Linear Models Final Project Report

Introduction

We are interested in determining the effects of demographic variables on crime. In this report, we will be evaluating the violent crimes using four predictor variables.

These four predictor variables are:

1. LemasSwFTFieldPerPop (the number of sworn full time police officers in field operations on the street per 100K population)
2. PctUnemployed (the percentage of people 16 and over, in the labor force, and unemployed)
3. PctLess9thGrade (the percentage of people 25 and over with less than a 9th grade education)
4. PctPopUnderPov (the percentage of people under the poverty level)

Response Variable: ViolentCrimesPerPop (the total number of violent crimes per 100K popuation)

The violent crimes variable was calculated using population and the sum of crime variables considered violent crimes in the United States: murder, rape, robbery, and assault.

The exploratory data analysis of these variables reveals that PctUnemployed,PctLess9thGrade and PctPop-UnderPov have a very strong linear relationship with the total number of violent crimes.

Exploratory Data Analysis

For our predictor variables we look at some basic statistics. The number of violent crimes ranged from 52 to 4026. Additionally, percentage of people with less education were from 1% to 40%, the percentage of people under the poverty level were 2% to 44% and the percentage of unemployed peopole were from 2% to 16%. The number of police officers in field variable shows a very weak relationship to our desired response variable (ViolentCrimesPerPop).

Below is a glimpse of the data set and the variables.

```
## Rows: 319
## Columns: 7
## $ communityname        <chr> "Wacocity", "PineBluffcity", "Glendalecity", "Arl~
## $ state                <chr> "TX", "AR", "CA", "TX", "NY", "TX", "TX", "CA", "~
## $ ViolentCrimesPerPop  <dbl> 1544.24, 1476.93, 374.07, 772.77, 2097.71, 689.42~
## $ LemasSwFTFieldPerPop <dbl> 173.33, 151.61, 105.36, 112.89, 307.34, 106.72, 1~
## $ PctUnemployed        <dbl> 8.39, 11.05, 6.95, 4.99, 8.98, 11.24, 7.71, 3.80,~
## $ PctLess9thGrade      <dbl> 13.01, 14.49, 11.54, 3.71, 14.10, 36.26, 14.84, 3~
## $ PctPopUnderPov       <dbl> 28.68, 27.71, 14.37, 8.21, 19.29, 37.29, 18.50, 6~
```

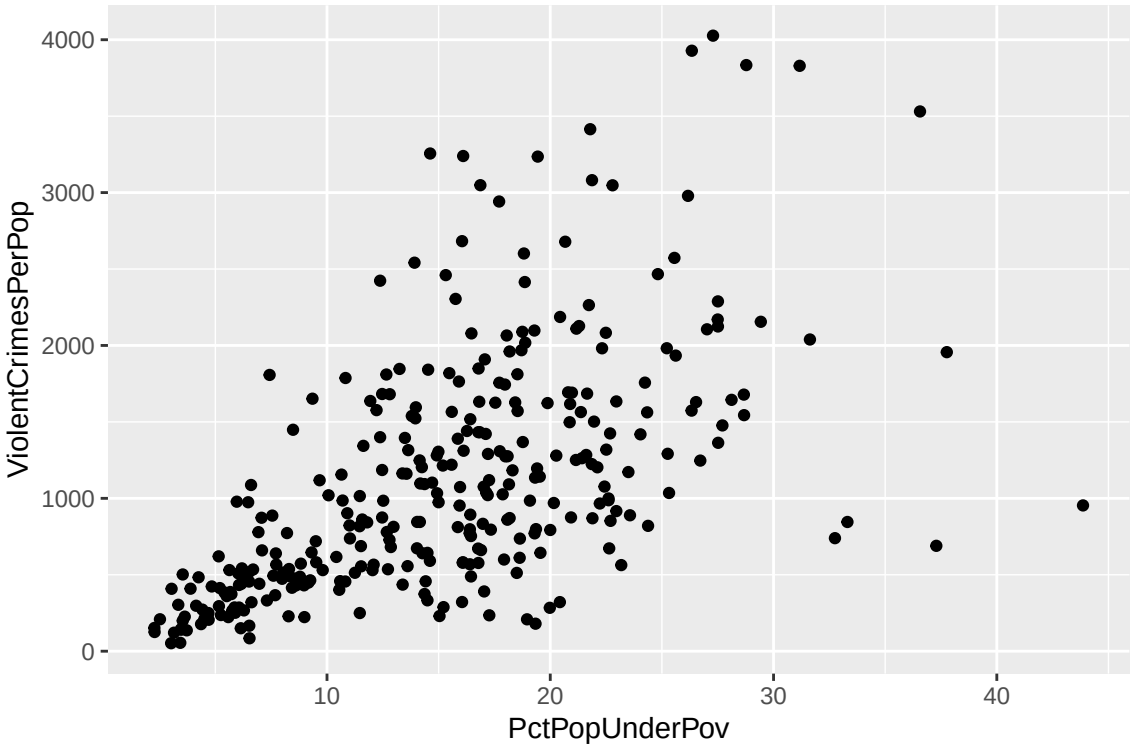
First we look at the summary statistics of each variable

```
## [1] "ViolentCrimesPerPop"
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   52.25  510.69  953.67 1114.83 1564.89 4026.59
```

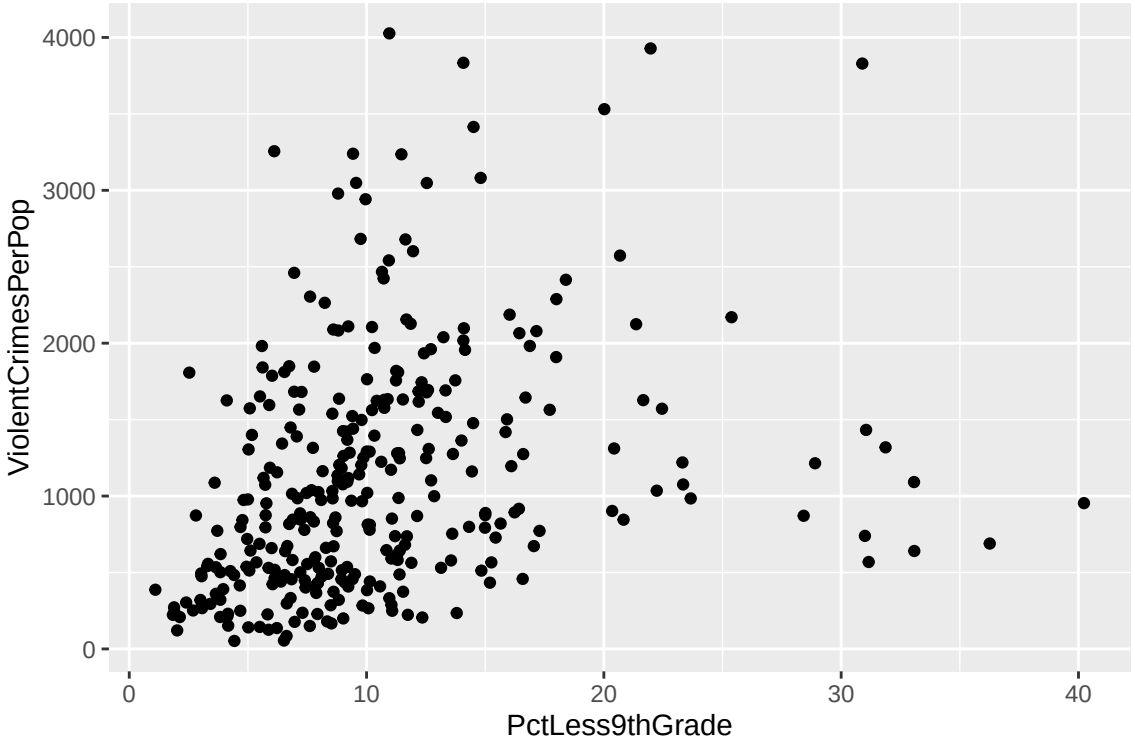
```
## [1] "LemasSwFTFieldPerPop"

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   19.21  130.43  170.16  211.32  226.81 3290.62

## [1] "PctLess9thGrade"

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1.10    6.65    9.23   10.47   12.38   40.23

## [1] "PctUnemployed"

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   2.110   5.305   6.580   7.061   8.610  16.600

## [1] "PctPopUnderPov"

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    2.27    9.08   15.32   15.15   19.48   43.86
```

Next, we look at the distribution of each predictor against the response to analyze the relationship between them:



There is a strong positive linear relationship between our predictor, PctPopUnderPov(percentage of people under poverty line) and the response, ViolentCrimesPerPop(total number of violent crimes per 100K population).

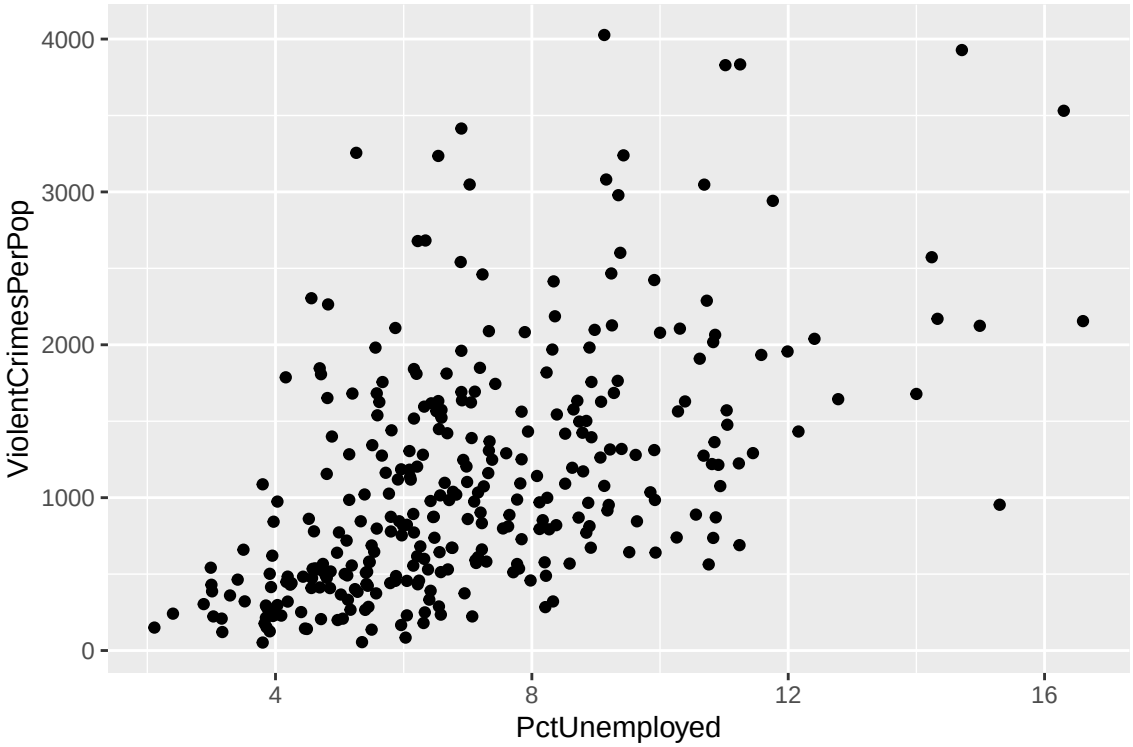
Next, we look at the relationship between percentage of people with education below 9th grade vs violent



