Linear Models Final Project Report

Ryan Maciej

May 2022

Introduction

The purpose of this report is to determine how well a selected number of demographic statistics help identitfy criminal activity in a given population. The specific crime we will be evaluating in this report is the number of larcenies per 100k people (larcPerPop) which will be our response variable. Below we have listed out the statistics that we will be using as predictive variables along with their definitions.

Predictor1: *PctUnemployed*
percentage of people 16 and over, in the labor force, and unemployed (numeric - decimal)

Predictor2: *PctKids2Par*
percentage of kids in family housing with two parents (numeric - decimal)

Predictor3: *MalePctNevMarr*
percentage of males who have never married (numeric - decimal)

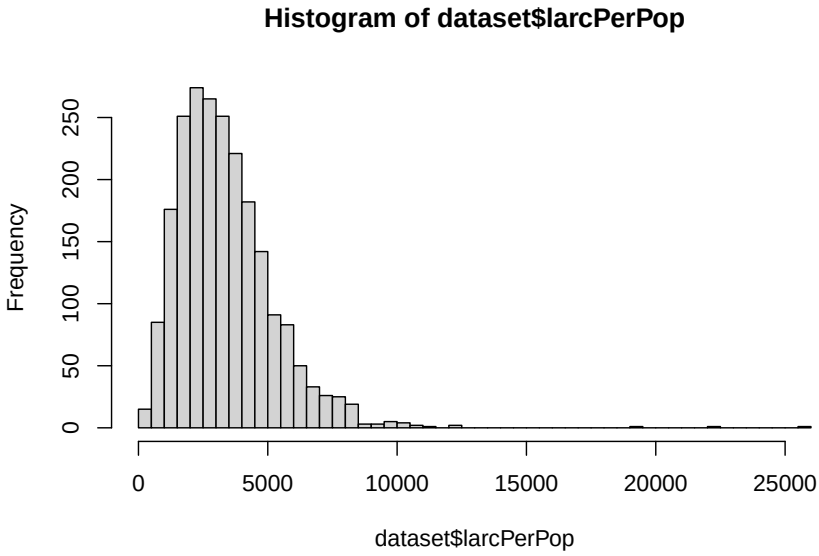
Predictor4: *pctWPubAsst*
percentage of households with public assistance income in 1989 (numeric - decimal)

We will be using linear regression techniques to build an appropriate model that uses these four predictor variables to better understand how they together can predict larcenies. Throughout this report we will discuss and demonstrate the techniques used for building, tuning, refining and evaluating the model as well as other aspects of the process when appropriate.

