Multivariate Regression Analysis

The goal of this situation was to find a good fit for the *uscrime* data set using linear regression, and to predict a new point given a set of new parameters.

Data Pre-Processing

Normality

Before fitting a model, the distribution of each variable needed to be examined. For linear regression, each variable following a normal distribution is an important assumption for fitting a model. To examine the normality of each variable, the Shapiro-Wilk Test for normality was used.

```
norm.df = data.frame()
for(i in 1:16){
  norm.df[i,'col'] = colnames(crime[i])
  norm.df[i,'normality p-value'] = round(shapiro.test(crime[[i]])$p.value,4)
}
norm.df
```

```
##       col normality p-value
## 1       M            0.0361
## 2      So            0.0000
## 3      Ed            0.0032
## 4     Po1            0.0043
## 5     Po2            0.0066
## 6      LF            0.1720
## 7     M.F            0.0038
## 8     Pop            0.0000
## 9      NW            0.0000
## 10     U1            0.0086
## 11     U2            0.2133
## 12 Wealth            0.3375
## 13   Ineq            0.0132
## 14   Prob            0.0179
## 15   Time            0.6132
## 16  Crime            0.0019
```

From this, it can be seen that the variables LF, Time, Wealth, and U2 all have p-values below the threshold of 0.05, showing that they are not normally distributed. When examining the residual diagnostics of the final model selected, if the residuals are not normally distributed then any of these variables used will be standardized to a normal distribution to help remedy this problem. If there are no issues in the residuals, however, these variables will remain un-transformed for ease of interpreting the coefficients of the linear regression model.

Outliers

Next, the data set needed to be examined for outliers. From the results below it can be seen that there are no collective outliers present, another important assumption for fitting a linear regression model.

```
sum(is.na(crime))

## [1] 0
```

Next, each variable was examined for point outliers. Significant outliers can worsen the fit of a linear regression model, and therefore any predictor variables with point outliers would be examined when fitting a model. To identify possible point outliers, the Grubbs Test was used to find both high and low outliers for each variable in the *uscrime* data set.

```
outlier.df = data.frame()
for(i in 1:16){
  outlier.df[i,'col'] = colnames(crime[i])
  outlier.df[i,'high outlier p-value'] = round(grubbs.test(x = crime[,i],
                                                  type = 10,
                                                  opposite = F,
                                                  two.sided = F)$p.value,4)
  outlier.df[i,'low outlier p-value'] = round(grubbs.test(x = crime[,i],
                                                  type = 10,
                                                  opposite = T,
                                                  two.sided = F)$p.value,4)
}
outlier.df

##        col high outlier p-value low outlier p-value
## 1        M               0.0303              1.0000
## 2       So               1.0000              1.0000
## 3       Ed               1.0000              1.0000
## 4      Po1               0.1083              1.0000
## 5      Po2               0.1009              1.0000
## 6       LF               0.9608              1.0000
## 7      M.F               0.0405              1.0000
## 8      Pop               0.0051              1.0000
## 9       NW               0.0223              1.0000
## 10      U1               0.1784              1.0000
## 11      U2               0.0702              1.0000
## 12  Wealth               0.2643              1.0000
## 13    Ineq               0.8519              1.0000
## 14    Prob               0.0166              1.0000
## 15    Time               0.2682              0.9063
## 16   Crime               0.0789              1.0000
```

From this, it can be seen that the predictor variables M, M.F, Pop, NW, Prob all contained possible high point outliers, as their p-values for high outliers were below the threshold of