crimes

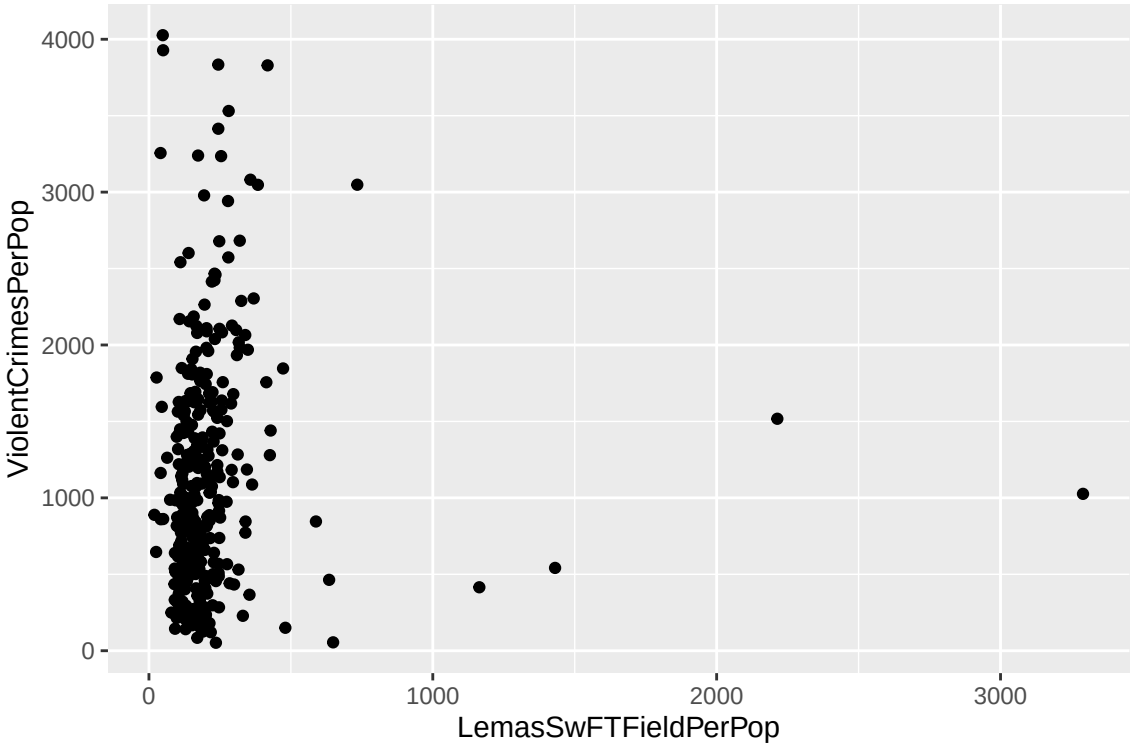
There is a positive linear correlation between PctLess9thGrade(percentage of people 25 and over with less than a 9th grade education) and the response, ViolentCrimesPerPop(total number of violent crimes per 100K population).

We look at the distribution of percentage of people unemployed vs violent crimes.



There is a positive linear correlation between PctUnemployed(percentage of people 16 and over, in the labor force, and unemployed) and the response, PctUnemployed(total number of violent crimes per 100K population).

We look at the distribution of sworn full time police officers in field vs violent crimes.



There is a very weak relationship between LemasSwFTFieldPerPop(sworn full time police officers in field operations (on the street as opposed to administrative etc) per 100K population) and the response, ViolentCrimesPerPop(total number of violent crimes per 100K population).

Next we explore the correlation values between all the predictors and response variable as shown below:

```
## [1] "Correlation between LemasSwFTFieldPerPop & ViolentCrimesPerPop"

## [1] 0.0659283

## [1] "Correlation between PctUnemployed & ViolentCrimesPerPop"

## [1] 0.5368987

## [1] "Correlation between PctLess9thGrade & ViolentCrimesPerPop"

## [1] 0.2982112

## [1] "Correlation between PctPopUnderPov & ViolentCrimesPerPop"

## [1] 0.5920328
```

Fitting the Linear Model

Next we fit a linear model with our response variable and the four predictors.

The linear regression equation is given as:

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

where y is the violent crimes per population $\beta_1, \beta_2, \beta_3, \beta_4$ are the slopes β_0 is the estimated intercept ϵ is the error

The summary of the model is shown below:

```
##
## Call:
## lm(formula = ViolentCrimesPerPop ~ LemasSwFTFieldPerPop + PctLess9thGrade +
##     PctPopUnderPov + PctUnemployed)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1661.6  -367.6  -102.5   213.5  2281.6
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)          -97.6777   110.2491  -0.886 0.376310
## LemasSwFTFieldPerPop   0.2326     0.1442   1.613 0.107653
## PctLess9thGrade      -20.8736     7.7676  -2.687 0.007587 **
## PctPopUnderPov        48.1211     7.6686   6.275 1.16e-09 ***
## PctUnemployed         92.4439    24.9675   3.703 0.000252 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 620.7 on 314 degrees of freedom
## Multiple R-squared:  0.383,  Adjusted R-squared:  0.3751
## F-statistic: 48.73 on 4 and 314 DF,  p-value: < 2.2e-16
```

As seen from the summary, LemasSwFTFieldPerPop pvalue is 0.108 which is greater than alpha of 0.05 and hence not statistically significant. All other predictors have p value less than 0.05 and are statistically significant. This seems accurate because from the scatter plot of ViolentCrimesPerPop and LemasSwFT-FieldPerPop, there is a weak relationship.

The R squared value is 0.38 which seems lower than expected since from the scatterplots atleast three out of four variables had good correlation with the response variable.

Model Selection

We want to keep meaningful predictors with predictve power. To eliminate unnecessary predictos we can use *fastbw()* or *stepAIC()*. Both are iterative processes. fastbw uses the p value but stepAIC uses the AIC value.

Next we perform fastbw() to select our variables and look at the summary below:

```
##
## No Factors Deleted
##
## Factors in Final Model
##
## [1] LemasSwFTFieldPerPop PctLess9thGrade      PctPopUnderPov
## [4] PctUnemployed
```

We see that this method does not delete any variables from our model.For futher confirmation we try the stepAIC method below:

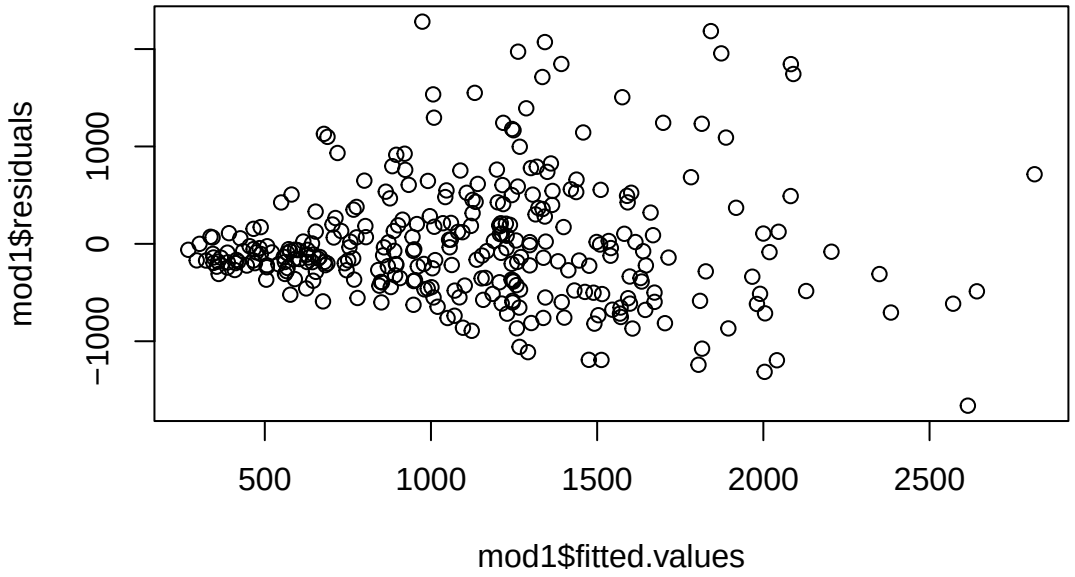
```
## Start:  AIC=4107.82
## ViolentCrimesPerPop ~ LemasSwFTFieldPerPop + PctLess9thGrade +
##     PctPopUnderPov + PctUnemployed
##
##                         Df Sum of Sq       RSS    AIC
## <none>                               120966702 4107.8
## - LemasSwFTFieldPerPop  1   1002864 121969566 4108.5
## - PctLess9thGrade       1   2782038 123748739 4113.1
## - PctUnemployed         1   5281295 126247997 4119.5
## - PctPopUnderPov        1  15169545 136136247 4143.5

## Stepwise Model Path
## Analysis of Deviance Table
##
## Initial Model:
## ViolentCrimesPerPop ~ LemasSwFTFieldPerPop + PctLess9thGrade +
##     PctPopUnderPov + PctUnemployed
##
## Final Model:
## ViolentCrimesPerPop ~ LemasSwFTFieldPerPop + PctLess9thGrade +
##     PctPopUnderPov + PctUnemployed
##
##
##   Step Df Deviance Resid. Df Resid. Dev      AIC
## 1                       314  120966702 4107.821
```

Again, this method also did not remove any variables from our model. This is surprising that both of the methods did not remove the variable LemasSwFTFieldPerPop as it was weakly related to our response variable.

