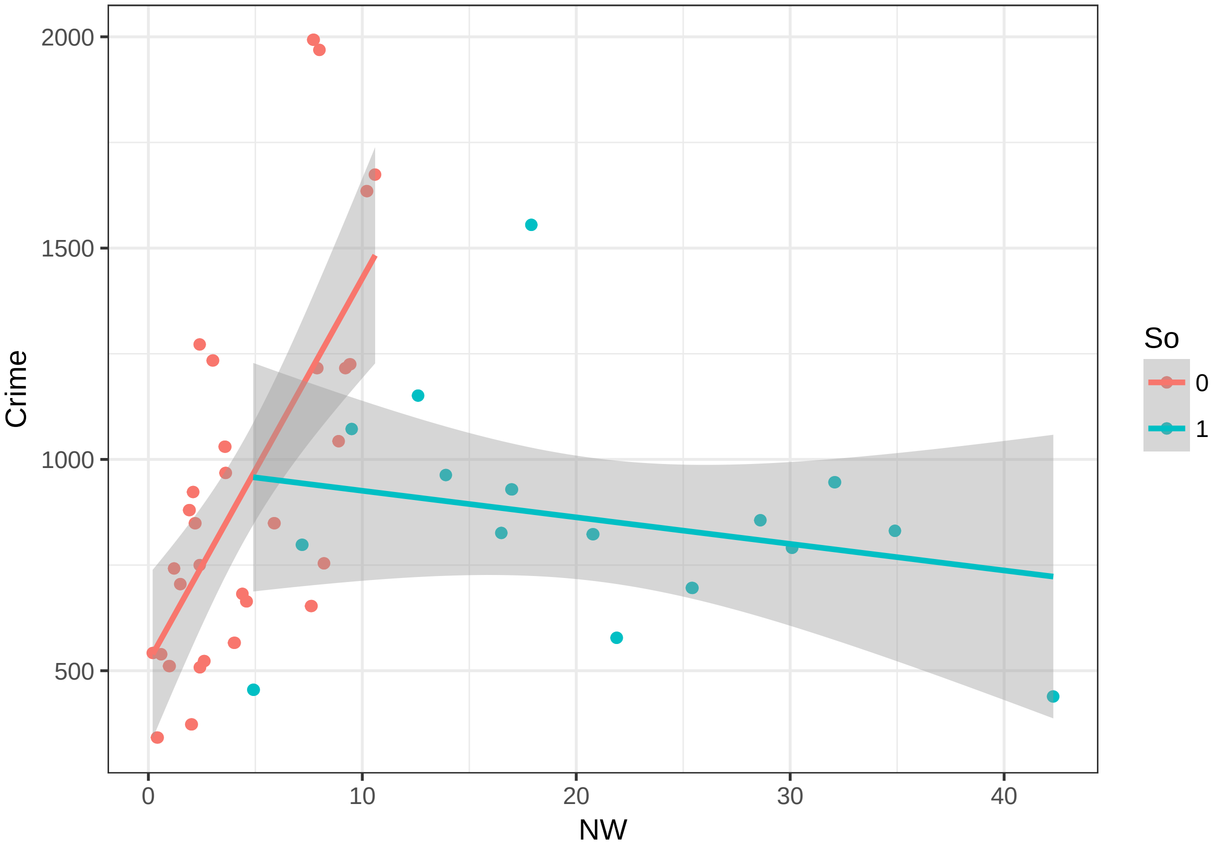
```
##
## Call:
## lm(formula = crime$Crime ~ crime$Wealth)
```

```
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -631.40 -272.84  -46.17  197.25  825.02
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -24.28261  286.31386  -0.085   0.9328
## crime$Wealth   0.17689    0.05362   3.299   0.0019 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 350.9 on 45 degrees of freedom
## Multiple R-squared:  0.1948, Adjusted R-squared:  0.1769
## F-statistic: 10.88 on 1 and 45 DF,  p-value: 0.001902
```

Wealth is the first variable where a relationship is apparent upon our initial visual inspection of the scatter plot. The data points are less dispersed than our previous plots. Also, we can see a slight positive relationship between wealth and crime. This relationship becomes even more apparent when we separate the data into southern and non-southern states. Southern states are represented by teal points while non-southern ones are represented by orange points. The relationship is far more pronounced among the non-southern states than the southern ones. The summary of the linear model confirms our initial suspicions. We have the our most statistically significant result yet, though admittedly a t-value of 3.299 and corresponding p-value of .0019 is not as statistically significant as we may have hoped for at the outset of our initial exploration of the data set. Never-the-less, wealth will definitely be included as a candidate for our final model.

Crime vs NW

```r
ggplot(data = crime, aes(x = NW, y = Crime, col = So)) +
  geom_point() +
  geom_jitter(height = 0.2) +
  geom_smooth(method = "lm") +
  theme_bw()
```



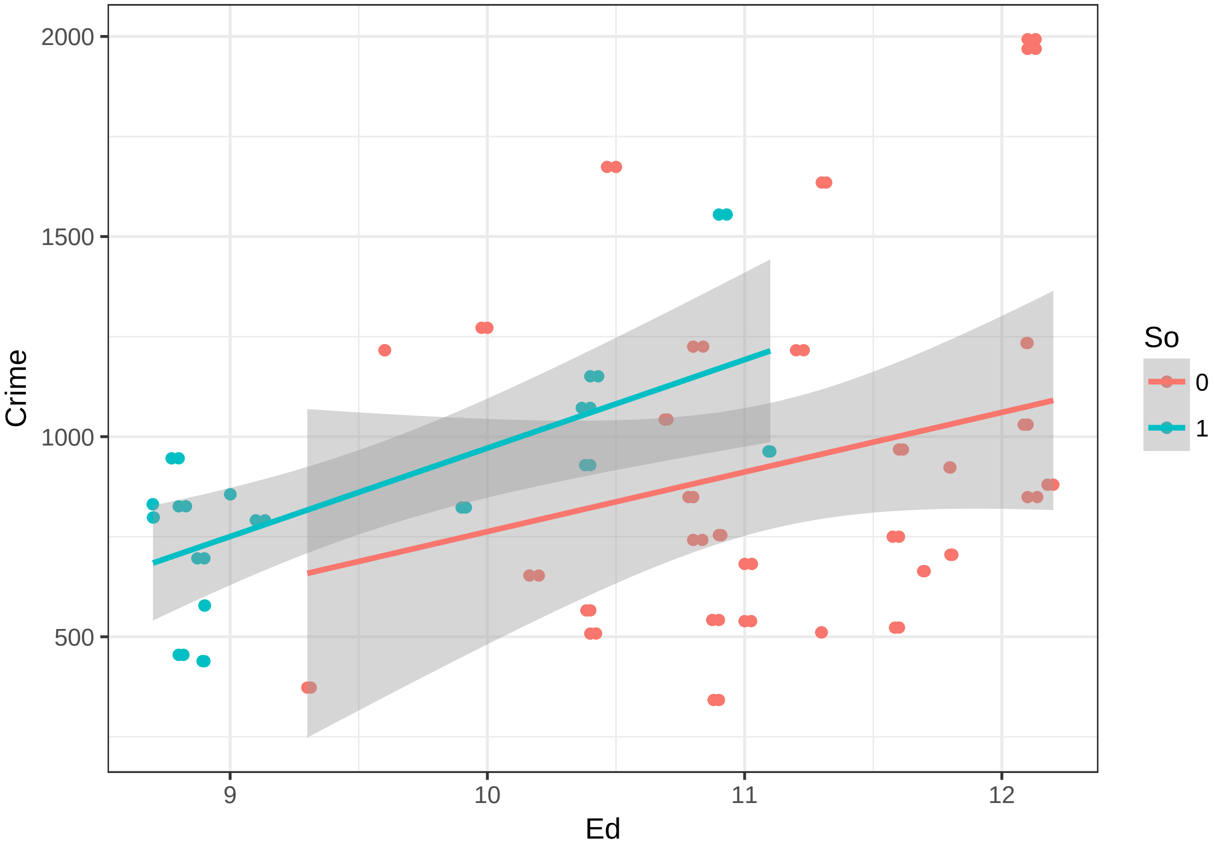
```
summary(lm(crime$Crime ~ crime$NW))
```

```
##
## Call:
## lm(formula = crime$Crime ~ crime$NW)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -551.18 -241.66  -86.92  153.53 1090.87
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  892.686     80.384  11.105 1.76e-14 ***
## crime$NW       1.226      5.604   0.219    0.828
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 390.8 on 45 degrees of freedom
## Multiple R-squared:  0.001063,	Adjusted R-squared:  -0.02114
## F-statistic: 0.04787 on 1 and 45 DF,  p-value: 0.8278
```

This was an interesting variable. If we take the data set as a whole, we see virtually no relationship between the percentage of non-whites in the population and crime. This is confirmed by the summary of the linear model. We have a very low t-value and a very high p-value paired with a virtually non-existent R^2. It seems like there is no statistical significance to this variable, what-so-ever. Upon further investigation, though, we find an extremely strong positive relationship between NW and crime within the non-southern states whereas there appears to be a slight negative relationship within the southern states. This discrepancy between the states is probably why there does not appear to be a relationship when observing the variable within the context of the entire data set. We would definitely like to explore this variable a bit further and so we will include it among our candidates for the final model.