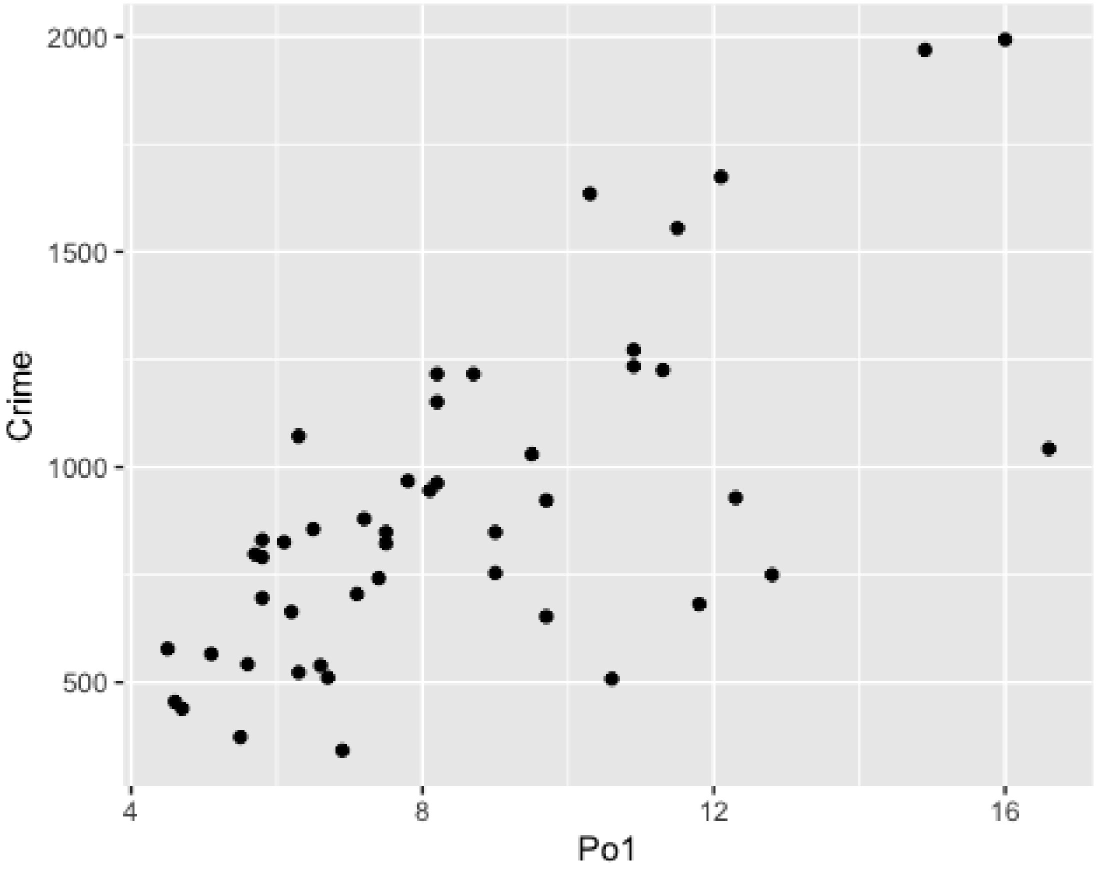
0.05. If any of these variables are selected for the final set of predictors, models with and without the outliers will be fit to examine the effect these point outliers may have.

Initial Plot

Next, an initial plot was created for the *uscrime* data set to examine if a linear relationship was present in the data. To do so, the variable with the strongest correlation with the response variable was selected as the predictor variable, since only one predictor and one response variable could be visualized.

```
corrs = c()
for(i in 1:15){
  corrs[i] = cor.test(crime[[i]],crime[[16]])$estimate[['cor']]
}
which.max(corrs)

## [1] 4
```

This variable turned out to be Po1. An initial plot of the response and a possible predictor variable for the *uscrime* data set can be seen below:



From this, a linear relationship between these two variables can be seen, which meets another important assumption for fitting a linear regression model.