Model Diagnostics

Next we verify our model assumptions which are: 1. Errors have constant variance 2. Residuals are normally distributed 3. Errors are uncorrelated and normally distributed

We perform diagnostics on the model to check constant variance of errors below:



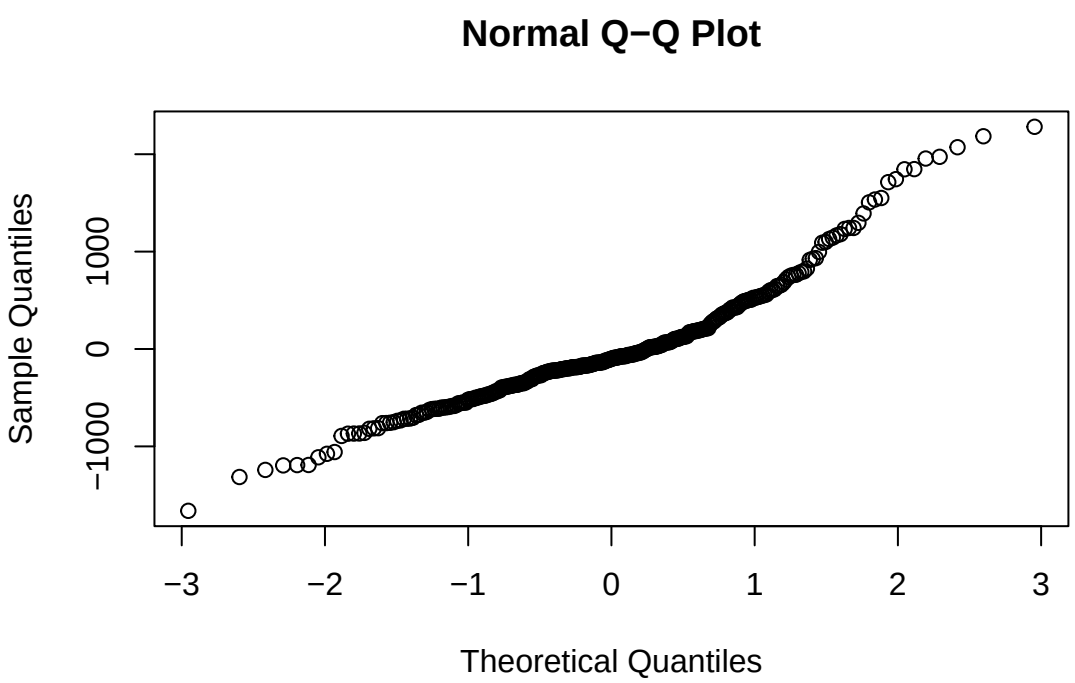
We see that the model assumption of errors having constant variance is not upheld. This is a cone shaped graph indicating *heteroscedasticity*.

We can also check for heteroscedasticity using the *Breush Pagan test*. The null hypothesis is that the errors have constant variance.

```
##
##  studentized Breusch-Pagan test
##
## data:  mod1
## BP = 38.386, df = 4, p-value = 9.326e-08
```

Since the p value is less than alpha significance of 0.05, we reject the null hypothesis and conclude that the model assumption of errors with constant variance is not upheld.

Now we look at a *QQ plot* to verify the normality of residuals below.



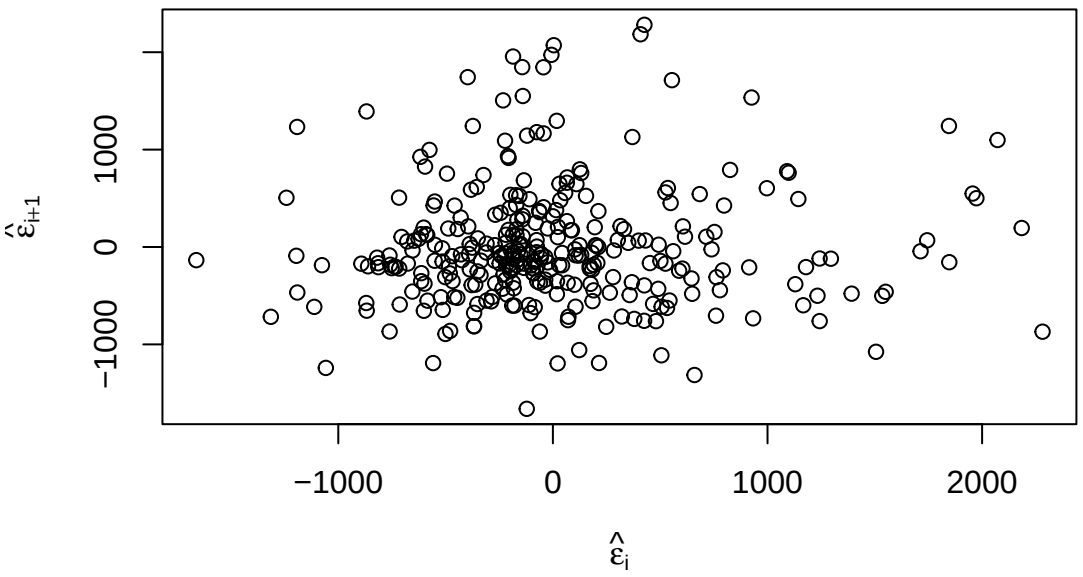
The residuals appear to be normally distributed but there is a slight departure from straight line in the center so we will need to perform further testing to confirm.

We can verify the normality of residuals using the *Shapiro Wilk test*. The null hypothesis states that the residuals are normally distributed.

```
##
##  Shapiro-Wilk normality test
##
## data:  mod1$residuals
## W = 0.93389, p-value = 1.038e-10
```

The p value of 1.04 e-10 is less than the alpha of 0.05, therefore we can reject the null hypothesis. However, tests can be sensitive when sample size is large so we can conclude from the QQ plot that residuals are normally distributed since we do not see a gross departure from the straight line.

We also check for correlation of errors by *plotting successive pairs of residuals.*



We see a random scatter of points suggesting uncorrelated errors.

We can also verify that errors are uncorrelated using the *Durbin Watson Test* below.

```
##
##  Durbin-Watson test
##
## data:  mod1
## DW = 1.9309, p-value = 0.269
## alternative hypothesis: true autocorrelation is greater than 0
```

The p value of 0.269 is greater than alpha of 0.05, therefore we fail to reject the null hypothesis that the errors are uncorrelated and the assumption of uncorrelated errors is upheld.

Verifying Outlier & Influential Observations

Next we will check for outliers and influential observations in our model. To look at the influential points we can calculate the *cooks distance* and then using the rule of thumb, if cooks distance is greater than the 50th percentile of F distribution and another rule of thumb that if cooks distance is greater than 1 we have influential points.

```
## named integer(0)
```

```
## named integer(0)
```

From above we do not find any influential observations.

We also check for outliers using *standardized residuals*.

```
## [1] 3.1922 3.6930 3.3592 3.0321 3.5503 3.2278
```

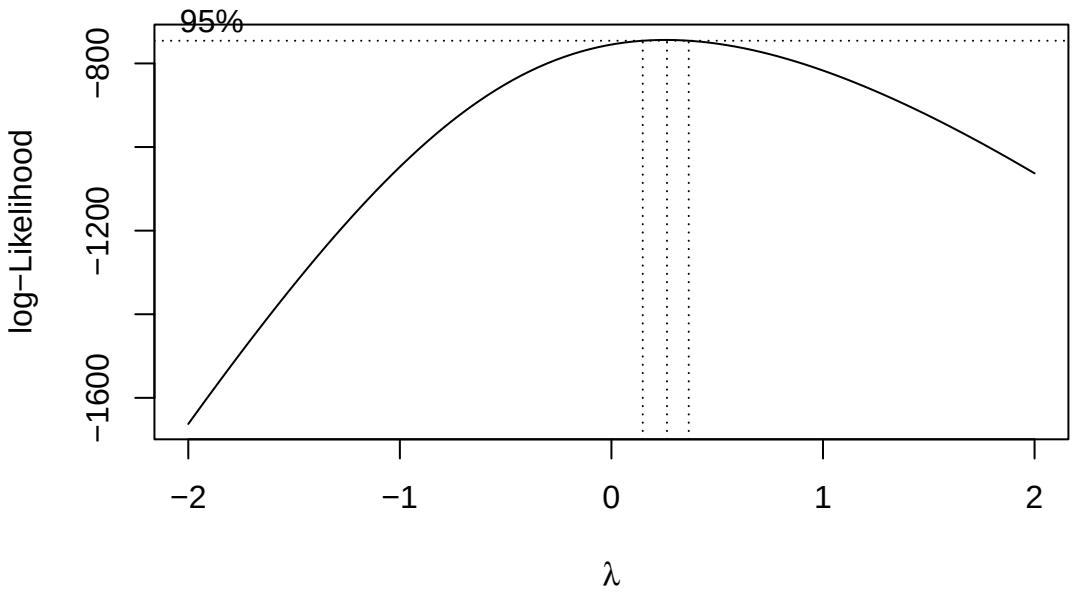
The rule of thumb states that an observation is considered an outlier if the absolute value of standardized residual is greater than 3. From the above, there are 6 outliers according to the standardized residual method. However they are not influential as we verified earlier using the cooks distance method. So even though we have a lot of outliers, they do not disproportionately affect the model.

Model Transformations

From above we know that the assumption of constant error variance was not upheld so we will try a few transformations to correct the heteroskedasticity first.

First we try the Box Cox transformation using all predictors below.

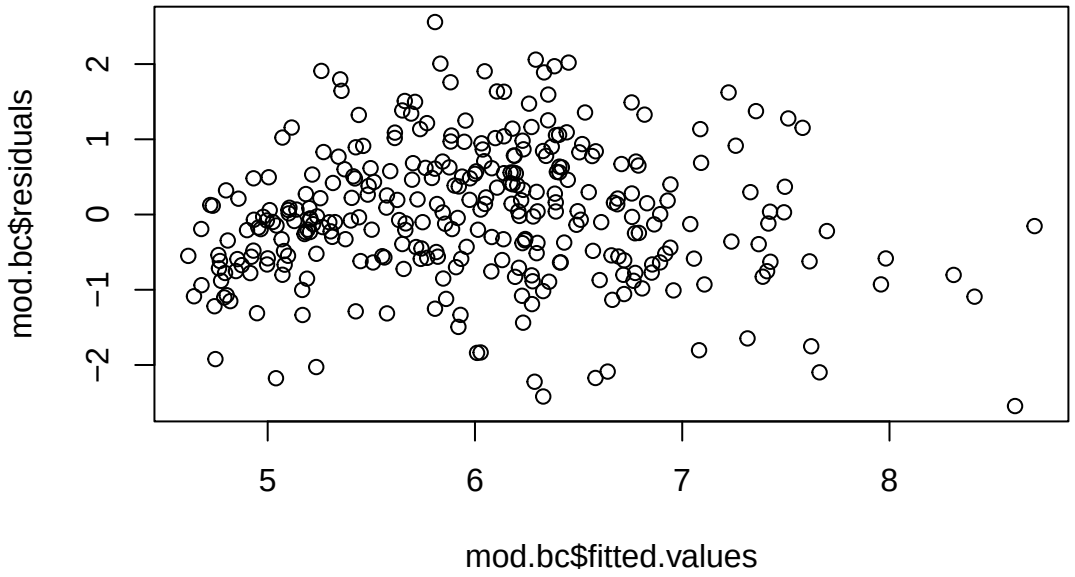
Box Cox transformation:




```
##
## Call:
## lm(formula = ViolentCrimesPerPop^lambda ~ LemasSwFTFieldPerPop +
##     PctLess9thGrade + PctPopUnderPov + PctUnemployed, data = tidy_data2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.54574 -0.58921 -0.03632  0.57516  2.55857
##
## Coefficients:
##                        Estimate Std. Error t value Pr(>|t|)
## (Intercept)           4.0819873  0.1593164  25.622  < 2e-16 ***
## LemasSwFTFieldPerPop  0.0003113  0.0002083   1.494  0.13610
## PctLess9thGrade      -0.0271158  0.0112246  -2.416  0.01627 *
## PctPopUnderPov        0.0854549  0.0110816   7.711 1.66e-13 ***
## PctUnemployed         0.1195709  0.0360796   3.314  0.00103 **
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8969 on 314 degrees of freedom
## Multiple R-squared:  0.4422, Adjusted R-squared:  0.4351
## F-statistic: 62.24 on 4 and 314 DF,  p-value: < 2.2e-16
```

We look at the plot of fitted values vs residuals to see if box cox was effective.