Crime vs Education

```
ggplot(data = crime, aes(x = Ed, y = Crime, col = So)) +
  geom_point() +
  geom_jitter(height = 0.2) +
  geom_smooth(method = lm) +
  theme_bw()
```



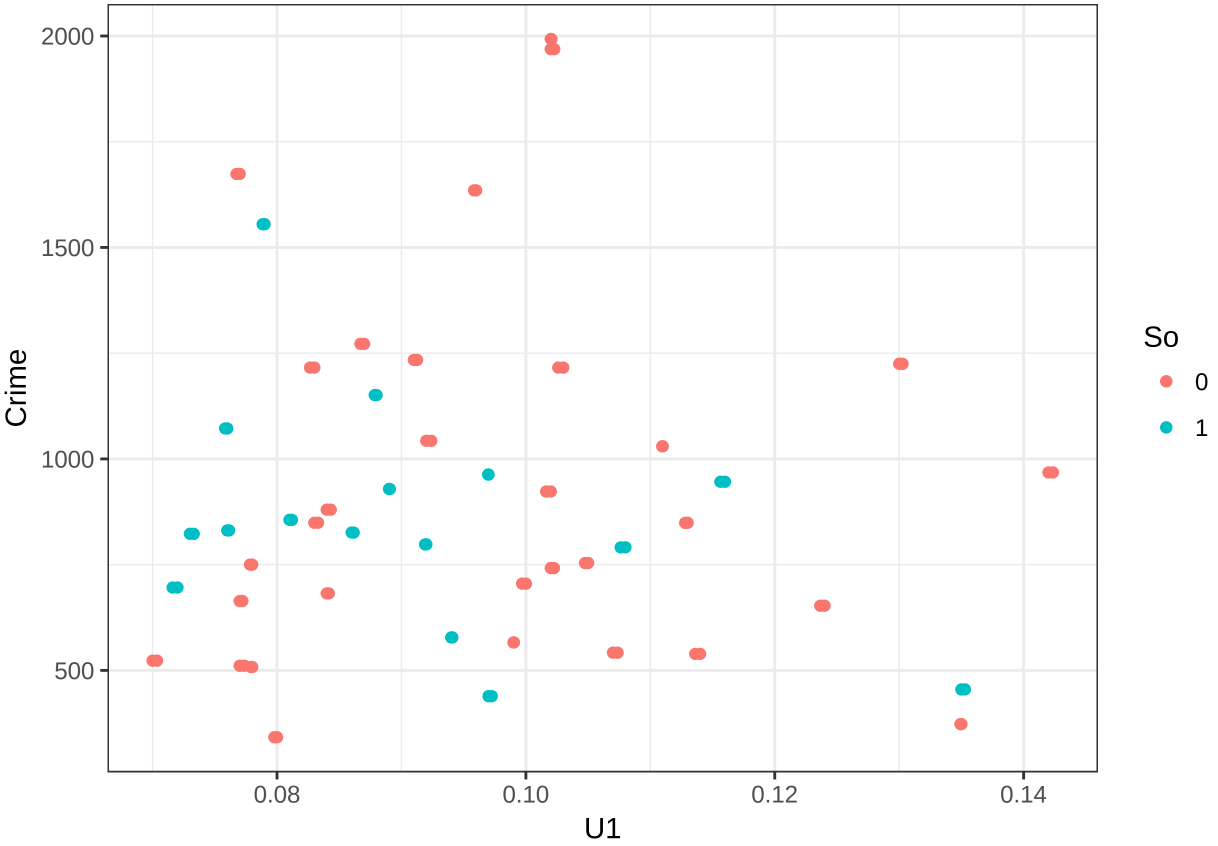
```
summary(lm(crime$Crime ~ crime$Ed))
```

```
##
## Call:
## lm(formula = crime$Crime ~ crime$Ed)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -600.61 -271.25  -46.54  171.33  916.46
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -273.97     518.10  -0.529   0.5996
## crime$Ed      111.61      48.78   2.288   0.0269 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 370.1 on 45 degrees of freedom
## Multiple R-squared:  0.1042, Adjusted R-squared:  0.08432
## F-statistic: 5.236 on 1 and 45 DF,  p-value: 0.02688
```

This was another variable for which the relationship was not when observing the data in the context of the full data set, but upon separating the data into southern and non-southerns states a relationship appeared to exist. The summary of the linear model supports the original conclusion that a strong relationship does not exist. The t-value is not terribly high and the p-value is not terribly low. We did notice, however, that the southern and non-southern states behaved as almost two distinct data sets. The non-southern states appeared to have a significantly higher level of education than the southern ones. In fact, there is only one southern state with an Ed > 11, whereas the majority of the non-southern states had an Ed > 11. We decide to keep Ed as a candidate for our final model.

Crime vs U1 & Crime vs U2

```
ggplot(data = crime, aes(x = U1, y = Crime, col = So)) +
  geom_point() +
  geom_jitter(height = 0.2) +
  theme_bw()
```



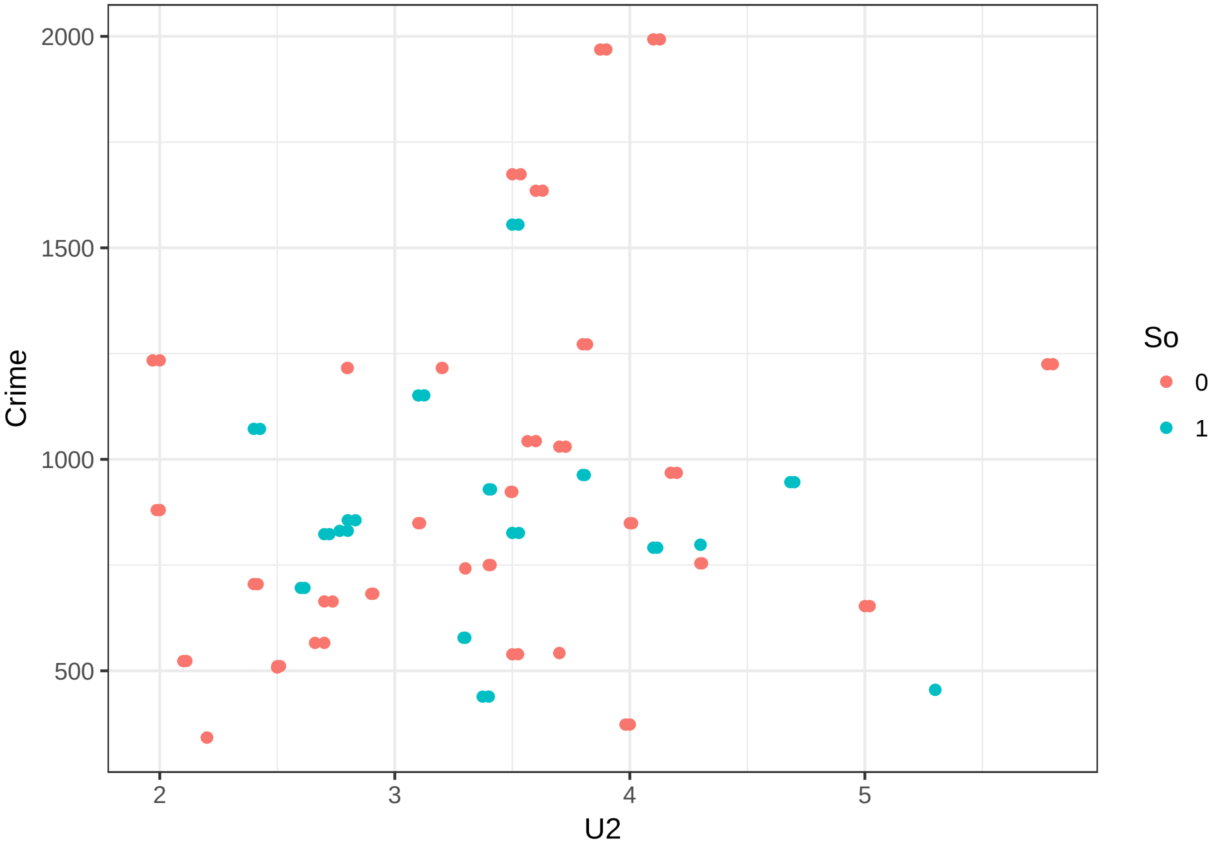
```
summary(lm(crime$Crime ~ crime$U1))
```

```
##
## Call:
## lm(formula = crime$Crime ~ crime$U1)
```

```
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -579.84 -248.29  -89.34  143.78 1094.99
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   1008.5      310.2   3.251  0.00218 **
## crime$U1     -1082.9     3193.9  -0.339  0.73615
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 390.5 on 45 degrees of freedom
## Multiple R-squared:  0.002548,   Adjusted R-squared:  -0.01962
## F-statistic: 0.115 on 1 and 45 DF,  p-value: 0.7362
```

Again we have a variable that does not appear to have a strong relationship with Crime. Separating by states does not make it any more clear, either. The summary of the linear model corroborates our conclusions from the visual inspection. The t-value is extremely low resulting in an extremely high p-value. Furthermore, the R^2 is virtually 0.

```
ggplot(data = crime, aes(x = U2, y = Crime, col = So)) +
  geom_point() +
  geom_jitter(height = 0.2) +
  theme_bw()
```



```
summary(lm(crime$Crime ~ crime$U2))
```

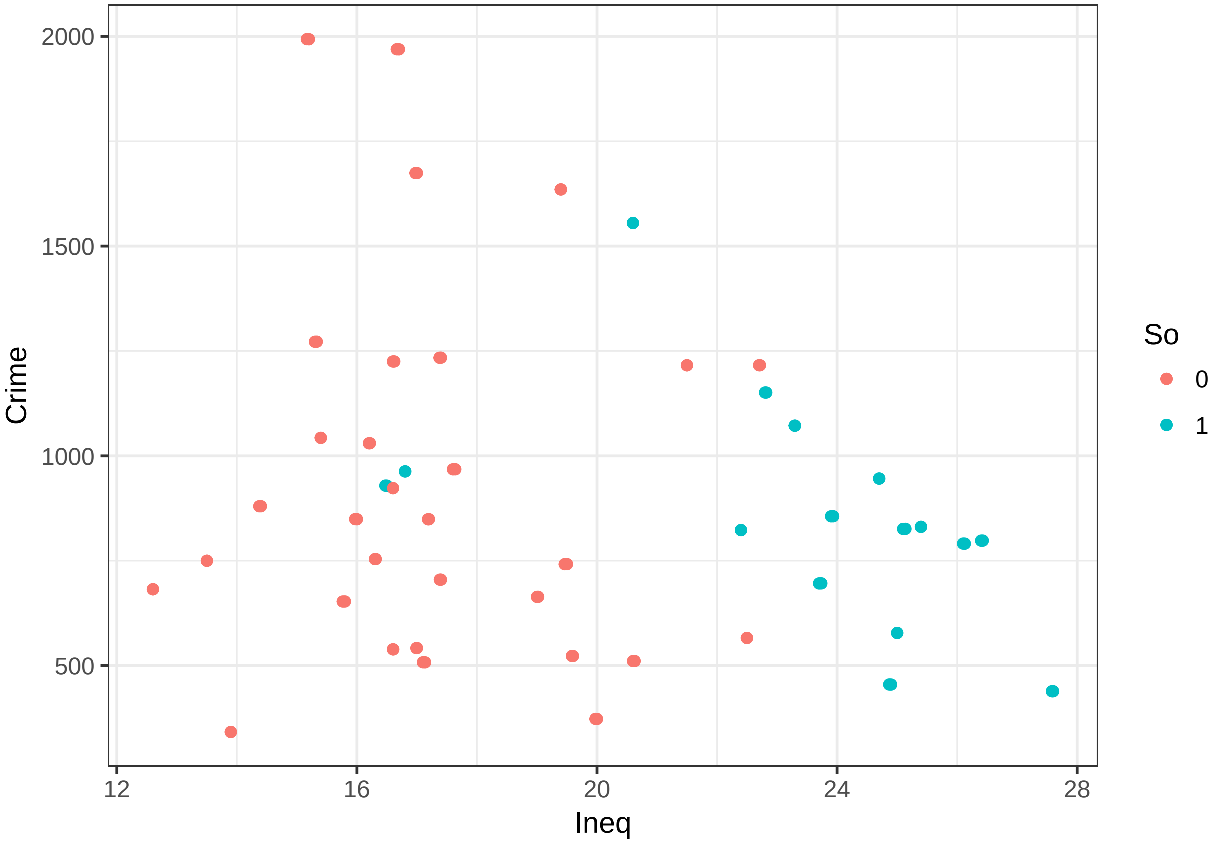
```
##
```

```
## Call:
## lm(formula = crime$Crime ~ crime$U2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -604.55 -250.52  -64.82  123.18 1030.90
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   629.16     235.09   2.676   0.0103 *
## crime$U2       81.20      67.19   1.209   0.2331
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 384.8 on 45 degrees of freedom
## Multiple R-squared:  0.03144,    Adjusted R-squared:  0.009919
## F-statistic: 1.461 on 1 and 45 DF,  p-value: 0.2331
```

As with U1, no relationship is immediately apparent with Crime from the initial plot. While the summary of the linear model shows a higher level of statistical significance for U2 than for U1, it still is not strong enough to merit including in our final model. We will eliminate both U1 and U2 from our list of candidates for our final model.