Model Fitting

Initial Model

The first model fit contained all possible predictor variables in the data set. While this may not end up being the best model, it provides a good starting point to compare other models to.

```
lm.fit.1 = lm(formula = Crime ~ .,
            data = crime)
```

Next, the fit of this model was examined. In this process, the SSE, AIC, and BIC were all examined to gain the most information possible about the fit of the model. When comparing two models, if a discrepancy exists between these metrics, then the BIC will be used to select a model. This decision was made to provide the largest penalty for including additional parameters, mitigate overfitting, and attempt to find the true model among the set of potential models.

```
# SSE
lm.1.SSE = sum(lm.fit.1$residuals^2)
lm.1.SSE

## [1] 1354946

# AIC
lm.1.AIC = AICc(lm.fit.1)
lm.1.AIC

## [1] 671.1325
## attr(,"nall")
## [1] 47

# BIC
lm.1.BIC = BIC(lm.fit.1)
lm.1.BIC

## [1] 681.4816

# Adj R Squared
lm.1.summary$adj.r.squared

## [1] 0.7078062
```

From this, it can be seen that the full model accounts for 70.78% of the variance in the data. Notably, the Adjusted R-Squared value will be used throughout this process to adjust for the inclusion of more parameters as well.

Next, the coefficients of the full model were examined to determine the relevance of each predictor variable.

```
##
## Call:
## lm(formula = Crime ~ ., data = crime)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -395.74  -98.09   -6.69  112.99  512.67
##
## Coefficients: (1 not defined because of singularities)
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -5.984e+03  1.628e+03  -3.675 0.000893 ***
## M            8.783e+01  4.171e+01   2.106 0.043443 *
## So          -3.803e+00  1.488e+02  -0.026 0.979765
## Ed           1.883e+02  6.209e+01   3.033 0.004861 **
## Po1          1.928e+02  1.061e+02   1.817 0.078892 .
## Po2         -1.094e+02  1.175e+02  -0.931 0.358830
## LF          -6.638e+02  1.470e+03  -0.452 0.654654
## M.F          1.741e+01  2.035e+01   0.855 0.398995
## Pop         -7.330e-01  1.290e+00  -0.568 0.573845
## NW           4.204e+00  6.481e+00   0.649 0.521279
## U1          -5.827e+03  4.210e+03  -1.384 0.176238
## U2           1.678e+02  8.234e+01   2.038 0.050161 .
## Wealth       9.617e-02  1.037e-01   0.928 0.360754
## Ineq         7.067e+01  2.272e+01   3.111 0.003983 **
## Prob        -4.855e+03  2.272e+03  -2.137 0.040627 *
## Time        -3.479e+00  7.165e+00  -0.486 0.630708

## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 209.1 on 31 degrees of freedom
## Multiple R-squared:  0.8031, Adjusted R-squared:  0.7078
## F-statistic: 8.429 on 15 and 31 DF,  p-value: 3.539e-07
```

From this, it can be see that the predictor variables Ineq, Ed, Prob, and M were the only ones with p-values below the threshold of 0.05. However, both U2 and Po1 has p-values close to this threshold and had relatively large coefficient values, showing that these predictor variables could still be important. These variables would be included in the next model, and their p-values would be reassessed after fitting it.

Second Model

The next model fit contained only the aftermentioned predictor variables, which can be seen below:

```
lm.fit.2 = lm(formula = Crime ~ Ineq + Ed + Prob + M + U2 + Po1,
              data = crime)
```

After fitting the model, the accuracy metrics for this model were compared to the full model.

From this, it can be see that the full model had a lower SSE.

```
# SSE
lm.2.SSE = sum(lm.fit.2$residuals^2)
lm.2.SSE

## [1] 1611057

lm.1.SSE

## [1] 1354946
```

However, the AIC of the reduced model was lower than that of the full model, and there was an incredibly small probability that the full model was more accurate than the reduced model.

```
# AIC
lm.2.AIC = AICc(lm.fit.2)
lm.2.AIC

## [1] 643.9556
## attr(,"nall")
## [1] 47

lm.1.AIC

## [1] 671.1325
## attr(,"nall")
## [1] 47

exp((lm.2.AIC - lm.1.AIC) / 2)

## [1] 1.254898e-06
## attr(,"nall")
## [1] 47
```

Similarly, the BIC of the reduced model was lower than that of the full model, and it was very likely that the reduced model was better than the full model.

```
# BIC
lm.2.BIC = BIC(lm.fit.2)
lm.2.BIC

## [1] 654.9673

lm.1.BIC
```

```
## [1] 681.4816

lm.1.BIC - lm.2.BIC

## [1] 26.51427
```