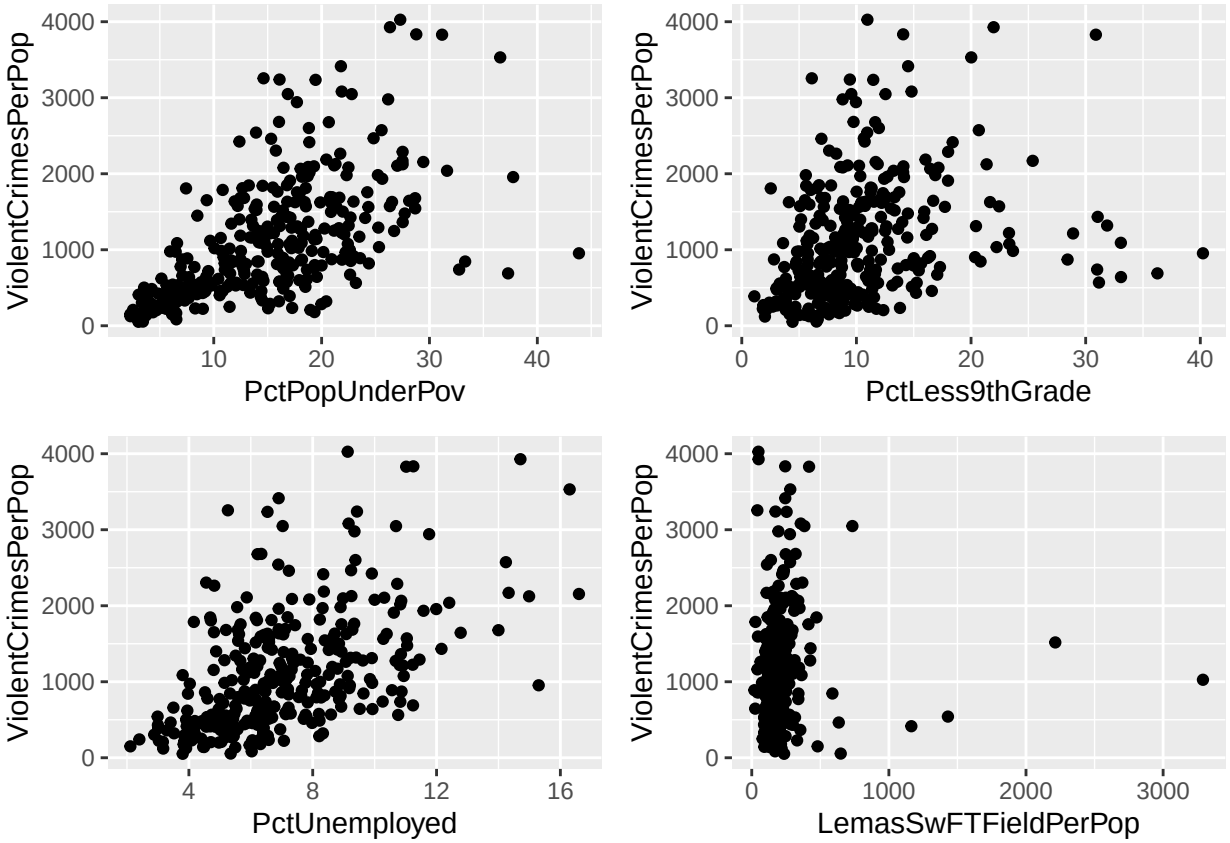


We still see a cone shape suggesting that our assumption of constant error variance is not upheld.We also use the BP test to confirm our findings.

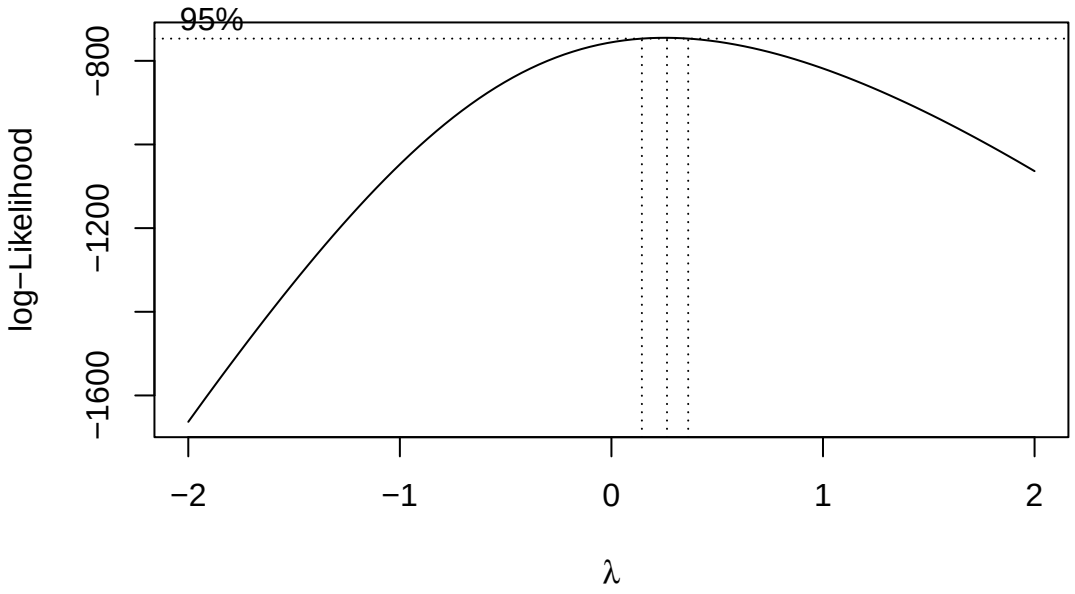
```
##
##  studentized Breusch-Pagan test
##
## data:  mod.bc
## BP = 12.489, df = 4, p-value = 0.01406
```

The p value is less than alpha of 0.05, we reject the null hypothesis. We conclude that model assumption of errors with constant variance is not upheld. Both, the *diagnostic plot* and *BP test* show that the box cox transformation did not help.

Again, lets look at the relationships between predictors and response to decide our next step

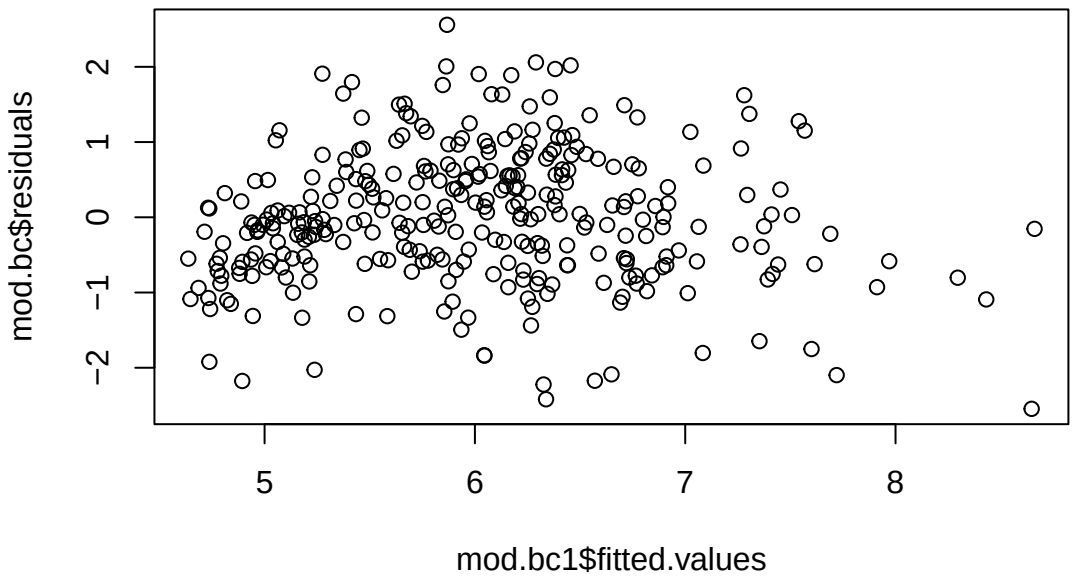


We see that the fourth variable, LemasSwFTFieldPerPop has a very weak correlation with the response variable and hence we can remove it from the model. We fit the reduced model without this variable and try out Box Cox transformation again.




```
##
## Call:
## lm(formula = ViolentCrimesPerPop^lambda ~ PctLess9thGrade + PctPopUnderPov +
##     PctUnemployed, data = tidy_data2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.58730 -0.59873 -0.04742  0.58080  2.49832
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)      4.16061    0.15067  27.614  < 2e-16 ***
## PctLess9thGrade -0.02647    0.01124  -2.356  0.01910 *
## PctPopUnderPov   0.08696    0.01106   7.864 5.97e-14 ***
## PctUnemployed    0.11357    0.03593   3.161  0.00172 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8987 on 315 degrees of freedom
## Multiple R-squared:  0.4383, Adjusted R-squared:  0.4329
## F-statistic: 81.92 on 3 and 315 DF,  p-value: < 2.2e-16
```

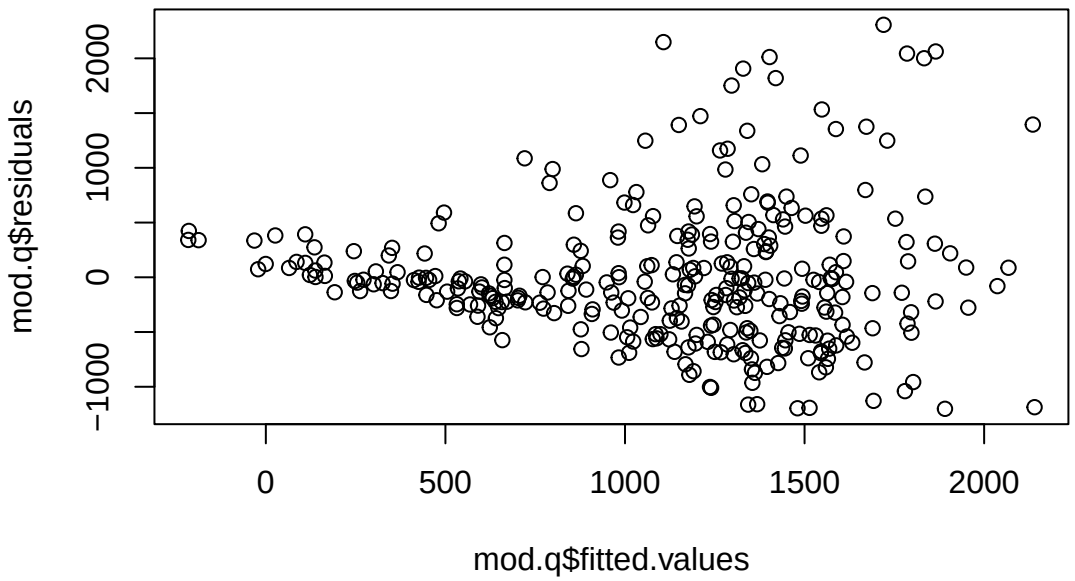
We fit our diagnostic plot and run the BP test for confirmation.




```
##
##  studentized Breusch-Pagan test
##
## data:  mod.bc1
## BP = 13.502, df = 3, p-value = 0.003667
```

We see the same conclusions as before. Since we did not see much improvement, we try other transformations: First we take log of all the variables , fit a linear model, run BP test and see the diagnostic plot.