Step 1

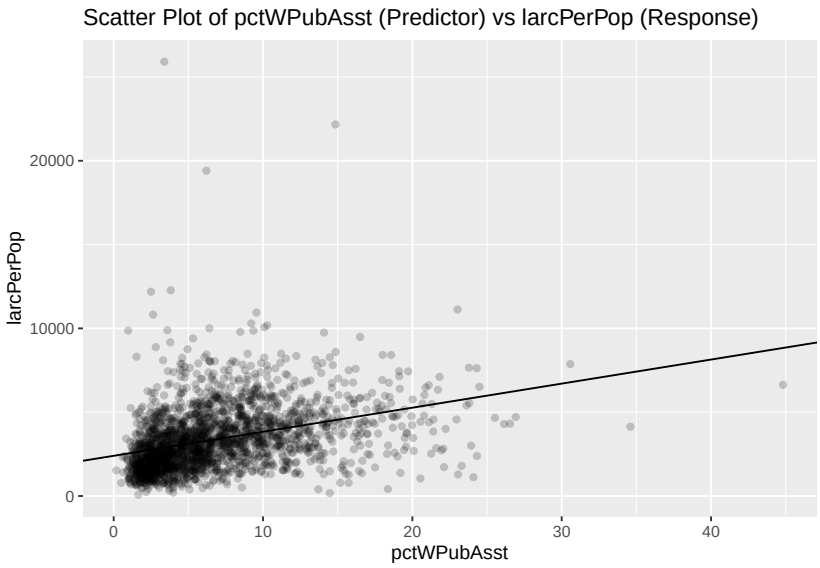
Exploratory Data Analysis

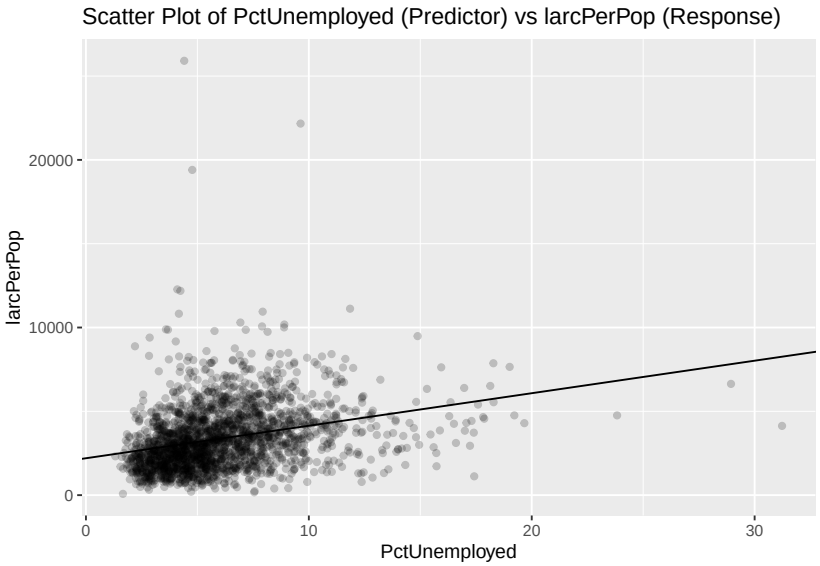
The first item we will look will be the distribution of larcenies with a histogram. This will give us an idea of how the data looks and can tell at a glance if there may be any immediate issues we want to address.