Crime vs Ineq

```
ggplot(data = crime, aes(x = Ineq, y = Crime, col = So)) +
  geom_point() +
  geom_jitter(height = 0.2) +
  theme_bw()
```



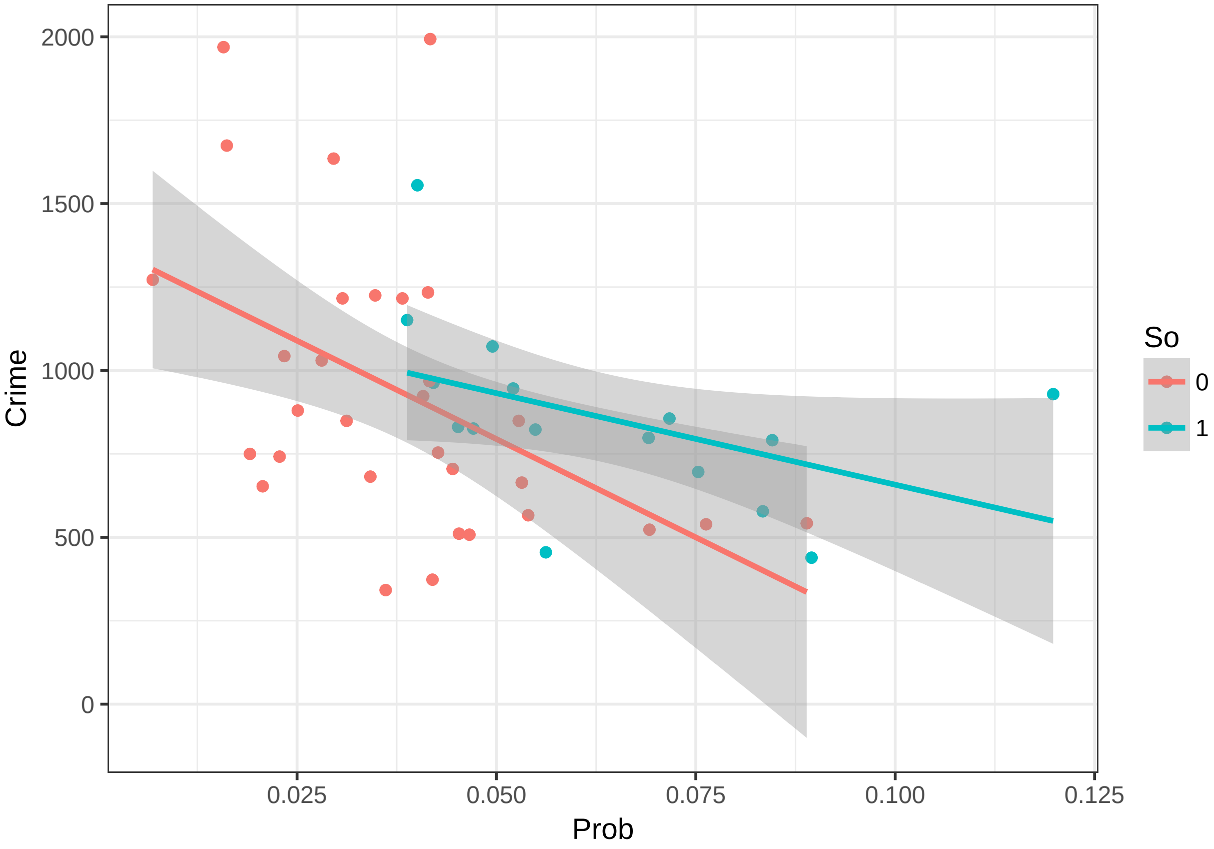
```
summary(lm(crime$Crime ~ crime$Ineq))
```

```
##
## Call:
## lm(formula = crime$Crime ~ crime$Ineq)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -658.54 -271.38  -30.02  183.75 1017.06
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1241.77     281.48   4.412 6.33e-05 ***
## crime$Ineq    -17.36      14.22  -1.221    0.229
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 384.7 on 45 degrees of freedom
## Multiple R-squared:  0.03205,    Adjusted R-squared:  0.01054
## F-statistic:  1.49 on 1 and 45 DF,  p-value: 0.2286
```

Upon inspecting the plot of Crime vs Ineq, no relationship is immediately recognizable. There is far too much scatter of the data to reasonably conclude any relationship. We noticed that from Ineq = 10 to Ineq = 15 the data virtually spanned the entire range of Crime. The summary of the linear model confirmed our initial conclusion. There was a very low R^2 coupled with a low t-value and a relatively high corresponding p-value. Ineq will not be included among our candidates for our final model.

Crime vs Prob

```
ggplot(data = crime, aes(x = Prob, y = Crime, col = So)) +
  geom_point() +
  geom_jitter(height = 0.2) +
  geom_smooth(method = "lm") +
  theme_bw()
```



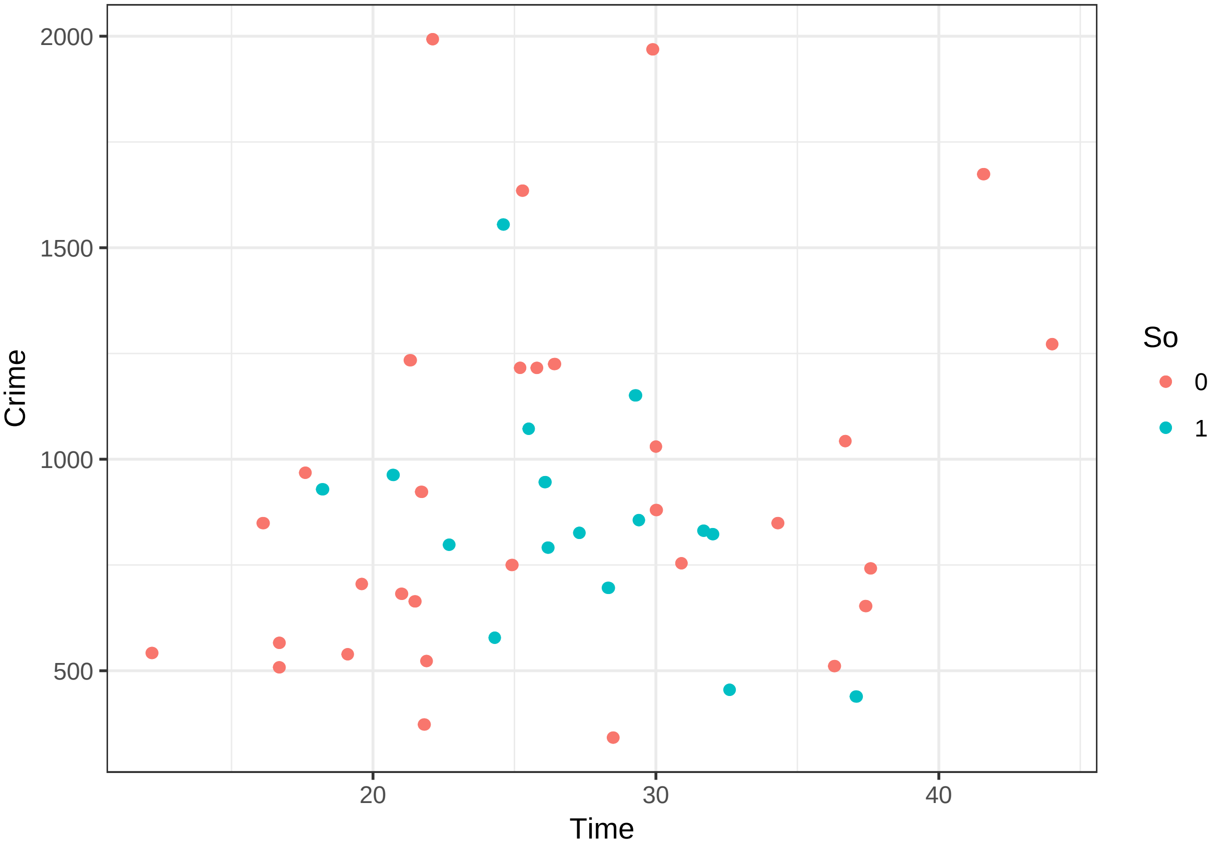
```
summary(lm(crime$Crime ~ crime$Prob))
```

```
##
## Call:
## lm(formula = crime$Crime ~ crime$Prob)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -643.01 -207.81  -27.83  171.53 1048.70
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   1247.5      119.6  10.427 1.38e-13 ***
## crime$Prob   -7270.6     2292.4  -3.172  0.00273 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 353.5 on 45 degrees of freedom
## Multiple R-squared:  0.1827, Adjusted R-squared:  0.1645
## F-statistic: 10.06 on 1 and 45 DF,  p-value: 0.00273
```

We noticed a negative relationship between Crime and Prob. While the data is significantly scattered in the early range of Prob, it is far less scattered as Prob increases. Furthermore, despite the relatively broad scatter of points in the early portion of Prob's range, we could still visualize a negative relationship. The analysis of the linear model confirmed our conclusion. Crime~Prob resulted in one of our highest R^2 values as well as providing statistically significant t-values and p-values. We will certainly consider Prob for our final model.