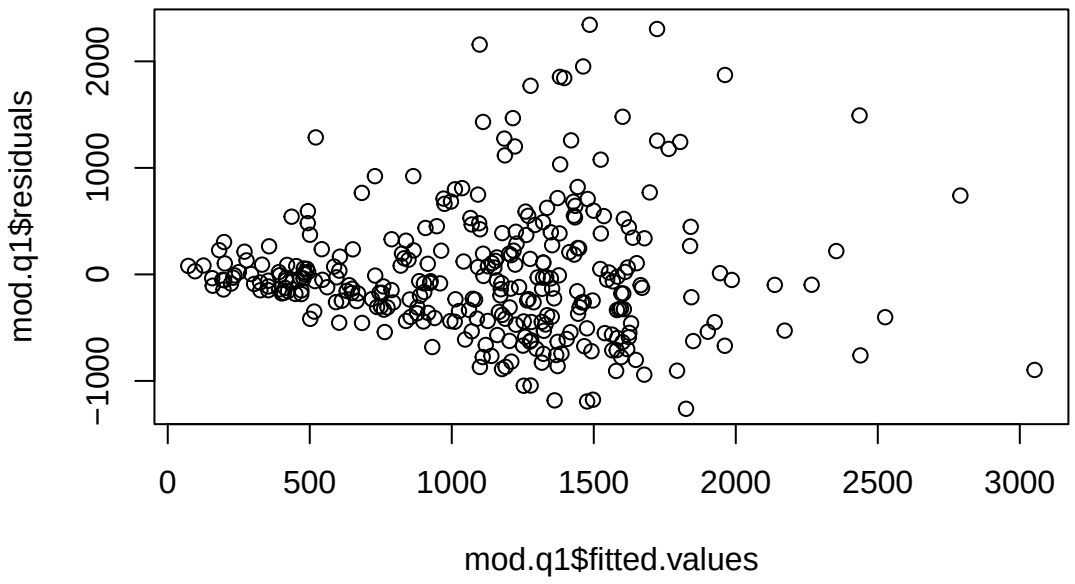
```
##
##  studentized Breusch-Pagan test
##
## data:  mod.q
## BP = 30.142, df = 3, p-value = 1.288e-06
```



We donot see any improvement in diagnostic plot. The p value of BP test is less than alpha of 0.05, we reject the null hypothesis. We conclude that model assumption of errors with constant variance is not upheld.

We try other transformations:

Polynomial regression with d=2 (quadratic model for all variables)

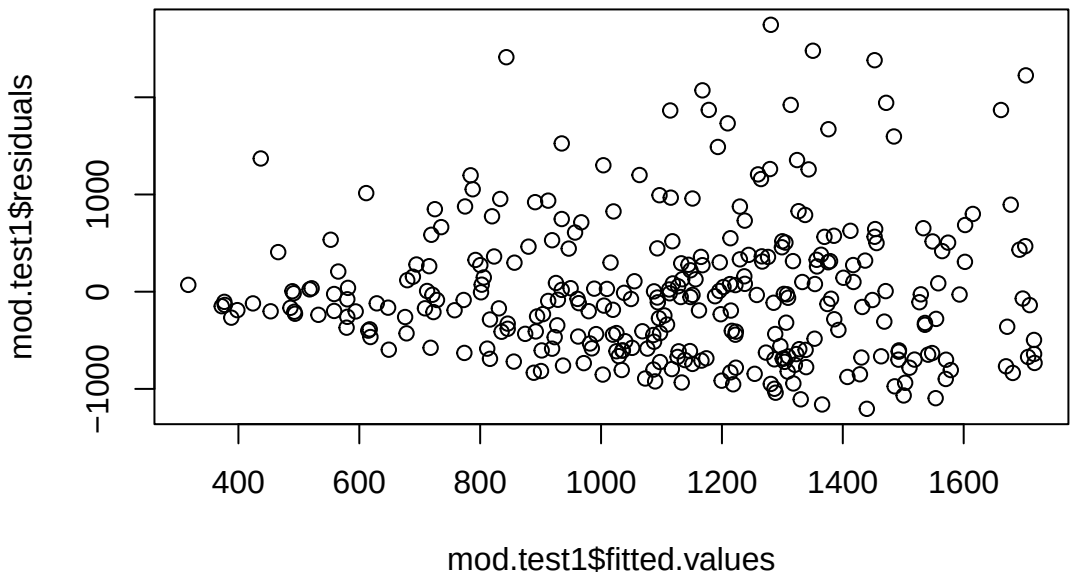


```
##
##  studentized Breusch-Pagan test
##
## data:  mod.q1
## BP = 28.748, df = 6, p-value = 6.789e-05
```

We donot see any improvement in diagnostic plot. The p value of BP test is less than alpha of 0.05, we reject the null hypothesis. We conclude that model assumption of errors with constant variance is not upheld.

We can try other transformations. We can try *polynomial regression but one variable* at a time.First we transform the predictor, PctLess9thGrade and try to find the optimal value of d using polynomial regression:

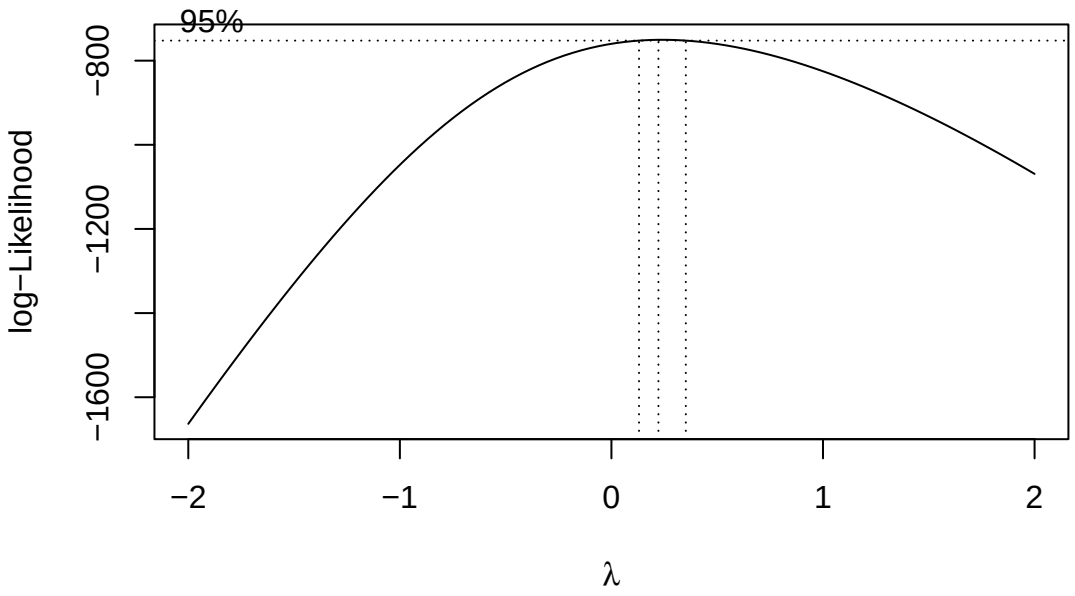
```
##
## Call:
## lm(formula = ViolentCrimesPerPop ~ PctLess9thGrade + I(PctLess9thGrade^2) +
##     I(PctLess9thGrade^3) + I(PctLess9thGrade^4) + I(PctLess9thGrade^5) +
##     I(PctLess9thGrade^6))
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1204.91  -577.35   -85.67   327.36  2745.69
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)           2.743e+02  6.766e+02   0.405    0.685
## PctLess9thGrade       1.164e+01  3.826e+02   0.030    0.976
## I(PctLess9thGrade^2)  2.824e+01  7.740e+01   0.365    0.716
## I(PctLess9thGrade^3) -3.342e+00  7.273e+00  -0.459    0.646
## I(PctLess9thGrade^4)  1.719e-01  3.408e-01   0.504    0.614
## I(PctLess9thGrade^5) -4.169e-03  7.677e-03  -0.543    0.587
## I(PctLess9thGrade^6)  3.802e-05  6.595e-05   0.576    0.565
##
## Residual standard error: 722.1 on 312 degrees of freedom
## Multiple R-squared:  0.1701, Adjusted R-squared:  0.1542
## F-statistic: 10.66 on 6 and 312 DF,  p-value: 8.844e-11
```



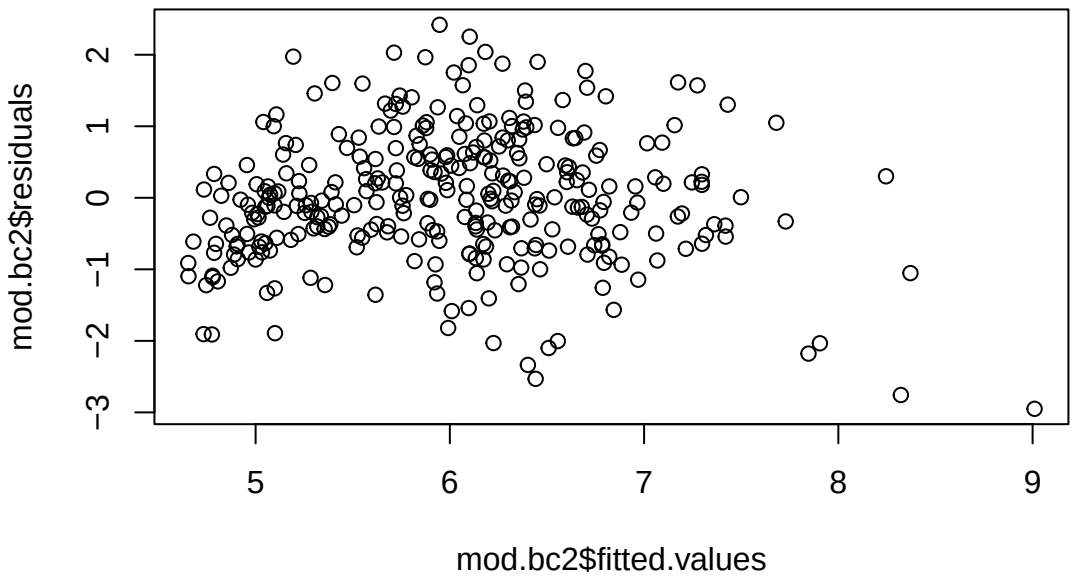

```
##
##  studentized Breusch-Pagan test
##
## data:  mod.test1
## BP = 12.44, df = 6, p-value = 0.05283
```

The p value of the polynomial terms sugges that we can choose a lower value of d. However, the bp test confirms that there is no heteroscedasticity but the diagnostic plot still looks cone shaped.