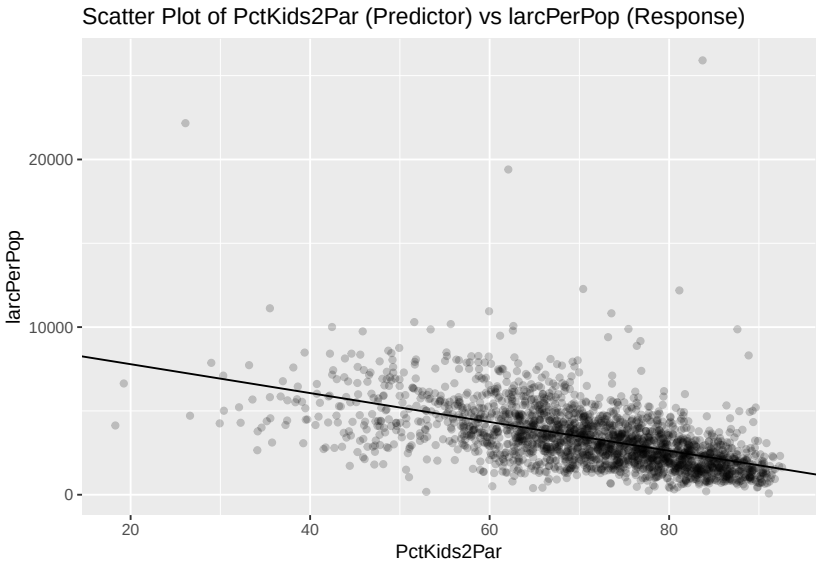


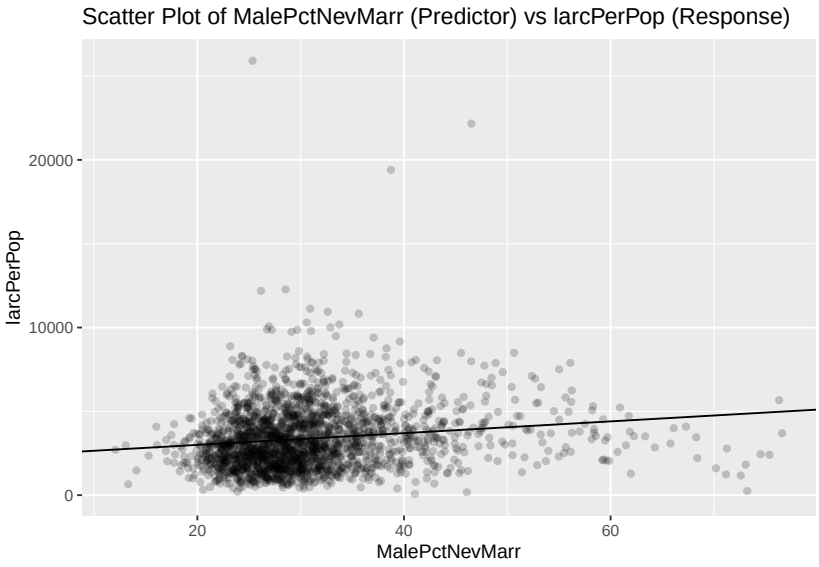
We can see that there are a few observations which have much higher values than the rest of the distribution. We will address these points later when we look, more formally, at outliers in our data.

To get some initial information about each of our four predictor variables and how they might be associated with the number of larcenies, we build scatter-plots to visualize the relationship between each of the predictors (each placed on x-axis) and the response (placed on y-axis). We've also included a basic linear regression line on each to show the relationships.




Scatter Plot of PctUnemployed (Predictor) vs larcPerPop (Response)
larcPerPop
20000
10000
0
0
10
20
30
PctUnemployed


Scatter Plot of PctKids2Par (Predictor) vs larcPerPop (Response)
larcPerPop
20000
10000
0
20
40
60
80
PctKids2Par



We can see each of the four predictor variables have decent linear relationships to the response, three of which have positive correlations and one with a negative correlation.

Step 2

Fitting the Linear Model

In this step, we fit the initial linear regression model using all of the 4 predictor variables as well as the response to produce a linear regression model.

The form of the regression equation is:

$y = \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_0 + \epsilon$

where y is larcenies, each β_1, β_2, β_3, β_4 is the estimated slope for each predictor, β_0 is an estimated intercept and ϵ is an error term.

```
##                   Estimate      Pr(>|t|)
## (Intercept)    12670.55128 5.773422e-146
## PctUnemployed    -61.83867  5.490698e-03
## PctKids2Par     -113.75353 7.850089e-118
## MalePctNevMarr   -17.27995  1.168341e-04
## pctWPubAsst      -42.77284  6.550019e-03
```

Above we can see the summary of the linear regression fit with this data, including its slope/ intercept in the 'Estimate' column as well as the p-values for each in the 'Pr(>|t|)' column. This shows that each of our predictor variables does have a significant p-value and will likely be kept in the regression.

We also have an R^2 value of 0.3167

To look at which variables we will keep in our model, we will move on to the next section to look at this more formally with a couple of methods.

Step 3

Model Selection

We want to be sure that we only keep predictor variable in our model which contribute to the predictive process in a meaningful way. To ensure we have the simplest, most powerful model we can, we may way to eliminate some predictor variables if they do not help in the regression. To determine which of the predictor variables should be included in the model we looked at two different method of variable selection using the functions in R: **fastbw()** and **stepAIC()** . Both are iterative processes where variables are removed from the model, and evaluated for improvement. The first (**fastbw()**) looks at p-values while the second (**stepAIC()**) calculates AIC values.