Crime vs Time

```
ggplot(data = crime, aes(x = Time, y = Crime, col = So)) +
  geom_point() +
  geom_jitter(height = 0.2) +
  theme_bw()
```



```
summary(lm(crime$Crime ~ crime$Time))
```

```
##
## Call:
## lm(formula = crime$Crime ~ crime$Time)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -578.64 -249.20  -84.83  156.20 1124.70
##
## Coefficients:
```

```
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  687.545    221.250   3.108  0.00326 **
## crime$Time     8.179      8.044   1.017  0.31468
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 386.6 on 45 degrees of freedom
## Multiple R-squared:  0.02246,    Adjusted R-squared:  0.0007367
## F-statistic: 1.034 on 1 and 45 DF,  p-value: 0.3147
```

Once again we see some pretty significant scattering of the data points through the range of Time. There is no immediately obvious relationship between Time and Crime even when considering the data separately depending on whether it is a southern or non-southern state. The summary of our linear model confirms our conclusion from our inspection of the plot generated. We will ignore Time when developing our final model.

Initial Formulation of Our Final Model

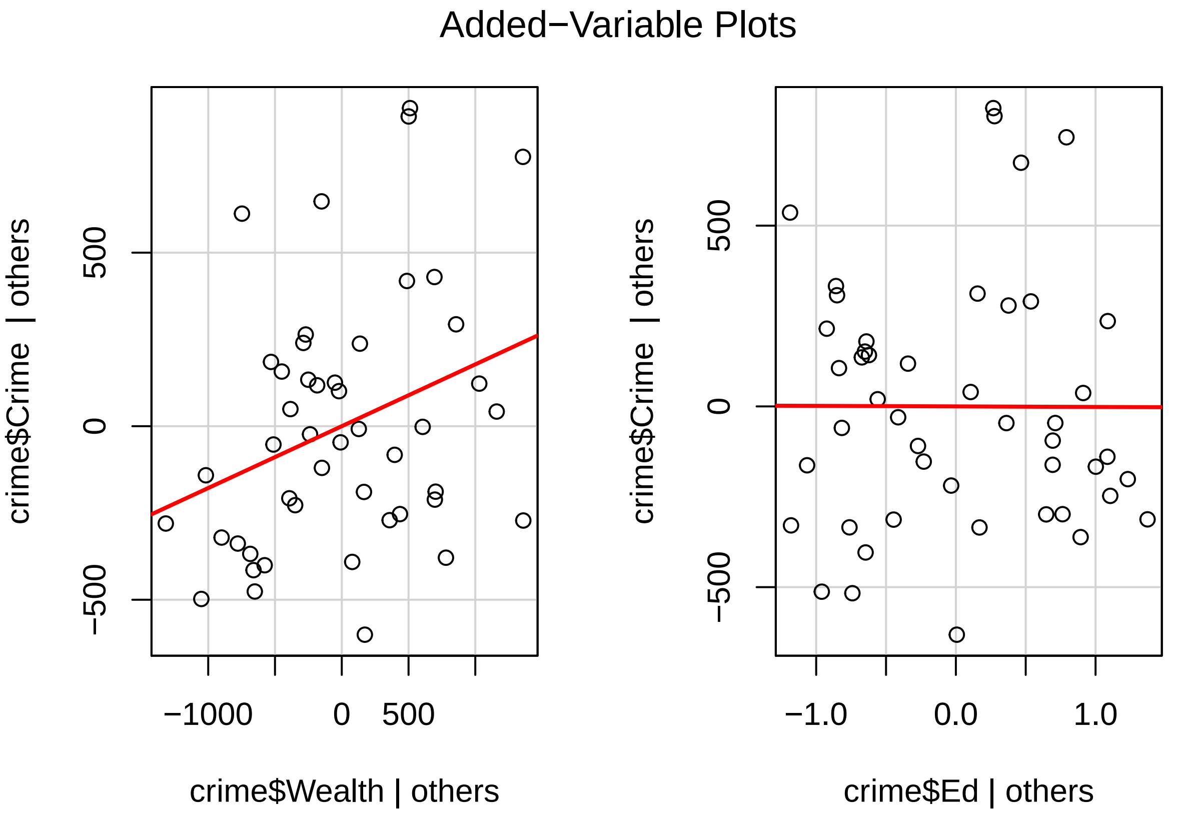
Having completed our initial analysis on our original variables of interest, we have narrowed our pool of possible predictors down to Population (Pop), Education (Ed), Probably of incarceration (Prob), Wealth (Wealth) and whether or not a state is considered Southern (So). The question was where to start our model. Given the analysis of each variable individually against Crime, we knew we would include Wealth in our final model. Wealth had the highest R^2 of any of our variables meaning it was the best at explaining the variation within our response variable, Crime. Since wealth would definitely be included in our final model, we decided it was the best place to start. Next we had to decide which other variables to include.

Added Variable Plots

Our plan was to formulate added variable plots for each of our other candidates and using the information from the plots to decide whether or not to include the variable. If a variable was included in our model, it would remain and further added variable plots would be generated.

Education

```
av.plots(lm(crime$Crime ~ crime$Wealth + crime$Ed))
```

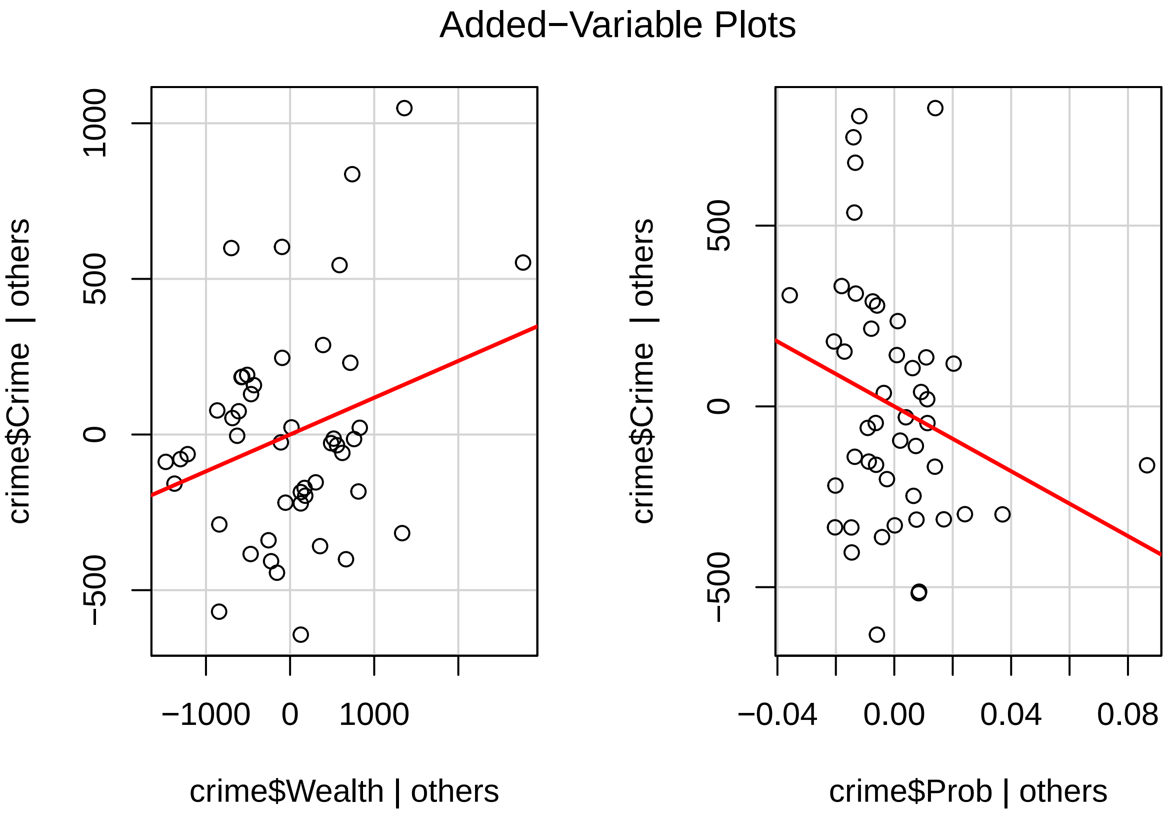


We are using the av.plots() function from the 'car' package. When given a linear model, av.plots() will construct added variable plots for each variable versus a model including all of the other variables. This being our first attempt at an added variable plot, there are only 2 added variable plots. We are really only concerned with the second plot since the first is added variable plot for Wealth and we have already decided to include wealth in our model. The second plot is what we are interested in. As we can see, we essentially have null-plot. The residuals of Ed ~ Wealth (along the x-axis) are not correlated in any way with the residuals from Crime ~ Wealth (our final model so far). We conclude there is no reason to include Education in our model and proceed to the other variables.

Probability of Incarceration

```
av.plots(lm(crime$Crime ~ crime$Wealth + crime$Prob))
```

```
## Warning: 'av.plots' is deprecated.
## Use 'avPlots' instead.
## See help("Deprecated") and help("car-deprecated").
```