Next we try to transform the variable, PctPopUnderPov using Box Cox.

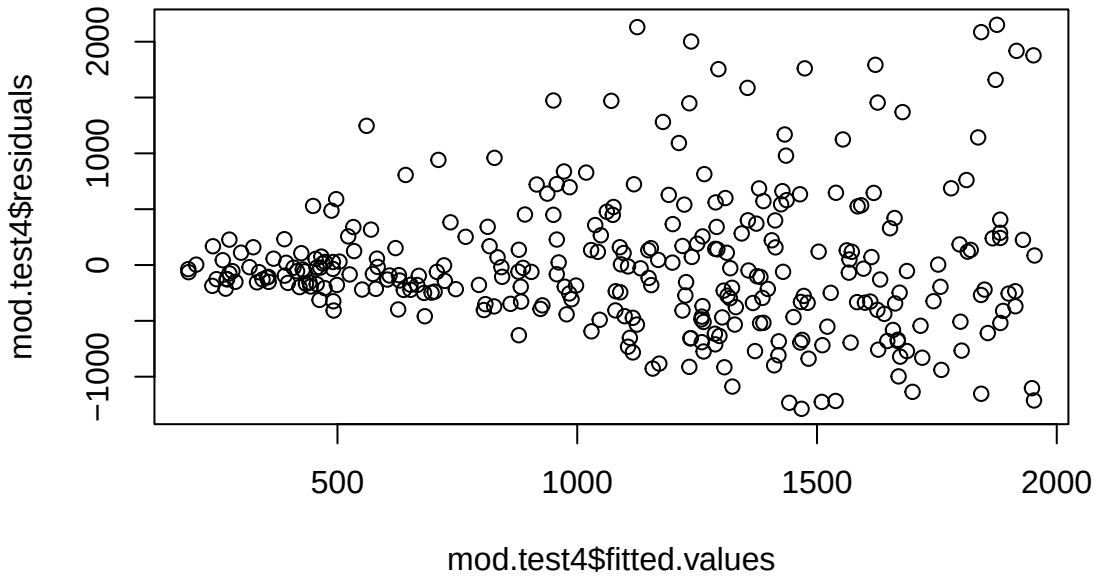



```
##
## Call:
## lm(formula = ViolentCrimesPerPop^lambda ~ PctPopUnderPov, data = tidy_data2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.95018 -0.57005 -0.02764  0.58265  2.41873
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)    4.415712   0.116835    37.8   <2e-16 ***
## PctPopUnderPov 0.104752   0.006936    15.1   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9115 on 317 degrees of freedom
## Multiple R-squared:  0.4185, Adjusted R-squared:  0.4166
## F-statistic: 228.1 on 1 and 317 DF,  p-value: < 2.2e-16
```




```
##
##  studentized Breusch-Pagan test
##
## data:  mod.bc2
## BP = 13.11, df = 1, p-value = 0.0002937
```

The p value is less than alpha of 0.05, we reject the null hypothesis. We conclude that model assumption of errors with constant variance is not upheld. We try polynomial transformation:

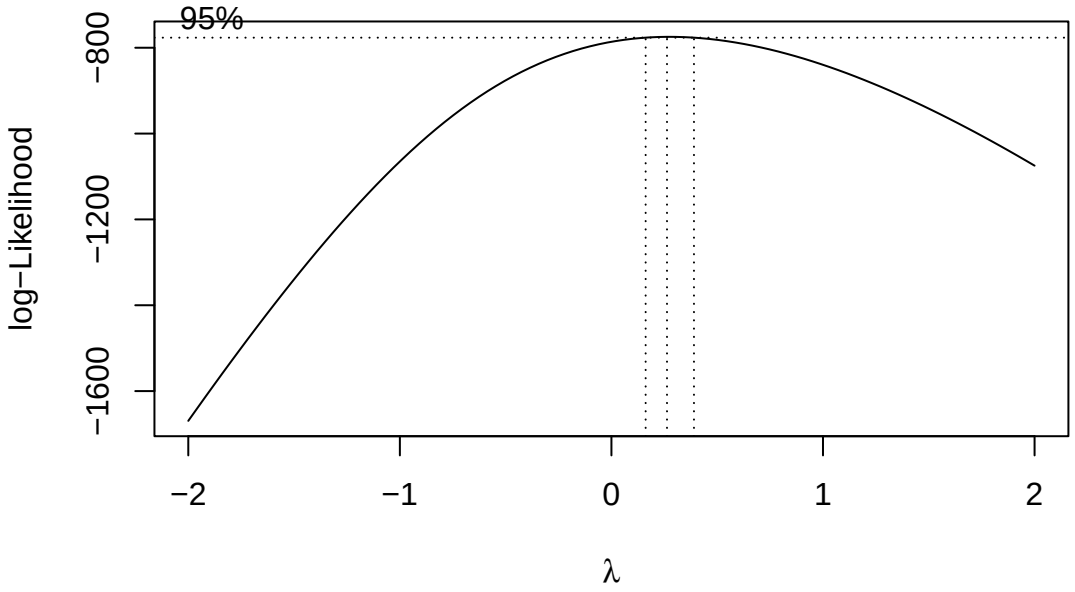



```
##
## Call:
## lm(formula = ViolentCrimesPerPop ~ polym(PctPopUnderPov, degree = 3))
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1287.9  -346.1   -79.9   229.6  2151.3
##
## Coefficients:
##                                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)                         1114.83      34.81  32.030  < 2e-16 ***
## polym(PctPopUnderPov, degree = 3)1  8289.65     621.66  13.335  < 2e-16 ***
## polym(PctPopUnderPov, degree = 3)2 -2171.95     621.66  -3.494 0.000544 ***
## polym(PctPopUnderPov, degree = 3)3  -941.58     621.66  -1.515 0.130869
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 621.7 on 315 degrees of freedom
## Multiple R-squared:  0.3791, Adjusted R-squared:  0.3732
## F-statistic: 64.11 on 3 and 315 DF,  p-value: < 2.2e-16


##
##  studentized Breusch-Pagan test
##
## data:  mod.test4
## BP = 35.516, df = 3, p-value = 9.477e-08
```

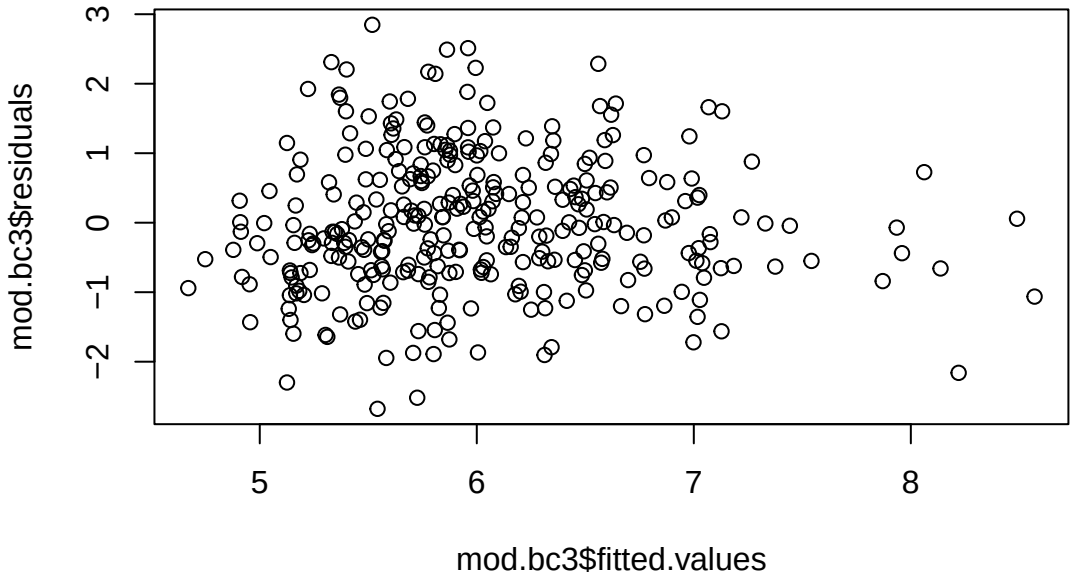
The p value is less than alpha of 0.05, we reject the null hypothesis. The diagnostic plot also suggests

heteroscedasticity. We conclude that model assumption of errors with constant variance is not upheld. Both box cox and poly function did not transform the variable.

Next we try box cox for the PctUnemployed variable:




```
##
## Call:
## lm(formula = ViolentCrimesPerPop^lambda ~ PctUnemployed, data = tidy_data2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.67812 -0.67231 -0.07418  0.63089  2.84754
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)    4.10285    0.16388   25.04   <2e-16 ***
## PctUnemployed  0.26911    0.02186   12.31   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9831 on 317 degrees of freedom
## Multiple R-squared:  0.3234, Adjusted R-squared:  0.3213
## F-statistic: 151.5 on 1 and 317 DF,  p-value: < 2.2e-16
```



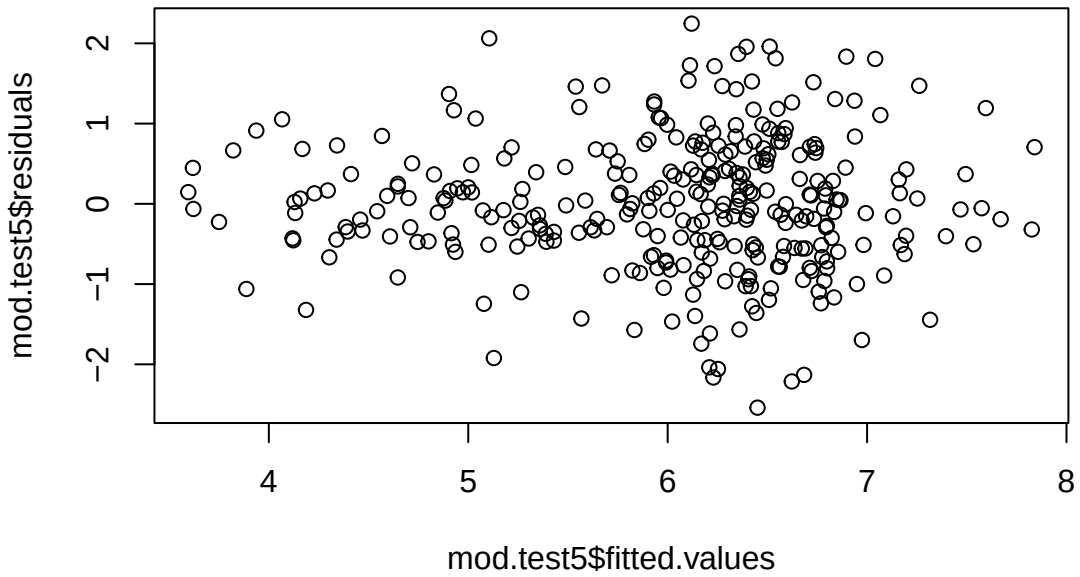

```
##
##  studentized Breusch-Pagan test
##
## data:  mod.bc3
## BP = 1.5025, df = 1, p-value = 0.2203
```

The p value is greater than alpha of 0.05, we fail to reject the null hypothesis. We conclude that model assumption of errors with constant variance is upheld. The diagnostic plot also suggests the same conclusion.