Finally, the reduced model accounted for more variation in the data than the full model did.

```
# Adj R Squared
lm.2.summary$adj.r.squared

## [1] 0.7307463

lm.1.summary$adj.r.squared

## [1] 0.7078062

##
## Call:
## lm(formula = Crime ~ Ineq + Ed + Prob + M + U2 + Po1, data = crime)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -470.68  -78.41  -19.68  133.12  556.23
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -5040.50     899.84  -5.602 1.72e-06 ***
## Ineq           67.65      13.94   4.855 1.88e-05 ***
## Ed            196.47      44.75   4.390 8.07e-05 ***
## Prob        -3801.84    1528.10  -2.488  0.01711 *
## M             105.02      33.30   3.154  0.00305 **
## U2             89.37      40.91   2.185  0.03483 *
## Po1           115.02      13.75   8.363 2.56e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 200.7 on 40 degrees of freedom
## Multiple R-squared:  0.7659, Adjusted R-squared:  0.7307
## F-statistic: 21.81 on 6 and 40 DF,  p-value: 3.418e-11
```

Additionally, it can be seen that in the reduced model all of the predictors selected were relevant as they had p-values below the threshold of 0.05.

Given that all the accuracy metrics other than the SSE favored the reduced model and that all predictors selecter were relevant, analysis would continue with this reduced model.

Better Model: Outlier Examination

Since both predictors variables M and Prob were selected for the reduced model and they contained possible point outliers, the effect of these outliers on the model needed to be examined. To do so, the observations containing each of these outliers were removed from the data set.

It can be seen below that after removing the outliers from the data set, no more high point outliers were present in either of these predcitor variables.

```
##
##  Grubbs test for one outlier
##
## data:  clean.crime$M
## G = 2.46701, U = 0.85853, p-value = 0.2449
## alternative hypothesis: highest value 16.6 is an outlier

##
##  Grubbs test for one outlier
##
## data:  clean.crime$Prob
## G = 2.15214, U = 0.89234, p-value = 0.625
## alternative hypothesis: highest value 0.089502 is an outlier
```

Next, the model was refit using the same predcitor variables as before, but without the outliers in M or Prob. Afterwards, the accuracy metric for this model were compared to the model with the outliers present.

```
lm.fit.3 = lm(formula = Crime ~ Ineq + Ed + Prob + M + U2 + Po1,
              data = clean.crime)
```

From this, is can be seen that the SSE of the model without outliers was lower than that of the model containing outliers.

```
# SSE
lm.3.SSE = sum(lm.fit.3$residuals^2)
lm.3.SSE

## [1] 1534109

lm.2.SSE

## [1] 1611057
```

Similarly, the AIC of this model was lower, and there was an incredibly small probability that the model with outliers was more accurate.

```
# AIC
lm.3.AIC = AICc(lm.fit.3)
lm.3.AIC
```

```
## [1] 617.3604
## attr(,"nall")
## [1] 45

lm.2.AIC

## [1] 643.9556
## attr(,"nall")
## [1] 47

exp((lm.3.AIC - lm.2.AIC) / 2)

## [1] 1.678494e-06
## attr(,"nall")
## [1] 45
```

Finally, the BIC of this model was also lower and it was very likely that the model without outliers was better than the model with outliers.

```
# BIC
lm.3.BIC = BIC(lm.fit.3)
lm.3.BIC

## [1] 627.8137

lm.2.BIC

## [1] 654.9673

lm.2.BIC - lm.3.BIC

## [1] 27.15363
```

Additionally, the model without outliers account for more variation in the data.