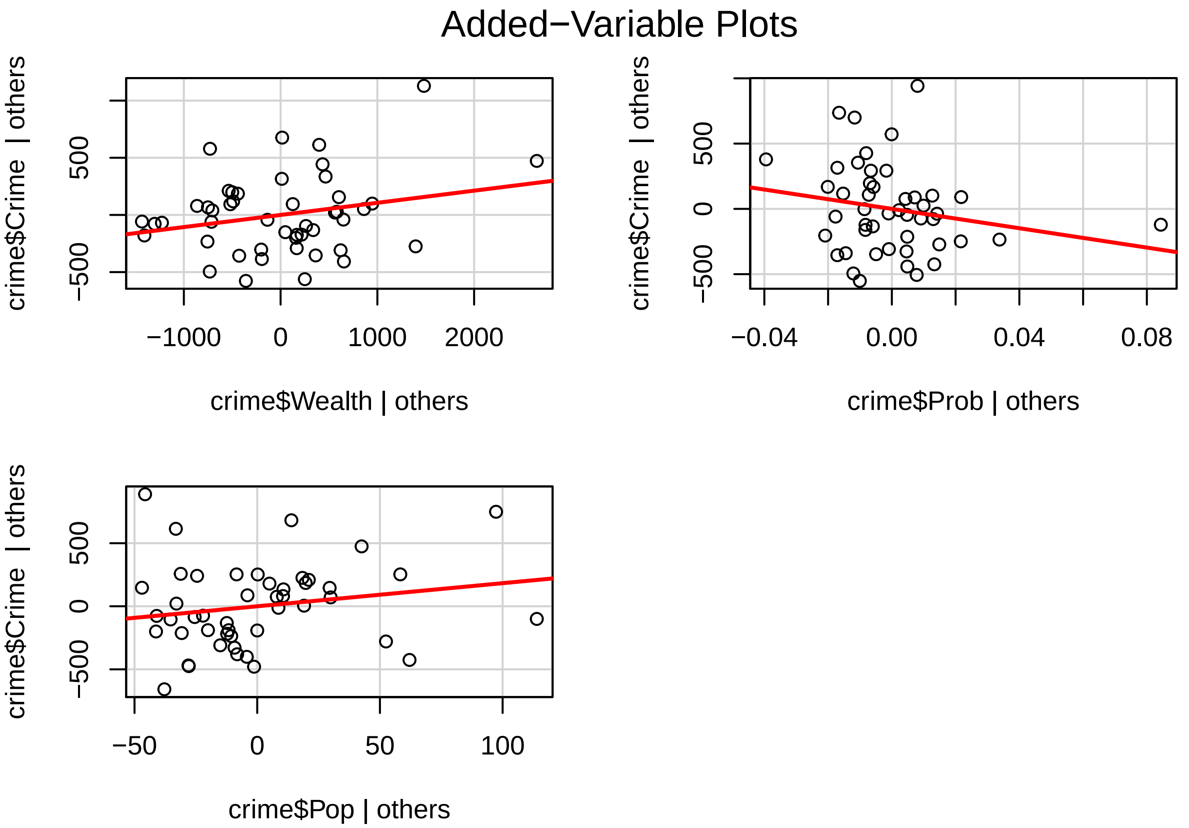


Once again, we will ignore the first plot since we have already decided to include Wealth in our model. Unlike the added variable plot for education, we notice what appears to be a negative relationship between the residuals of Crime ~ Wealth and Prob ~ Wealth. The added variable plot reveals that part of the variation in Crime ~ Wealth is explained by the variation within Prob ~ Wealth. We decide to include Prob in our final model and proceed to the next variable.

Population

```
av.plots(lm(crime$Crime ~ crime$Wealth +
              crime$Prob +
              crime$Pop))
```

```
## Warning: 'av.plots' is deprecated.
## Use 'avPlots' instead.
## See help("Deprecated") and help("car-deprecated").
```

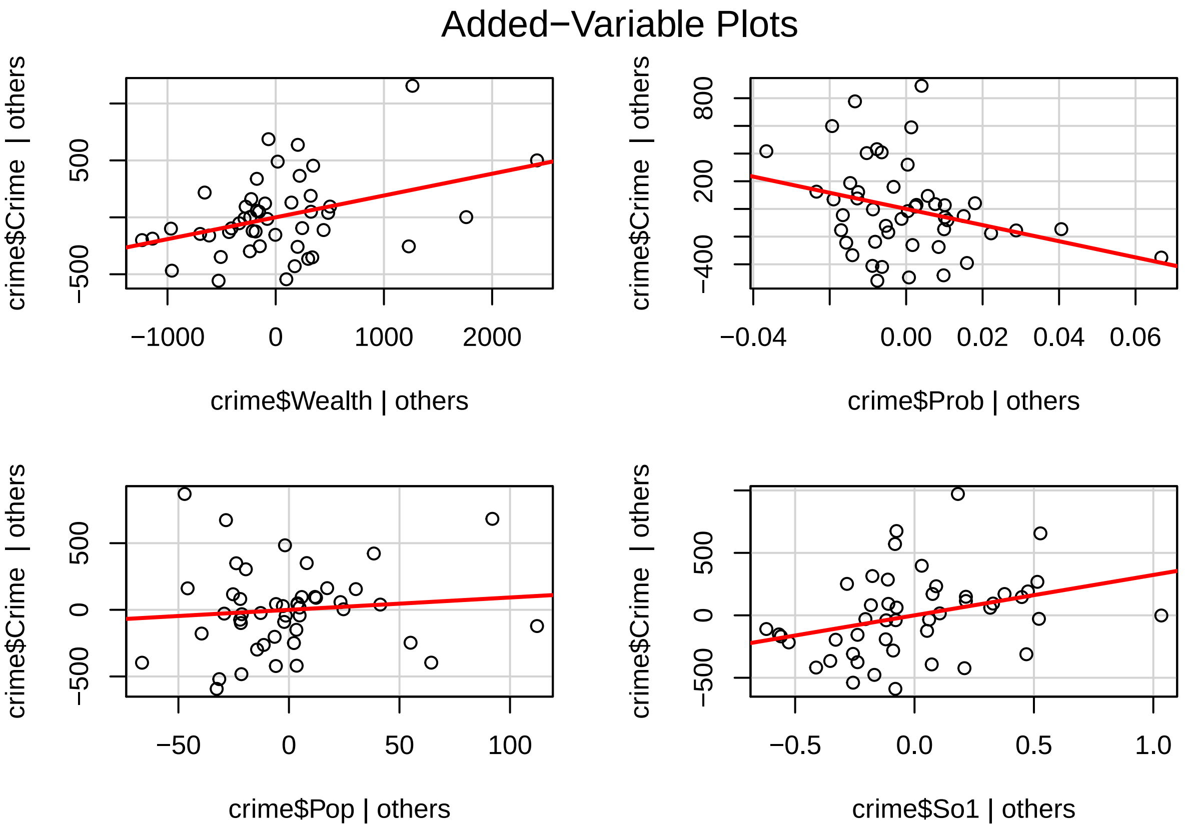


Since we decided to keep Prob for our final model, our next added variable plots will include both Wealth and Prob. Since we've already determined Wealth and Prob will be included in our model, the only plot we are focusing on for our decision on whether to keep or discard Population is the last one. The added variable plot is not great, but it's also not a null plot. As we can see by the regression line included in the plot, there is a slight positive relationship between the residuals of Crime ~ Wealth + Prob and the residuals of Pop ~ Wealth + Prob. Therefore, population does help explain at least some part of the variation that exist in our current incarnation of what will ultimately be our final model. We will add population and proceed.

Southern States/Non-southern states

```
av.plots(lm(crime$Crime ~
              crime$Wealth +
              crime$Prob +
              crime$Pop +
              crime$So))
```

```
## Warning: 'av.plots' is deprecated.
## Use 'avPlots' instead.
## See help("Deprecated") and help("car-deprecated").
```



So was the only factor variable available to us in our data. We noticed in our single variable analysis that including the factor in our analysis revealed relationships that were not apparent when only analyzing the data in the context of the entire data set. Once again, we will only focus on the last added variable plot when making our decision to keep or discard So. This added variable plot does show a relationship between the residuals of Crime ~ Wealth + Prob + Pop and the residuals of So ~ Wealth + Prob + Pop indicating So should be include in our final model.

Refining our Final Model

At this point we have settled on the predictor variables we would like to focus our attention on for our final model. We have narrowed it down to wealth (Wealth), population (Pop), probability of an arrest leading to incarceration (Prob) and the factor indicating whether or not a state is considered southern (So). We decide to put the predictors together into our final model and run a summary on the multi-variable regression model.

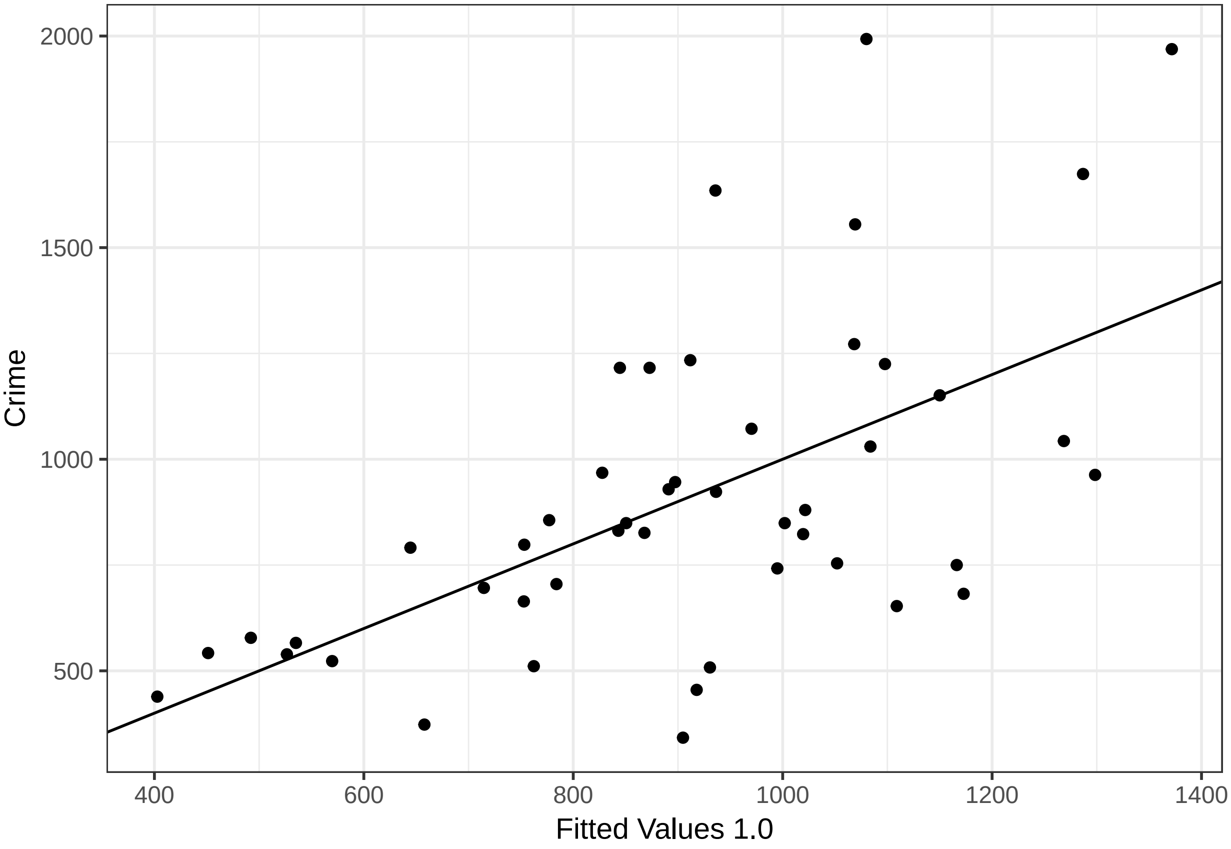
```r
final.model.1 <-lm(crime$Crime ~
                     crime$Wealth +
                     crime$Pop +
                     crime$Prob +
                     crime$So)
summary(final.model.1)
```

```
##
## Call:
## lm(formula = crime$Crime ~ crime$Wealth + crime$Pop + crime$Prob +
##     crime$So)
##
## Residuals:
```

```
##     Min      1Q  Median      3Q     Max
## -562.81 -211.07   -1.54  114.52  913.00
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)   2.989e+01  4.388e+02   0.068   0.9460
## crime$Wealth  1.915e-01  7.107e-02   2.695   0.0101 *
## crime$Pop     9.252e-01  1.417e+00   0.653   0.5173
## crime$Prob   -5.834e+03  2.767e+03  -2.109   0.0410 *
## crime$So1     3.232e+02  1.410e+02   2.293   0.0269 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 325.9 on 42 degrees of freedom
## Multiple R-squared:  0.3519, Adjusted R-squared:  0.2902
## F-statistic: 5.701 on 4 and 42 DF,  p-value: 0.0009299
```

Considering our best R^2 from our single variable analysis was less than .20 (Wealth), these findings are not terrible. We have an $R^2 = .3519$ and 3 of our 4 are statistically significant. We decide to explore how our fitted values compare to the actual values.

```
crime$'Fitted Values 1' <- fitted(final.model.1)
ggplot(crime, aes(x = crime$`Fitted Values 1`, y = crime$Crime)) +
  geom_point() +
  geom_abline() +
  ylab("Crime") +
  xlab("Fitted Values 1.0") +
  theme_bw()
```



The line represents perfect correlation between fitted values and Crime. We actually are happy with the progress we have made thus far. There is definitely a relationship that is emerging from our model, yet we do see some slight curvature in the data points. There is almost a periodic quality to the data points. We decide to try and improve our model by exploring the options of power and logarithmic transformations for our variables.