Next, we will combine all the transformations and fit a model. That is, we use polymonial regression for PctLess9thGrade with d =6, use PctPopUnderPov with d=4 and PctUnemployed with Box Cox transformation and lambda = 0.262

```
##
## Call:
## lm(formula = ViolentCrimesPerPop^lambda ~ PctUnemployed + PctLess9thGrade +
##     I(PctLess9thGrade^2) + I(PctLess9thGrade^3) + I(PctLess9thGrade^4) +
##     I(PctLess9thGrade^5) + I(PctLess9thGrade^6) + PctPopUnderPov +
##     I(PctPopUnderPov^2) + I(PctPopUnderPov^3) + I(PctPopUnderPov^4),
##     data = tidy_data2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.54001 -0.50591 -0.02801  0.52387  2.24566
##
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)           1.789e+00  8.822e-01   2.028  0.04344 *
## PctUnemployed         9.513e-02  3.547e-02   2.682  0.00772 **
```

```
## PctLess9thGrade       1.640e-01  4.513e-01   0.363  0.71658
## I(PctLess9thGrade^2) -2.103e-02  9.161e-02  -0.230  0.81856
## I(PctLess9thGrade^3)  6.028e-04  8.650e-03   0.070  0.94449
## I(PctLess9thGrade^4)  3.269e-05  4.082e-04   0.080  0.93621
## I(PctLess9thGrade^5) -1.950e-06  9.285e-06  -0.210  0.83378
## I(PctLess9thGrade^6)  2.546e-08  8.085e-08   0.315  0.75304
## PctPopUnderPov        5.412e-01  1.632e-01   3.315  0.00103 **
## I(PctPopUnderPov^2)  -3.275e-02  1.595e-02  -2.053  0.04088 *
## I(PctPopUnderPov^3)   9.670e-04  6.159e-04   1.570  0.11739
## I(PctPopUnderPov^4)  -1.068e-05  8.100e-06  -1.318  0.18834
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8404 on 307 degrees of freedom
## Multiple R-squared:  0.5212, Adjusted R-squared:  0.5041
## F-statistic: 30.38 on 11 and 307 DF,  p-value: < 2.2e-16
```

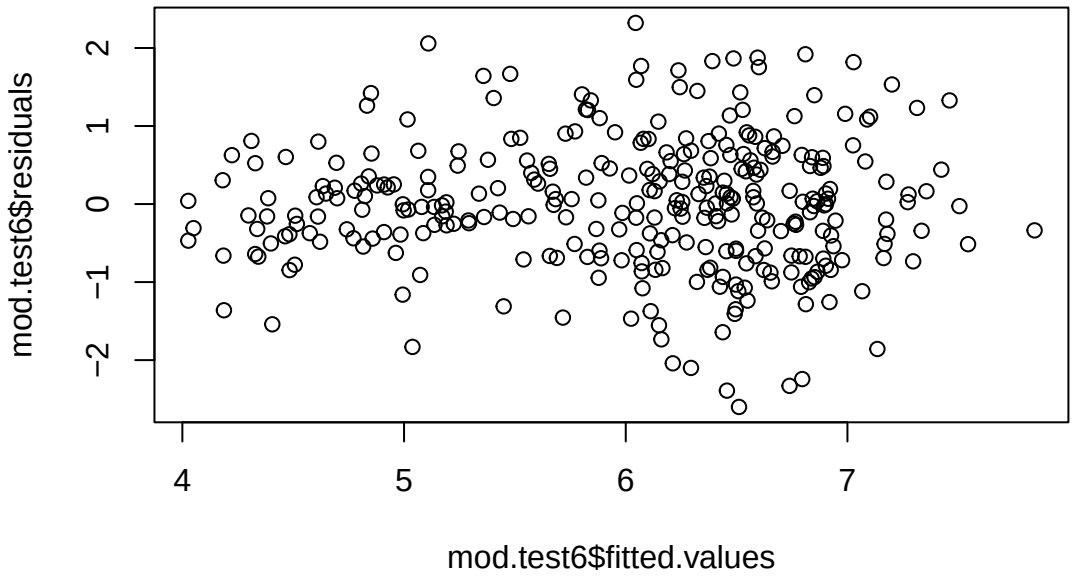



```
##
##  studentized Breusch-Pagan test
##
## data:  mod.test5
## BP = 19.222, df = 11, p-value = 0.05723
```

We see that there is no heteroscedasticity from the diagnostic plot and the BP test also confirms that since the p value is greater than alpha of 0.05. However, the p values of some polynomial terms are not significant suggesting we can reduce the model.

As we can see from the above model, there are many terms that are insignificant that we can remove, the p values are greater than alpha of 0.05, after removing those terms, we get the model below and again use the BP test and diagnostic plot for confirmation.

```
##
## Call:
## lm(formula = ViolentCrimesPerPop^lambda ~ PctUnemployed + PctLess9thGrade +
##     PctPopUnderPov + I(PctPopUnderPov^2), data = tidy_data2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.60029 -0.54248  0.00688  0.52401  2.32126
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)           3.207475   0.201588  15.911  < 2e-16 ***
## PctUnemployed         0.114074   0.033701   3.385 0.000802 ***
## PctLess9thGrade      -0.018296   0.010614  -1.724 0.085734 .
## PctPopUnderPov        0.220095   0.022599   9.739  < 2e-16 ***
## I(PctPopUnderPov^2) -0.004065   0.000613  -6.631 1.46e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.843 on 314 degrees of freedom
## Multiple R-squared:  0.5073, Adjusted R-squared:  0.501
## F-statistic: 80.82 on 4 and 314 DF,  p-value: < 2.2e-16
```




```
##
##  studentized Breusch-Pagan test
##
## data:  mod.test6
## BP = 15.223, df = 4, p-value = 0.00426
```

Even though we fail the BP test, from the diagnostic plot, we no longer see a cone shape and can conclude that the assumption of errors with constant variance is upheld.

Below is the table of the *parameter estimates and p-values of final model*

```
##             Predictor Parameter Estimate      p-value
## 1         (Intercept)        3.207475296 3.255899e-42
## 2       PctUnemployed        0.114074433 8.023342e-04
## 3     PctLess9thGrade       -0.018296233 8.573445e-02
## 4      PctPopUnderPov        0.220095525 9.291880e-20
## 5 I(PctPopUnderPov^2)       -0.004064998 1.456057e-10
```

Below is the R squared for the model.

```
## [1] 0.5072693
```

Next, we compute and report a 95% confidence interval for the slope of PctPopUnderPov predictor:

```
## [1] 0.2643895
```

```
## [1] 0.1758015
```

The 95% confidence interval for the slope of PctPopUnderPov is 0.175 to 0.264.

Next, we compute and report a 95% confidence interval for a prediction at the median value of all predictors.

```
##        fit      lwr      upr
## 1 6.207009 6.090084 6.323935
```

Translating that in y(ViolentCrimesPerPop) scale below:

```
## [1] 1062.352
```

```
## [1] 987.9736
```

```
## [1] 1140.785
```

We are 95% confident that the true mean ViolentCrimesPerPop for all crimes with median predictor values lies in the interval (987,1140)

Next we compute and report a 95% prediction interval for a particular observation.In this case, it is the median value of predictors.

```
##        fit      lwr      upr
## 1 6.207009 4.544227 7.869792
```

Translating that into the y scale:

```
## [1] 1062.352
```

```
## [1] 323.139
```

```
## [1] 2628.495
```

We are 95% confident that the ViolentCrimesPerPop for specific predictor values (median value of each predictor) lies in the interval (323,2628).

Summary

To summarize the model, we discovered that three variables had a strong linear relationship with our response variable, ViolentCrimesPerPop(total number of violent crimes per 100K population). We also used the forward selection method and the StepAIC method to determine if we used predictors with high predictive value. Both methods suggested using all four variables. However, we know that one variable, LemasSwFTFieldPerPop had a very weak relationship with the response and we decided to remove it from our model. We also investigated outliers and influential observations. There were 6 outliers according to the standardized residual method. There were no influential points according to cooks distance method. So even though we have a lot of outliers, they do not disproportionately affect the model. For our linear model, we verified the model assumptions which were as follows: The model had uncorrelated errors as seen from the lagged residual plot and Durbin Watson test. The model's residuals were normally distributed per the QQ plot. However the Shapiro Wilk test did not agree. We know that with large sample sizes the test can be overly sensitive so we sided with the QQ plot. The model assumption of constant error variance was not upheld and we tried different transformations to correct heteroscedasticity. We used both Box Cox transformation and polynomial regression in the final model. The final model's diagnostic plot showed a significant improvement in heteroscedasticity but the Breusch Pagan test did not agree. We concluded that we were successful in correcting heteroscedasticity. In conclusion, we were able to validate our model inferences by verifying all model assumptions. We further used our model to find confidence and prediction intervals for parameter estimates and specific values.