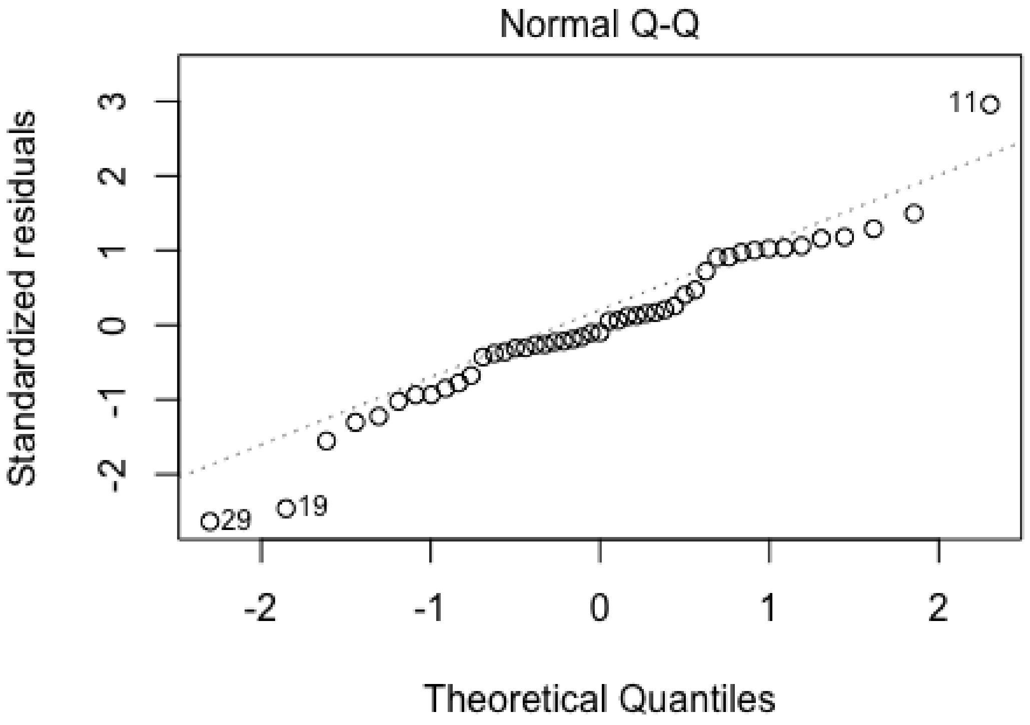
```
# Adj R Squared
lm.3.summary$adj.r.squared

## [1] 0.7416166

lm.2.summary$adj.r.squared

## [1] 0.7307463
```

Since all of the accuarcy metrics favored the model without outliers, this model was selected for further analysis.