The Range Rule and Power Transform

```
sapply(crime, range)
```

```
##      M      So  Ed     Po1    Po2    LF      M.F     Pop   NW     U1     
## [1,] "11.9" "0" "8.7"  "4.5"  "4.1"  "0.48"  "93.4"  "3"   "0.2"  "0.07" 
## [2,] "17.7" "1" "12.2" "16.6" "15.7" "0.641" "107.1" "168" "42.3" "0.142"
##      U2    Wealth Ineq   Prob       Time      Crime  Fitted Values 1   
## [1,] "2"   "2880" "12.6" "0.0069"   "12.1996" "342"  "402.754097976135"
## [2,] "5.8" "6890" "27.6" "0.119804" "44.0004" "1993" "1371.65930616165"
```

To determine if any of our variables would benefit from transformation, we apply the range rule. The range rule dictates that any variable that spans less than one order of magnitude would not be helped by transformation. So, any that do span more than one may benefit from a transformation. After finding the ranges of each variable, we find that NW, Pop and Prob would all potentially benefit from a transformation.

Logarithmic transformation is not the only transformation available to us. It is also common to use power transformations in order to create more linear data models.

PowerTransform

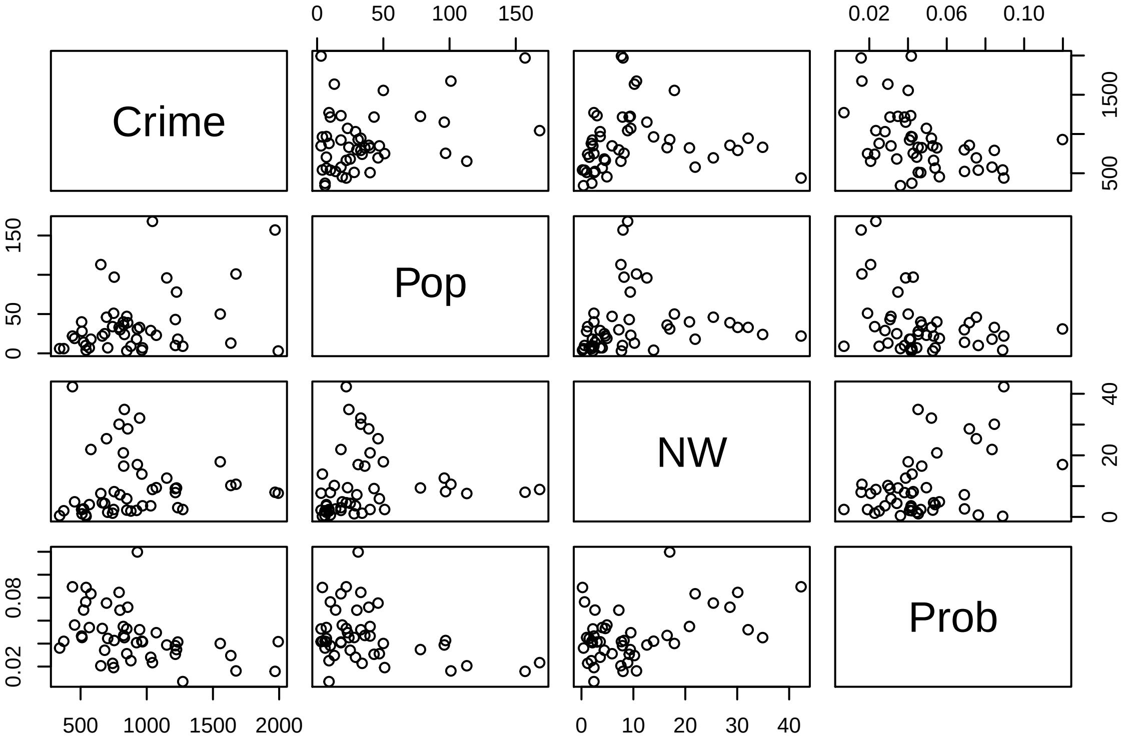
```
suppressMessages(library(GGally))
crime.subset <- subset(crime, select = c('Crime', 'Pop', 'NW', 'Prob'))
tmp <- powerTransform ( cbind ( Pop, NW, Prob ) ~ 1, data = crime.subset)
summary(tmp)
```

```
## bcPower Transformations to Multinormality
##      Est Power Rounded Pwr Wald Lwr bnd Wald Upr Bnd
## Pop     0.0797        0.00      -0.1695       0.3289
## NW      0.2166        0.33       0.0022       0.4310
## Prob    0.4039        0.50       0.0109       0.7968
##
## Likelihood ratio tests about transformation parameters
##                                 LRT df       pval
## LR test, lambda = (0 0 0)  8.555257  3 0.03582744
## LR test, lambda = (1 1 1) 93.382544  3 0.00000000
```

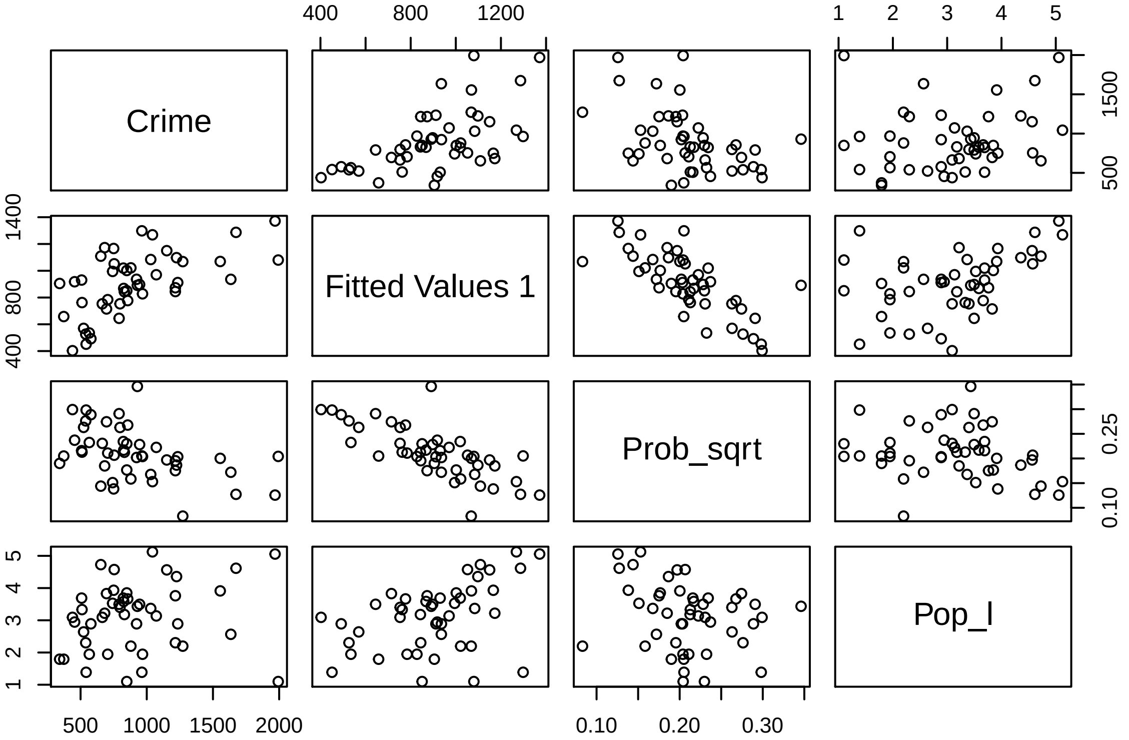
```
tmp2 <- powerTransform ( cbind ( Crime, Pop, NW, Prob ) ~ 1, data = crime.subset)
```

This Power Transform tells us that we should be using $log(Pop)$, $\sqrt[3]{NW}$, and $\sqrt{Prob}$. These look like they might straighten out the plots, so let's look at the before and after pairs plots.

```
pairs(crime.subset)
```



```
crime <- within(crime, {NW_cbrt <- sign(NW) * abs(NW)^(1/3)
                        Pop_l = log(Pop)
                        Prob_sqrt = sqrt(Prob)})
pairs(crime[16:19])
```



These transformations really helped straighten these relationships out, so we think they will positively impact our model. After using Power Transform to see what transformations we should make to our predictor variables, we used Power Transform again to see if we should transform our response variable.

```
crime.subset <- within(crime.subset, {NW_cbrt <- sign(NW) * abs(NW)^(1/3)
                        Pop_l = log(Pop)
                        Prob_sqrt = sqrt(Prob)})
tmp2 <- powerTransform ( cbind ( Crime, Pop_l, NW_cbrt, Prob_sqrt)
                        ~ 1, data = crime.subset)
summary(tmp2)
```

```
## bcPower Transformations to Multinormality
##           Est Power Rounded Pwr Wald Lwr bnd Wald Upr Bnd
## Crime       -0.0238           0      -0.5697       0.5221
## Pop_l        1.0844           1       0.3656       1.8032
## NW_cbrt      0.4788           1      -0.1080       1.0656
## Prob_sqrt    0.7273           1       0.0195       1.4351
##
## Likelihood ratio tests about transformation parameters
##                                  LRT df        pval
## LR test, lambda = (0 0 0 0) 15.82886  4 0.003257615
## LR test, lambda = (1 1 1 1) 16.02379  4 0.002987404
```