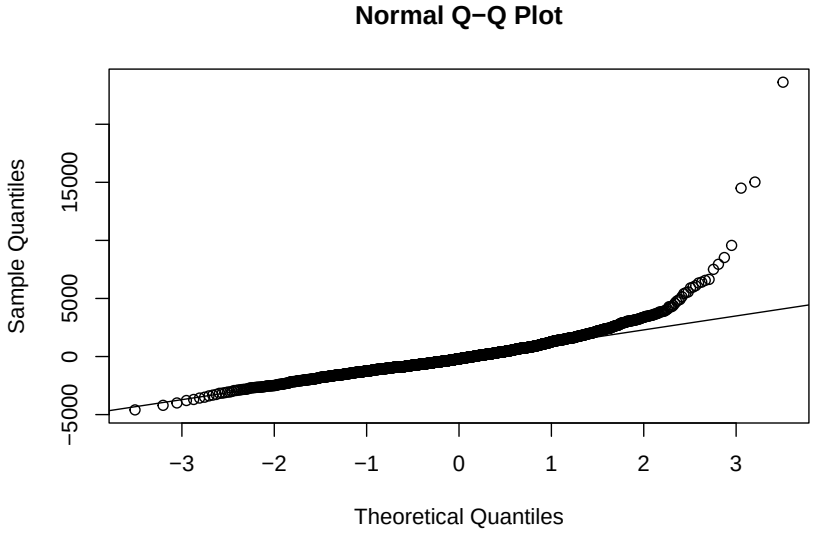
Both of these methods indicated that all 4 predictor variables were contributing positively to the regression analysis and should be kept in the model.

Step 4

Model Diagnostics

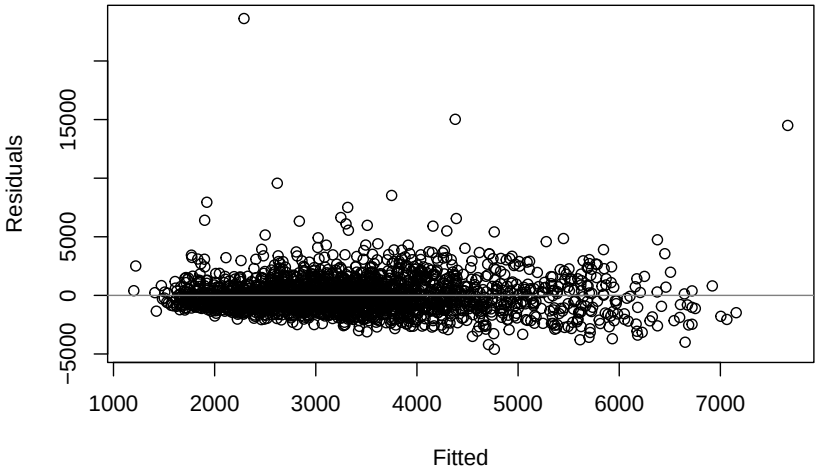
An important aspect of building linear regression models is to ensure that certain assumptions of the model are met so that the statistics we look at in relation to the model can be relied on with confidence. The model assumptions we need to account for are that the error terms are uncorrelated, normally distributed, have a mean of zero, and constant variance as well as that there is a linear association.

First we will look at the assumption that the errors are normally distributed. To check this, we can use the QQnorm() function in R. If the plot looks to be a straight line, or close to, we can say that the model assumption is upheld.

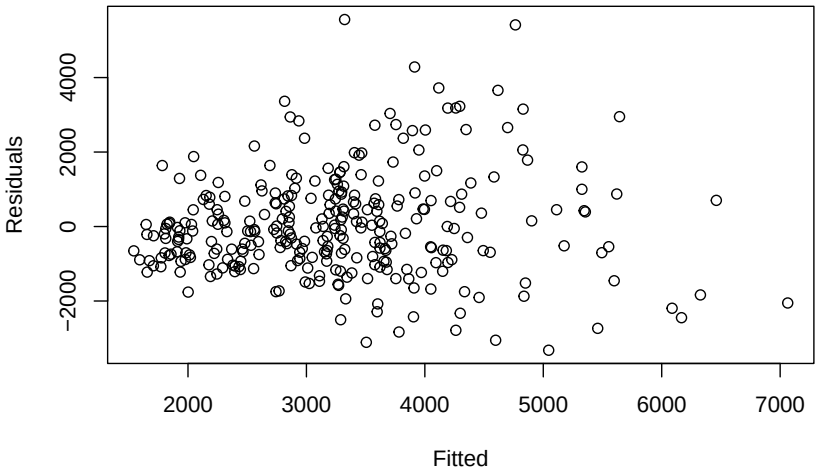


From the plot above we can see that there is a significant curve on the tail end. While this is a subjective interpretation, we could say that the model assumuption of normally distributed errors is not met. Later we will try to correct this.