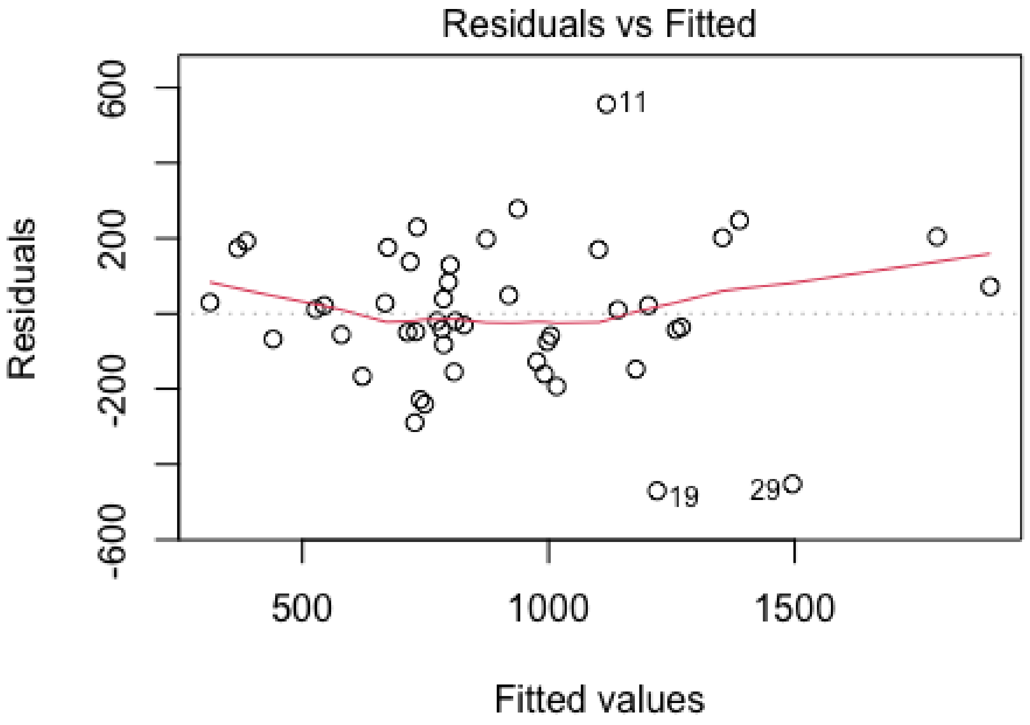
Best Model: Quality of Fit

After determining the best set of predictors for the model, the residual diagnostics needed to be examined to ensure the assumptions for linear regression were met.

Below, it can be seen that while perhaps long-tailed and with a few possible outliers, the residuals for this model mostly followed a normal distribution.

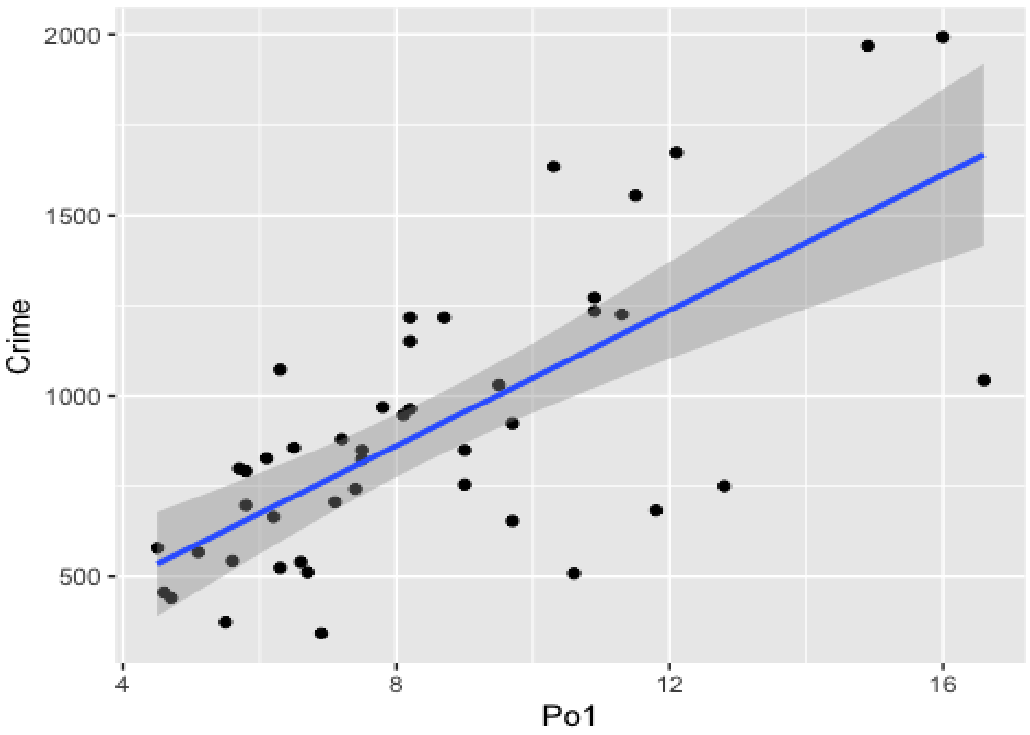


Additionally, it can be seen below that the residuals were fairly homogeneous with a few possible outliers present.



Since the assumptions appeared to be met from the residual diagnostics, predictions would take place using this model.