R suggested we use $log(Crime)$ instead of $Crime$, so we tried it and it helped out the model quite a bit:

```
summary(lm(log(crime$Crime) ~ crime$NW_cbrt + crime$Pop_l +
             crime$Wealth + crime$Prob_sqrt + So))
```

```
##
## Call:
## lm(formula = log(crime$Crime) ~ crime$NW_cbrt + crime$Pop_l +
##     crime$Wealth + crime$Prob_sqrt + So)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.6632 -0.1482  0.0208  0.2002  0.4833
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)       5.594e+00  5.233e-01  10.692 1.99e-13 ***
## crime$NW_cbrt     3.852e-01  1.050e-01   3.667 0.000698 ***
## crime$Pop_l      -6.591e-02  5.116e-02  -1.288 0.204878
## crime$Wealth      2.320e-04  6.386e-05   3.634 0.000770 ***
## crime$Prob_sqrt -3.005e+00  1.131e+00  -2.656 0.011210 *
## So1               2.809e-02  1.723e-01   0.163 0.871289
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2961 on 41 degrees of freedom
## Multiple R-squared:  0.5374, Adjusted R-squared:  0.481
## F-statistic: 9.527 on 5 and 41 DF,  p-value: 4.364e-06
```

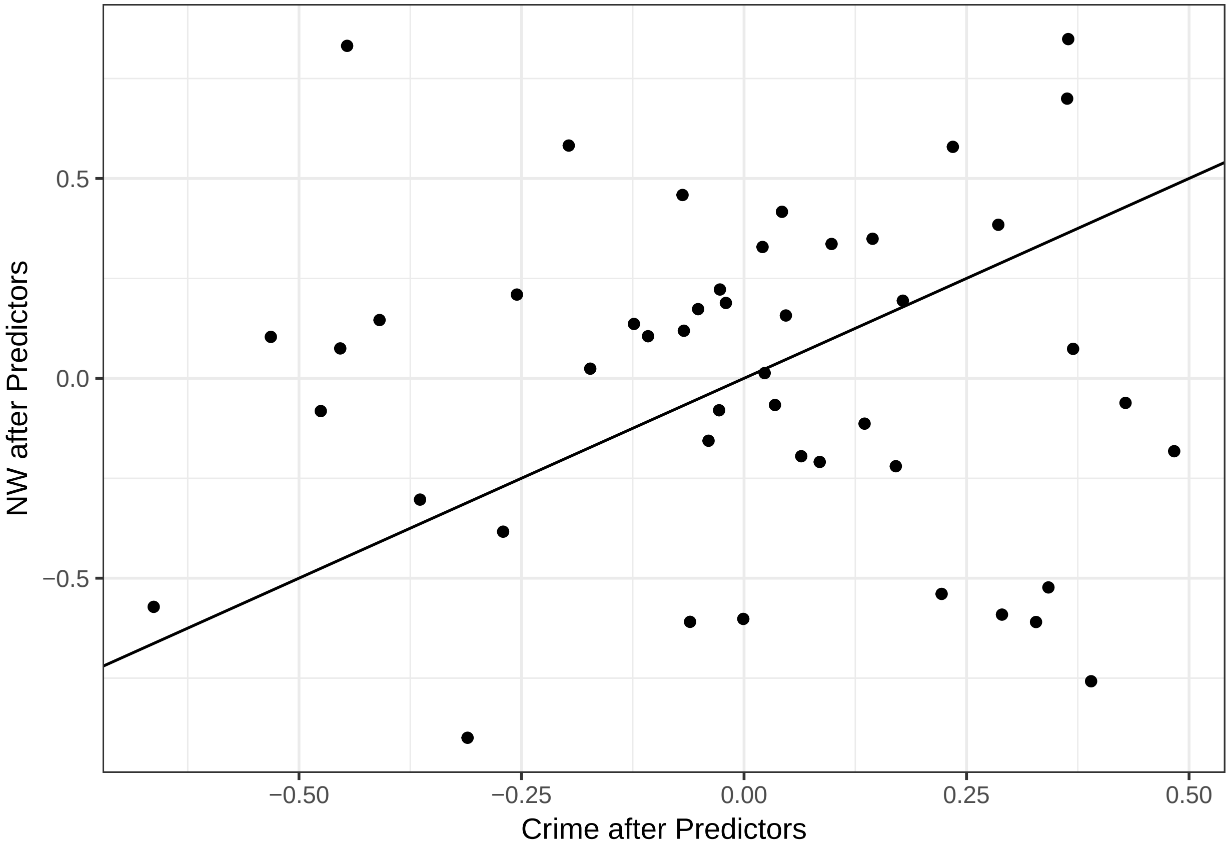
We then went back and looked at the newly transformed variable NW. We didn't use it in our previous model because it didn't really have any relationship with Crime, but since we transformed them both we decided to go back and have another look.

```
resid1 <- residuals(lm(log(crime$Crime) ~ crime$NW_cbrt + crime$Pop_l +
             crime$Wealth + crime$Prob_sqrt + So))
resid2 <- residuals(lm(crime$NW_cbrt ~ crime$NW_cbrt + crime$Pop_l +
            crime$Wealth + crime$Prob_sqrt + So))
```

```
## Warning in model.matrix.default(mt, mf, contrasts): the response appeared
## on the right-hand side and was dropped
```

```
## Warning in model.matrix.default(mt, mf, contrasts): problem with term 1 in
## model.matrix: no columns are assigned
```

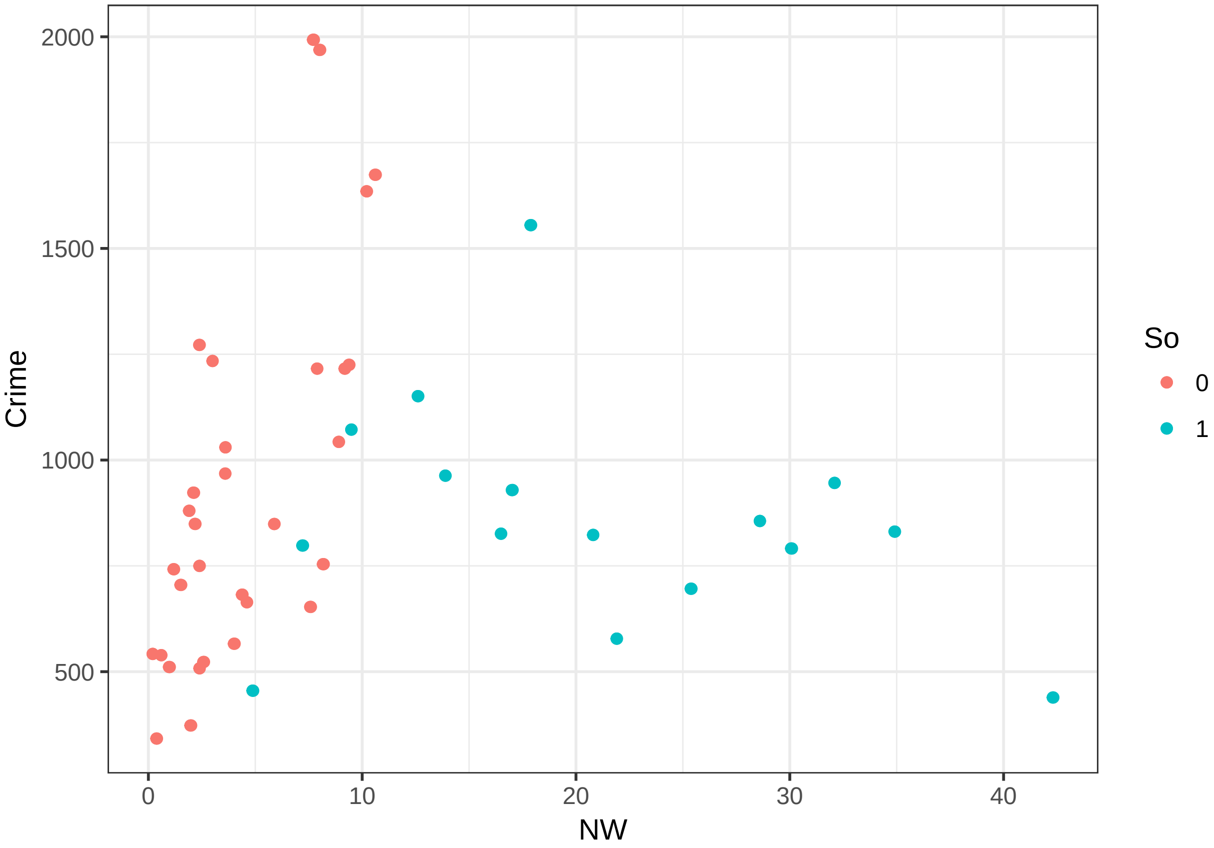
```
ggplot(crime, aes(x = resid1, y = resid2)) +
  geom_point() +
  geom_abline() +
  ylab("NW after Predictors") +
  xlab("Crime after Predictors") +
  theme_bw()
```



This is a better relationship than most other predictors we have seen so far, so we were happy to add it to our model.

After looking at our final model, we realized that So was not significant to any degree and quickly realized it was due to our audition of NW. Looking at the plot of NW ~ Crime with colors indicating So, it was quite obvious that if we know that $NW < 12$ then the state is northern. This means that knowing NW will tell us the value of So in all but three cases.