The second assumption we will look at is that of constant error variance. To evaluate this we can use the a plot that looks at our fitted values vs the residuals. If we happen to see something like a cone shape, we could say that the model assumption is not upheld.



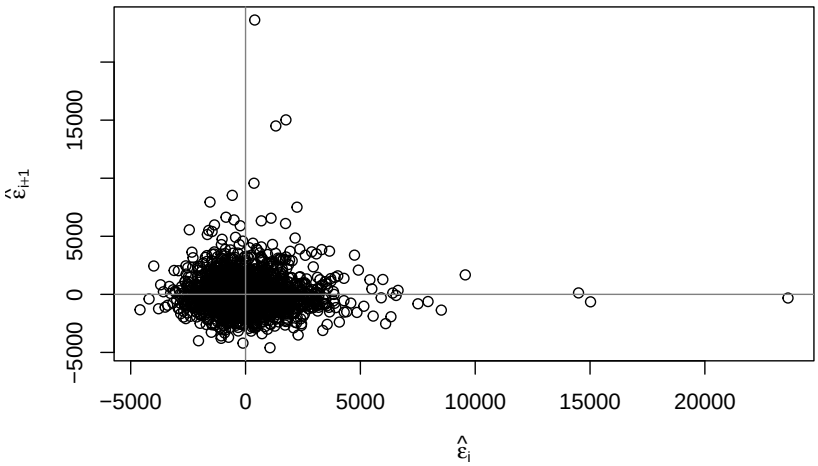
We can see the bit of a cone shape tip off to the left hand side of the plot. To visualize this easier without so many points, we can look at the same plot with only 300 points:



This better shows the data points forming more of a point or cone shape on the left hand side indicating this model assumption is also not upheld and we will try to correct for that later aswell.

Finally we can look at the model assumption that the errors are uncorrelated.

To evaluate this model assumption, we can use a plot that looks at the relationship of the errors to the previous error, or more specifically called the Lagged Residual Plot. Here we would expect the cluster of data points to be rather random and centered around (0,0)



We can see that the points seem to be centered around (0,0) and are relatively random around that, therefore we can say this model assumption, of uncorrelated errors is upheld.

Step 5

Assessing Possible Outliers

The existence of outliers can have an affect on our model and we want to determine how significant individual data points have on our model as well as whether or not it would be warranted to remove them. We will look at two different values to determine if a point might be considered an outlier: the Cooks Distance (cooks), and the studentized residuals (rstan).

```
##    larcPerPop PctUnemployed PctKids2Par MalePctNevMarr pctWPubAsst
## 1    25910.55          4.41       83.73          25.35        3.39
## 2    19401.00          4.77       62.08          38.73        6.21
## 3    22164.78          9.63       26.11          46.53       14.85
## 4    12187.98          4.24       81.14          26.16        2.50
## 5    12274.59          4.10       70.42          28.54        3.82
## 6     9868.63          3.68       87.61          29.67        0.97
## 7    10822.35          4.17       73.57          35.62        2.64
## 8     9888.57          3.59       75.47          26.73        3.60
## 9    10945.14          7.93       59.94          32.62        9.56
## 10    8310.36          2.83       88.86          24.34        1.53
##              communityname state population   cooks   rstan
## 1       EastLongmeadowtown    MA      13367 0.04506 15.0228
## 2             Tukwilacity    WA      11874 0.03275  9.5581
## 3        AtlanticCitycity    NJ      37986 0.15175  9.2577
## 4         HarperWoodscity    MI      14903 0.00696  6.0856
## 5         MyrtleBeachcity    SC      24848 0.00650  5.4221
## 6           Paramusborough    NJ      25067 0.00773  5.0549
## 7     RichmondHeightscity    MO      10448 0.00579  4.7752
## 8           Springdalecity    OH      10621 0.00322  4.2235
## 9       SouthSaltLakecity    UT      10129 0.00315  4.1695
## 10 LaCanadaFlintridgecity    CA      19378 0.00488  4.0772
```

Included in the table above are the top 10 most likely outliers in the data according the the studenized residuals since all of the cooks values came back far below the threshold of .8 to be considered outliers.