Below, the fit of the model can be seen on the predictor variable Po1 and the response varaible:



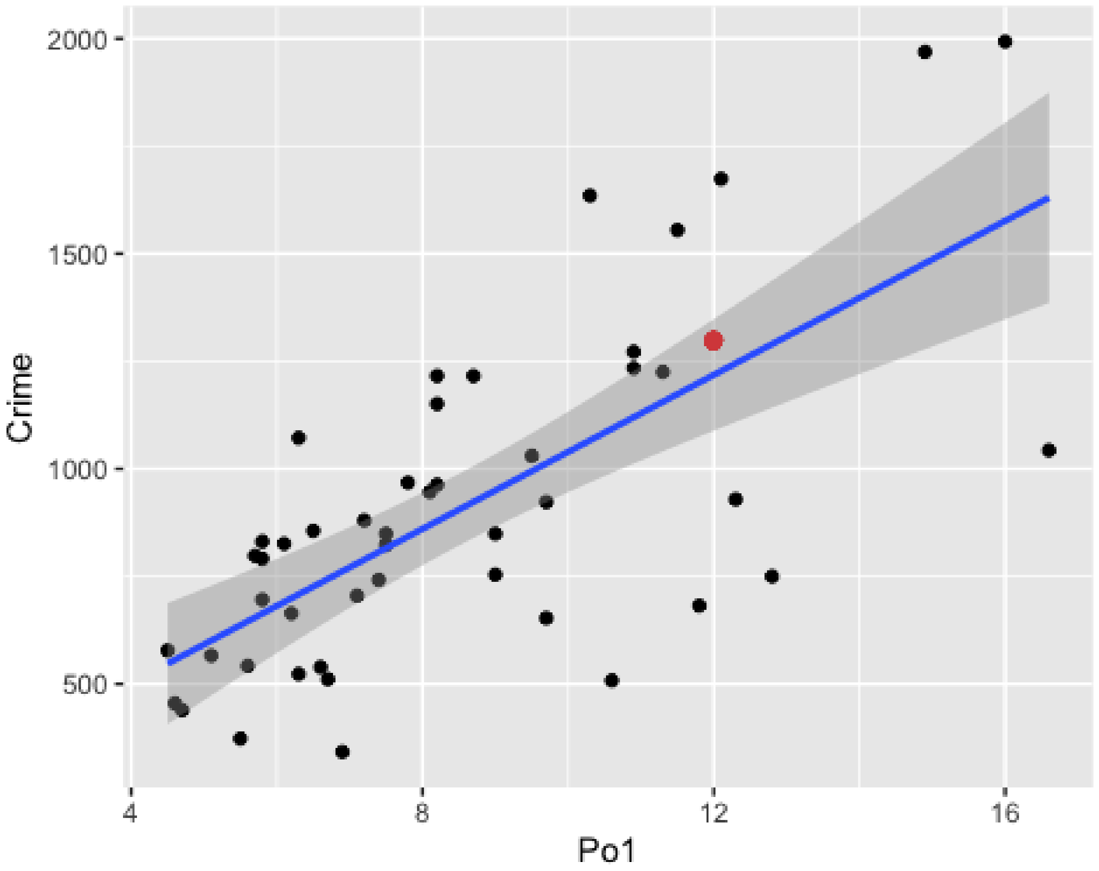
Prediction

After finding the best model and ensuring that the asusmption of linear regression were met, a new data point was predicted using a set of new values for the predictors.

```
## 1297.503
```

It can be seen that with this new set of predictor values, the model predicted a value of 1297.503 (rounded up to 1298) for the response variable.

This predicted data point can be seen in context with the other data points in the plot below:



From this, it can be seen that the predicted data point fits fairly well with the body of the rest of the data.