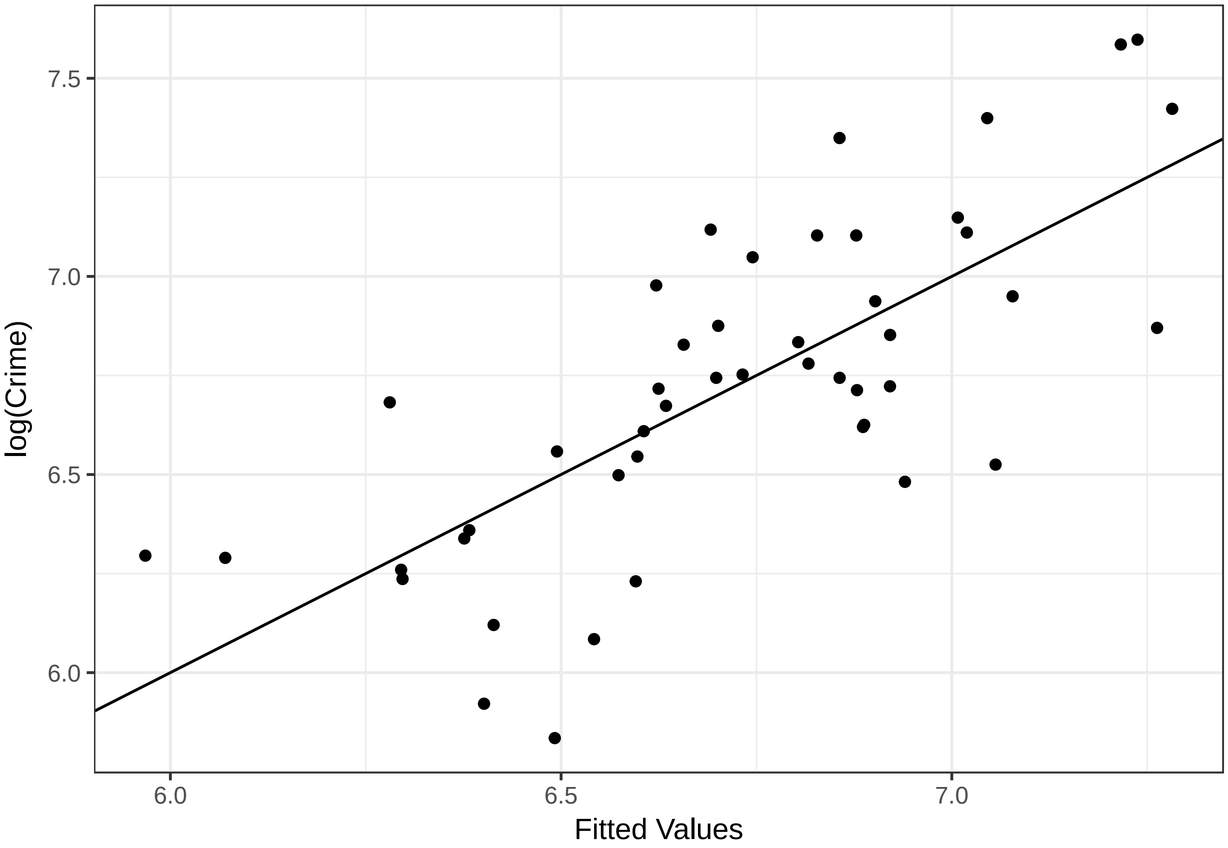
```
ggplot(data = crime, aes(x = NW, y = Crime, col = So)) +
  geom_point() +
  geom_jitter(height = 0.2) +
  theme_bw()
```



As a result we dropped So as a predictor.

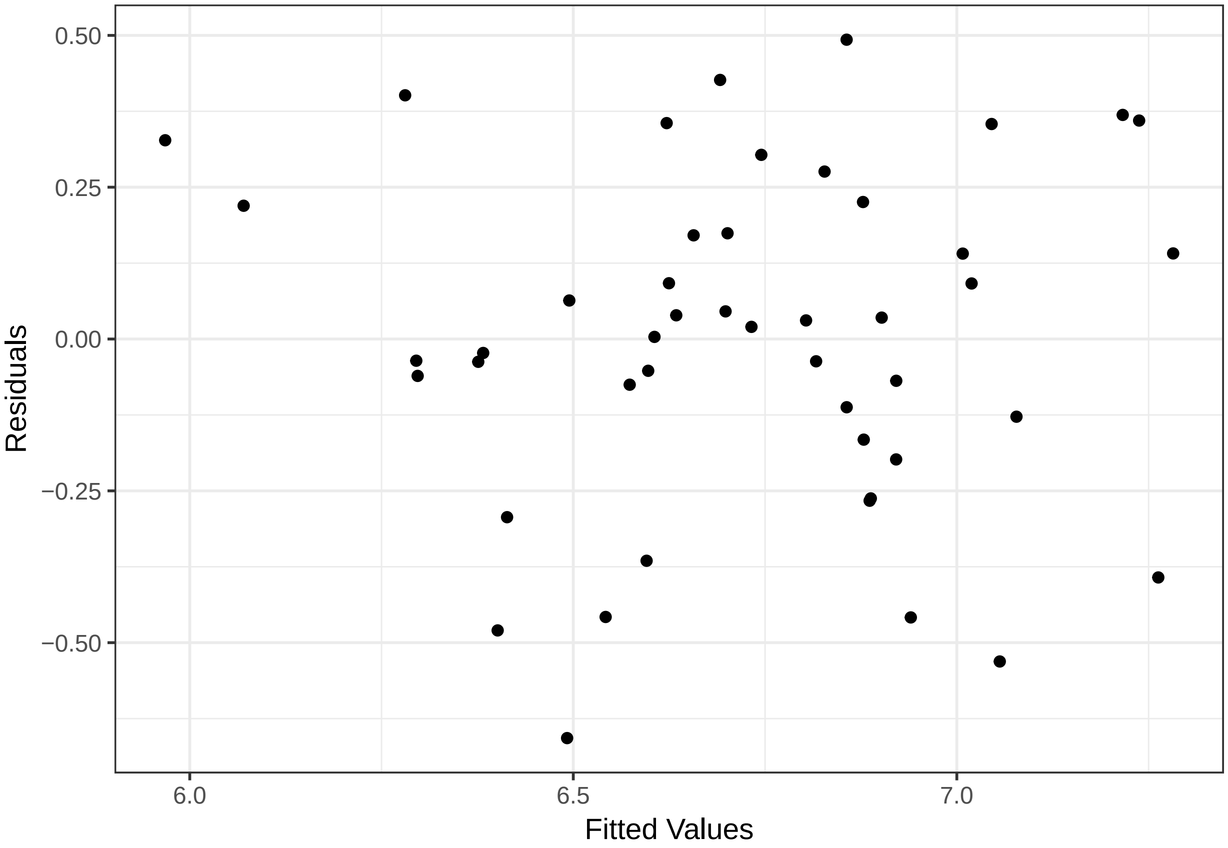
Our final plot of shows a much better linear correlation, but there was a small problem; the graph quite clearly had a sinusoidal or polynomial pattern:

```
final.model.2 <- lm(log(crime$Crime) ~ crime$NW_cbrt + crime$Pop_l +
             crime$Wealth + crime$Prob_sqrt)
fitted1 <- fitted(final.model.2)
ggplot(data = crime, aes(x = fitted1, y = log(Crime))) +
  geom_point() +
  theme_bw() +
  xlab("Fitted Values") +
  ylab("log(Crime)") +
  geom_abline()
```



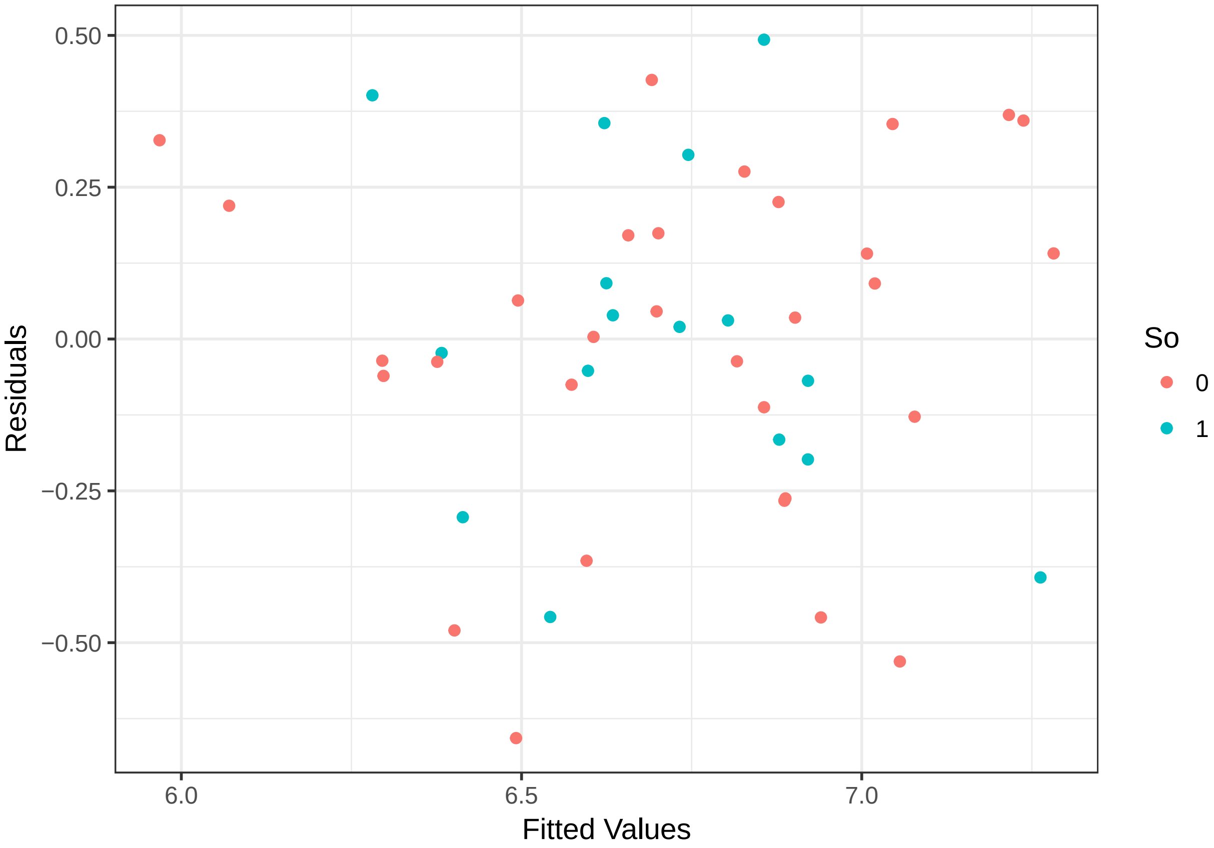
This is also quite clear in the graph of Fitted ~ Residuals:

```
resid3 <- residuals(final.model.2)
ggplot(data = crime, aes(x = fitted1, y = resid3)) +
  geom_point() +
  theme_bw() +
  xlab("Fitted Values") +
  ylab("Residuals") +
  geom_abline()
```



We tried sinusoidal regressions, splines, polynomial regressions, and ARIMA, but nothing we tried could help solve our problem. Lastly, we colored out data points by So and saw the same polynomial pattern in both the southern and non-southern states, so we knew we made the right choice in dropping it as a predictor:

```
ggplot(data = crime, aes(x = fitted1, y = resid3, col = So)) +
  geom_point() +
  theme_bw() +
  xlab("Fitted Values") +
  ylab("Residuals") +
  geom_abline()
```



We think that we could have done better finding a model to predict the crime rate of a state if the data had not been compressed into less specific variables, or if we could do more research into what was causing the polynomial pattern in our model.