Code

```
library(tidyverse)
library(MASS)
library(gridExtra)
library(lmtest)

#Read in the file
data1 <- data.frame(read.csv("/Users/disha/Documents/Linear Models/Final
↪  Project/CommViolPredUnnormalizedData.txt", header = F))

#Select predictor and response variables with meaningful names
tidy_data <- data1%>% dplyr::select(V1, V2, V146,V107,V38,V35,V34) %>%
↪  rename(communityname= V1, state = V2, ViolentCrimesPerPop =V146,
↪  LemasSwFTFieldPerPop=V107,PctUnemployed = V38, PctLess9thGrade = V35, PctPopUnderPov
↪  = V34)

#Clean up the data
tidy_data1 <- tidy_data
tidy_data1[tidy_data1 == "?"] <- NA
tidy_data1 <- na.omit(tidy_data1)
tidy_data2 <- tidy_data1 %>% mutate(ViolentCrimesPerPop =
↪  as.numeric(ViolentCrimesPerPop),LemasSwFTFieldPerPop =
↪  as.numeric(LemasSwFTFieldPerPop) )
attach(tidy_data2)
glimpse(tidy_data2)

#summary statistics of the variables
print("ViolentCrimesPerPop")
```

```
summary(ViolentCrimesPerPop)
print("LemasSwFTFieldPerPop")
summary(LemasSwFTFieldPerPop)
print("PctLess9thGrade")
summary(PctLess9thGrade)
print("PctUnemployed")
summary(PctUnemployed)
print("PctPopUnderPov")
summary(PctPopUnderPov)

#Distribution of percentage of people under the poverty level vs violent crimes
plot1 <- ggplot(data = tidy_data2, aes(x = PctPopUnderPov, y =
↪  ViolentCrimesPerPop))+geom_point()
#require(scales)
#plot1 + scale_x_continuous(labels = comma)
plot1

#Distribution of percentage of people with education below 9th grade vs violent crimes
plot2 <- ggplot(data = tidy_data2, aes(x = PctLess9thGrade, y =
↪  ViolentCrimesPerPop))+geom_point()
plot2

#Distribution ofpercentage of people unemployed vs violent crimes
plot3 <- ggplot(data = tidy_data2, aes(x =PctUnemployed, y =
↪  ViolentCrimesPerPop))+geom_point()
plot3

#Distribution of sworn full time police officers in field vs violent crimes
plot4 <- ggplot(data = tidy_data2, aes(x = LemasSwFTFieldPerPop, y =
↪  ViolentCrimesPerPop))+geom_point()
plot4

#Correlation values
print("Correlation between LemasSwFTFieldPerPop & ViolentCrimesPerPop")
cor(LemasSwFTFieldPerPop,ViolentCrimesPerPop)
print("Correlation between PctUnemployed & ViolentCrimesPerPop")
cor(PctUnemployed,ViolentCrimesPerPop)
print("Correlation between PctLess9thGrade & ViolentCrimesPerPop")
cor(PctLess9thGrade,ViolentCrimesPerPop)
print("Correlation between PctPopUnderPov & ViolentCrimesPerPop")
cor(PctPopUnderPov,ViolentCrimesPerPop)

# Fitting a linear model with the response and predictor variables
mod1 <-
↪  lm(ViolentCrimesPerPop~LemasSwFTFieldPerPop+PctLess9thGrade+PctPopUnderPov+PctUnemployed)
summary(mod1)

#performing fastbw on our linear model for variable selection
require(rms)
ols.vp<-
ols(ViolentCrimesPerPop~LemasSwFTFieldPerPop+PctLess9thGrade+PctPopUnderPov+PctUnemployed,data
↪  = tidy_data2)
#Perform p-value based selection using fastbw() function
```

```
fastbw(ols.vp, rule = "p", sls = 0.75)

#using stepAIC
aicmod <- stepAIC(mod1)
aicmod$anova

#checking constant variance of errors
plot(mod1$fitted.values,mod1$residuals)

#qq plot to check normality of model residuals
qqnorm(mod1$residuals)

#plot of succesive pairs of residuals
n <- length(residuals(mod1))
plot(tail(residuals(mod1),n-1)~head(residuals(mod1),n-1), xlab =
↪  expression(hat(epsilon)[i]), ylab = expression(hat(epsilon)[i+1]) )

#cooks distance
n <- dim(model.matrix(mod1))[1]
p <- dim(model.matrix(mod1))[2]
numdf <- p
dendf <- n-p
fthresh <- qf(0.5,numdf,dendf)
which(cooks.distance(mod1)>fthresh)
which(cooks.distance(mod1)>1)

#standardized residuals
stdr <- data.frame(round(rstandard(mod1),4))
abs(stdr)[abs(stdr)>3]

#BP Test
lmtest::bptest(mod1)

#Shapiro Wilk Test
shapiro.test(mod1$residuals)

#Durbin Watson Test
lmtest::dwtest(mod1)

#Box Cox Transformation, diagnostic plot and BP test
require(MASS)
bc <- MASS::boxcox(mod1, plotit=T)
lambda <- bc$x[which.max(bc$y)]
lambda
mod.bc <- lm(ViolentCrimesPerPop^lambda~LemasSwFTFieldPerPop+PctLess9thGrade+
               PctPopUnderPov+PctUnemployed,data = tidy_data2)
summary(mod.bc)
plot(mod.bc$fitted.values,mod.bc$residuals)

#exploring relationships between variables
grid.arrange(plot1,plot2,plot3,plot4,ncol = 2)

#fit model after removing fourth variable
```

```
mod.reduced <- lm(ViolentCrimesPerPop~PctLess9thGrade+PctPopUnderPov+PctUnemployed,data =
↪  tidy_data2)

#box cox transformation on model without the fourth variable
require(MASS)
bc1 <- boxcox(mod.reduced, plotit=T)
lambda <- bc$x[which.max(bc$y)]
lambda
mod.bc1 <-
↪  lm(ViolentCrimesPerPop^lambda~PctLess9thGrade+PctPopUnderPov+PctUnemployed,data =
↪  tidy_data2)
summary(mod.bc1)
plot(mod.bc$fitted.values,mod.bc$residuals)
lmtest::bptest(mod.bc, data = tidy_data2)


# Transforming predictors by taking the log of all predictors
mod.q <- lm(formula = ViolentCrimesPerPop~ log(PctLess9thGrade) + log(PctPopUnderPov) +
    log(PctUnemployed), data = tidy_data2)
bptest(mod.q,data = tidy_data2)
plot(mod.q$fitted.values,mod.q$residuals)

#Trying polynomial regression for all variables
mod.q1 <- lm(formula = ViolentCrimesPerPop~ PctLess9thGrade +I(PctLess9thGrade^2)+
↪  PctPopUnderPov +I(PctPopUnderPov^2) +PctUnemployed +I(PctUnemployed^2), data =
↪  tidy_data2)
plot(mod.q1$fitted.values,mod.q1$residuals)
bptest(mod.q1,data = tidy_data2)

#transforming variable PctLess9thGrade using polynomial regression
mod.test1 <- lm(ViolentCrimesPerPop~
↪  PctLess9thGrade+I(PctLess9thGrade^2)+I(PctLess9thGrade^3)+I(PctLess9thGrade^4)+
                  I(PctLess9thGrade^5)+I(PctLess9thGrade^6))
summary(mod.test1)
plot(mod.test1$fitted.values,mod.test1$residuals)
bptest(mod.test1,data = tidy_data2)

#transforming variable PctPopUnderPov using box cox transformation
require(MASS)
#Modeling variable PctPopUnderPov
mod.test2 <- lm( ViolentCrimesPerPop~ PctPopUnderPov)
bc2 <- boxcox(mod.test2, plotit=T)
lambda <- bc$x[which.max(bc$y)]
lambda
mod.bc2 <- lm(ViolentCrimesPerPop^lambda~PctPopUnderPov,data = tidy_data2)
summary(mod.bc2)
plot(mod.bc2$fitted.values,mod.bc2$residuals)
bptest(mod.bc2,data = tidy_data2)

#Trying the poly function for PctPopUnderPov variable
mod.test4 <- lm( ViolentCrimesPerPop~ polym(PctPopUnderPov,degree=3))
plot(mod.test4$fitted.values,mod.test4$residuals)
summary(mod.test4)
```

```
bptest(mod.test4,data = tidy_data2)

#Trying box cox transformation for variable PctUnemployed
require(MASS)
mod.test5 <- lm(ViolentCrimesPerPop~PctUnemployed,data = tidy_data2)
bc3 <- boxcox(mod.test5, plotit=T)
lambda <- bc$x[which.max(bc$y)]
lambda
mod.bc3 <- lm(ViolentCrimesPerPop^lambda~PctUnemployed,data = tidy_data2)
summary(mod.bc3)
plot(mod.bc3$fitted.values,mod.bc3$residuals)
bptest(mod.bc3,data = tidy_data2)
lambda #lambda value

#Model combining all the variable transformations explored above
mod.test5 <-
↪  lm(ViolentCrimesPerPop^lambda~PctUnemployed+PctLess9thGrade+I(PctLess9thGrade^2)+
                I(PctLess9thGrade^3)+I(PctLess9thGrade^4)+I(PctLess9thGrade^5)
              +I(PctLess9thGrade^6)+PctPopUnderPov+I(PctPopUnderPov^2)
              +I(PctPopUnderPov^3)+I(PctPopUnderPov^4),data = tidy_data2)
summary(mod.test5)
plot(mod.test5$fitted.values,mod.test5$residuals)
bptest(mod.test5,data = tidy_data2)

#Reduced final model
mod.test6 <- lm(ViolentCrimesPerPop^lambda~PctUnemployed+
                PctLess9thGrade+PctPopUnderPov+I(PctPopUnderPov^2),data = tidy_data2)
summary(mod.test6)
plot(mod.test6$fitted.values,mod.test6$residuals)
bptest(mod.test6,data = tidy_data2)

#Table for model estimates and p values
mod.test6 <- lm(ViolentCrimesPerPop^lambda~PctUnemployed+
                PctLess9thGrade+PctPopUnderPov+I(PctPopUnderPov^2),data = tidy_data2)

table1 <- data.frame(summary(mod.test6)$coefficients[,-2:-3])
table2 <- cbind(Predictor = rownames(table1), table1)
rownames(table2) <- 1:nrow(table2)
colnames(table2) <- c("Predictor","Parameter Estimate","p-value")
table2

#R squared value for the model
summary(mod.test6)$r.squared

#  95% Confidence Interval for slope of PctPopUnderPov:
upper.beta1 <- table2[4,2] + 1.96*summary(mod.test6)$coef[[4,2]]
lower.beta1 <- table2[4,2] - 1.96* summary(mod.test6)$coef[[4,2]]

upper.beta1
lower.beta1

#95% confidence interval for prediction at median values of predictors
tidy_data3 <- tidy_data2 %>%
↪  dplyr::select(ViolentCrimesPerPop,PctUnemployed,PctLess9thGrade,PctPopUnderPov)
```

```
mod.test6 <- lm(ViolentCrimesPerPop^lambda~PctUnemployed+
                  PctLess9thGrade+PctPopUnderPov+I(PctPopUnderPov^2),data = tidy_data3)
predict(mod.test6, new = data.frame(PctUnemployed =
↪  median(PctUnemployed),PctLess9thGrade= median(PctLess9thGrade),PctPopUnderPov =
↪  median(PctPopUnderPov)), interval = "confidence")

#Confidence interval in response variable scale
#lambda = 0.262
6.207009^(1/0.262)
6.090084^(1/0.262)
6.323935^(1/0.262)

#95% prediction interval for an observation(median values of predictor)
predict(mod.test6, new = data.frame(PctUnemployed =
↪  median(PctUnemployed),PctLess9thGrade= median(PctLess9thGrade),PctPopUnderPov =
↪  median(PctPopUnderPov)), interval = "prediction")

#Prediction interval in the response variable scale
#lambda = 0.262
6.207009^(1/0.262)
4.544227^(1/0.262)
7.869792^(1/0.262)
```