Best Model: Conclusions

Below, the values of the coefficients for this model can be seen:

```
##
## Call:
## lm(formula = Crime ~ Ineq + Ed + Prob + M + U2 + Po1, data = clean.crime)
##
## Coefficients:
## (Intercept)         Ineq           Ed         Prob            M
U2
##    -5227.79        70.30       200.25     -5520.41       120.57        92.
39
```

```
##         Po1
##      109.17

##
## Call:
## lm(formula = Crime ~ Ineq + Ed + Prob + M + U2 + Po1, data = clean.crime)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -467.62  -92.55   -3.83  110.23  554.15
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -5227.79     911.44  -5.736 1.31e-06 ***
## Ineq           70.30      14.44   4.868 2.00e-05 ***
## Ed            200.25      45.86   4.366 9.39e-05 ***
## Prob        -5520.41    2113.34  -2.612  0.01281 *
## M             120.57      36.44   3.309  0.00206 **
## U2             92.39      41.18   2.244  0.03075 *
## Po1           109.17      15.17   7.195 1.34e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 200.9 on 38 degrees of freedom
## Multiple R-squared:  0.7769, Adjusted R-squared:  0.7416
## F-statistic: 22.05 on 6 and 38 DF,  p-value: 5.479e-11
```

From this, it can be seen notably that each predictor value selected was relevant. Additionally, the effect of each of the predictor variables changing by 1 on the response variable can be seen in the Estimate column (the estimate of the coefficient values).

Affirming Conlcusions: Model Combinations

I wanted to affirm that I had found the best combination of predictor variables using the approach above by following a slightly different method. To do so, I created all possible combinations of the predictor variables from the *uscrime* data set, fit a model for each set of predcitors, and then returned the model with the lowest BIC (to start, I assigned the first set of predcitors as the “best model”). The results can be seen below:

```
best.lm.fit = lm(formulas[1],crime)
best.BIC = BIC(best.lm.fit)
best.param = formulas[1]
for(i in 2:length(formulas)){
  lm.fit = lm(formulas[i],crime)
  lm.BIC = BIC(lm.fit)
  if(lm.BIC < best.BIC){
    best.lm.fit = lm.fit
    best.BIC = lm.BIC
    best.param = formulas[i]
  }
}
best.param

## [1] "Crime ~ M+Ed+Po1+U2+Ineq+Prob"

lm.fit.3$call$formula

## Crime ~ Ineq + Ed + Prob + M + U2 + Po1
```

The best set of predictors found was the same that I had found using a more supervised method!

Next, I used the same approach by selected a model using the AIC and then the SSE.

```
best.lm.fit = lm(formulas[1],crime)
best.AIC = BIC(best.lm.fit)
best.param = formulas[1]
for(i in 2:length(formulas)){
  lm.fit = lm(formulas[i],crime)
  lm.AIC = AIC(lm.fit)
  if(lm.AIC < best.AIC){
    best.lm.fit = lm.fit
    best.AIC = lm.AIC
    best.param = formulas[i]
  }
}
best.param

## [1] "Crime ~ M+Ed+Po1+M.F+U1+U2+Ineq+Prob"

lm.fit.3$call$formula

## Crime ~ Ineq + Ed + Prob + M + U2 + Po1

best.lm.fit = lm(formulas[1],crime)
best.SSE = sum(best.lm.fit$residuals^2)
best.param = formulas[1]
for(i in 2:length(formulas)){
  lm.fit = lm(formulas[i],crime)
  lm.SSE = sum(lm.fit$residuals^2)
  if(lm.SSE < best.SSE){
    best.lm.fit = lm.fit
    best.SSE = lm.SSE
    best.param = formulas[i]
  }
}
best.param

## [1] "Crime ~ M+So+Ed+Po1+Po2+LF+M.F+Pop+NW+U1+U2+Wealth+Ineq+Prob+Time"

lm.fit.3$call$formula
```

```
## Crime ~ Ineq + Ed + Prob + M + U2 + Po1
```

As it can be seen above, the approach using AIC selected more parameters than the method using the BIC, and the SSE method selected even more than the AIC method. This was to be expected, as the AIC penalizes the inclusion of more parameters less than the BIC, and the SSE does not penalize at all. Notably, this shows how using different accuracy metrics can result in the selection of vastly different models. When deciding which method to use, the goal of fitting a linear regression model must be considered to choose an accuracy metric which best meets those goals.