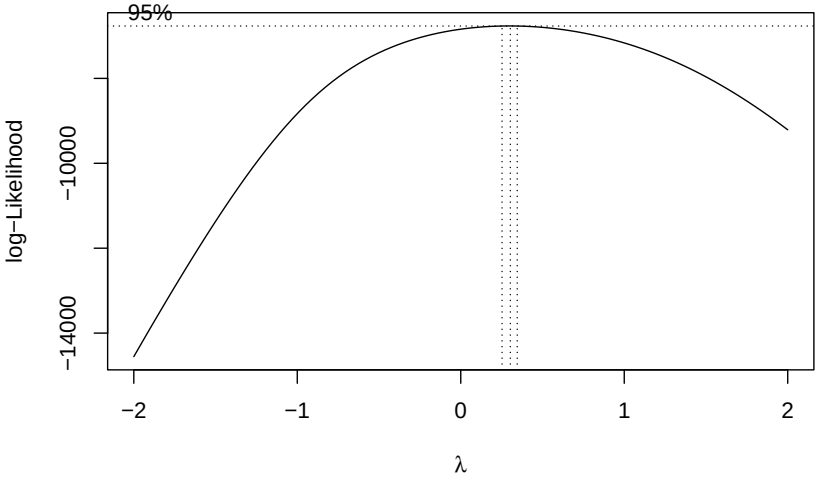
While some values do have a large value for their studenized residual, we do not know of any obvious justification to remove them from the model and will leave all data as it.

Step 6

Model Transformation

As mentioned earlier, adjustments are needed to be made to the model to ensure certain model assumptions are met. Earlier we discussed how two of our model assumptions were not met: normality of errors and constant error variance. To correct these issues, we first used the Box-Cox method to determine if we could make an adjustment to our response variable.

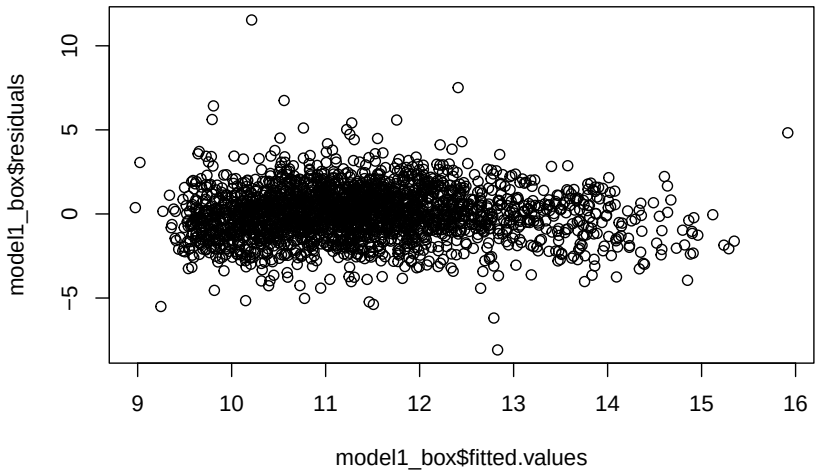
To visualize the possibility of this transformation we can look at the plot below, which gives us a confidence interval in which our transformation might optimally be. Given that it does not include zero or one in that interval, we will use the transformation of our response raised to the power of λ, which will be .3 in this case.

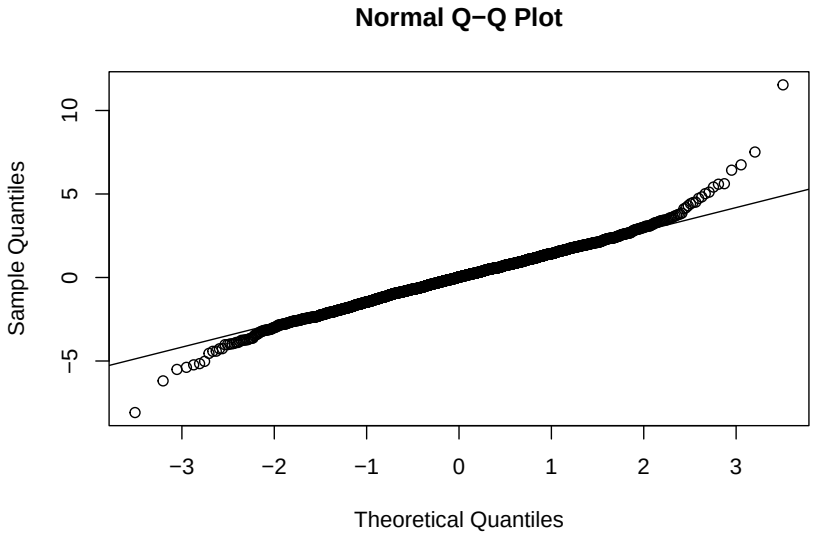


When we build the new model with the transformation of y^λ we will want to again look at our diagnostic plots to determine if we have succesfully corrected any of the assumptions that were not previously updheld.

The equation with the transformation would be:

$y^\lambda = \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_0 + \epsilon$





When looking at both our QQ Plot & Fitted vs Residual Plots to evaluate Normality of Errors and Constant Error Variance respectively, we can see that they look improved significantly and we'd be able to more confidently say that the model assumptions are now upheld.

Lastly we looked at weather including polynomial terms in our regression would improve it at all.

We found that including squared terms of both pctWPubAsst & PctKids2Par in our model improved the regression.

The final model equation being:

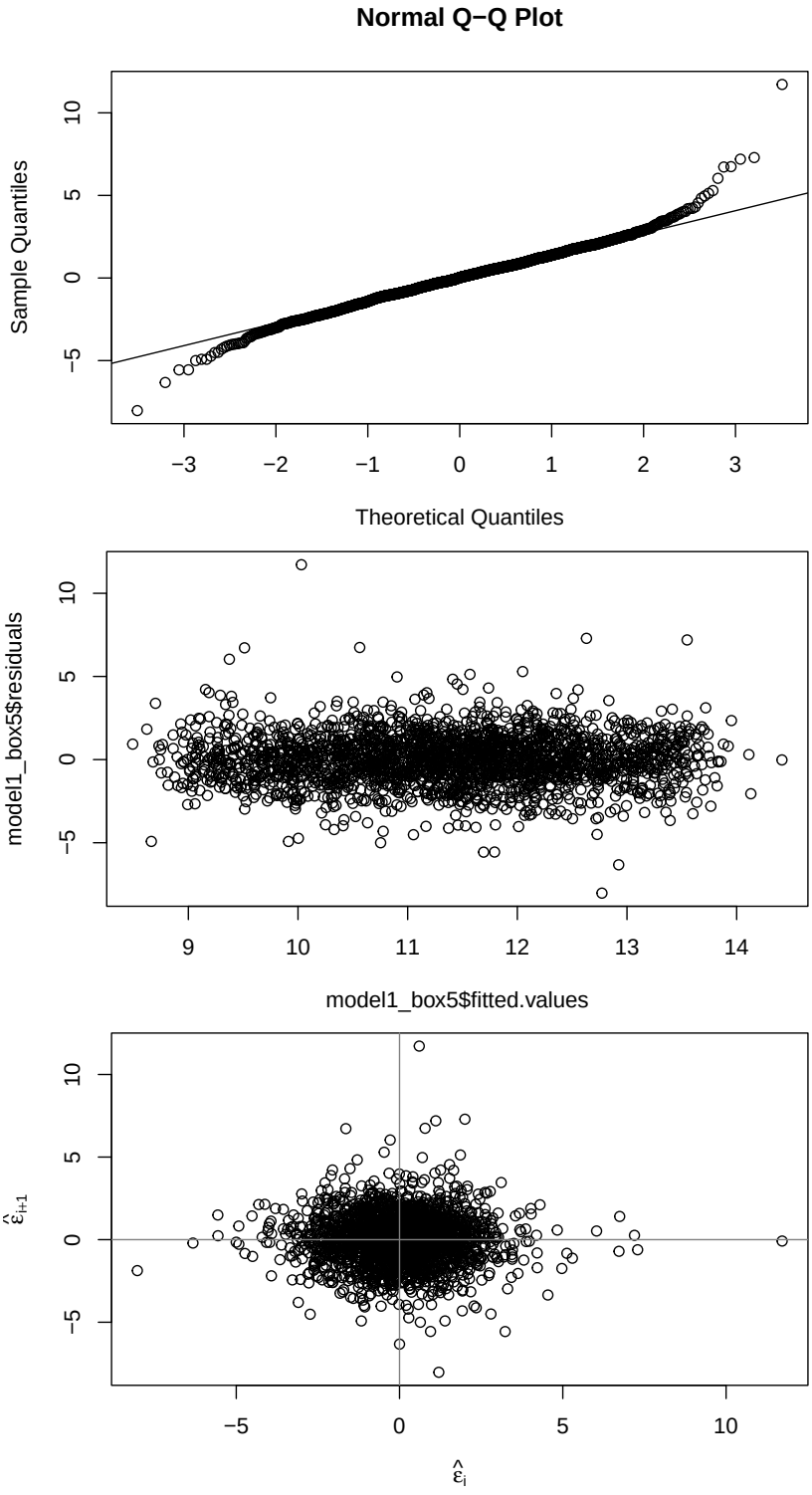
$larcPerPop^{.303} = PctUnemployed * x_1 + PctKids2Par * x_2 + MalePctNevMarr * x_3 + pctWPubAsst * x_4 + pctWPubAsst^2 * x_5 + PctKids2Par^2 * x_6 + \beta_0$

The estimated slopes for the model along with their p-value are:

```
##                       Estimate     Pr(>|t|)
## (Intercept)      14.728065563 1.855065e-60
## PctUnemployed    -0.068133964 1.557596e-03
## PctKids2Par       0.094482441 2.203318e-04
## MalePctNevMarr   -0.019699230 5.247356e-06
## pctWPubAsst      -0.111913128 4.005406e-05
## I(pctWPubAsst^2)  0.003188786 1.411644e-03
## I(PctKids2Par^2) -0.001635481 4.680269e-17
```

The R^2 value is: 0.3776 which is an improvement over our initial R^2 of 0.3167

Again we can verify our model assumptions are met with the QQPlot, Fitted vs Residual Plot and Lagged Residual Plots:



Each of the three diagnostic plots of our final model indicate that our assumptions have been met and we can declare victory over linear regression!

Conclusion

Over the course of this report we have discussed and demonstrated our approach to building a linear regression model to help understand what some of the underlying predictors of larcenies within a population might be. We looked at four different measurements of a population to determine if they were suitable to assist in making a meaningful assessment. We found each of these four variables to be useful and kept all of them in our model. We also included two additional variations of those variables, or more specifically polynomial terms, to help improve our model. We have listed below each of the variables that were used in the model along with there associated slope estimates and p-values.

```
##                       Estimate     Pr(>|t|)
## (Intercept)      14.728065563 1.855065e-60
## PctUnemployed    -0.068133964 1.557596e-03
## PctKids2Par       0.094482441 2.203318e-04
## MalePctNevMarr   -0.019699230 5.247356e-06
## pctWPubAsst      -0.111913128 4.005406e-05
## I(pctWPubAsst^2)  0.003188786 1.411644e-03
## I(PctKids2Par^2) -0.001635481 4.680269e-17
```

The R^2 and Adjusted R^2 for this model are 0.3776 & 0.3759 respectively.

That concludes our report on building a linear regression model to help assess larcenies in a population. Thank you for your time and consideration.