Violas Group Presentation V1

Joe, Jason, Pam

August 4, 2018

Preliminaries

Review of the Guns dataset from the "AER"" package.

- **Dataset is a balanced panel of data on 50 US states, plus the District of Columbia (for a total of 51 states), by year for 1977-1999, and contains 1,173 observations on 13 variables. The variables are:**
  - ***state:*** factor indicating state.
  - ***year:*** factor indicating year.
  - ***violent:*** violent crime rate (incidents per 100,000 members of the population).
  - ***murder:*** murder rate (incidents per 100,000).
  - ***robbery:*** robbery rate (incidents per 100,000).
  - ***prisoners:*** incarceration rate in the state in the previous year (sentenced prisoners per 100,000 residents;value for the previous year).
  - ***afam:*** percent of state population that is African-American, ages 10 to 64.
  - ***cauc:*** percent of state population that is Caucasian, ages 10 to 64.
  - ***male:*** percent of state population that is male, ages 10 to 29.

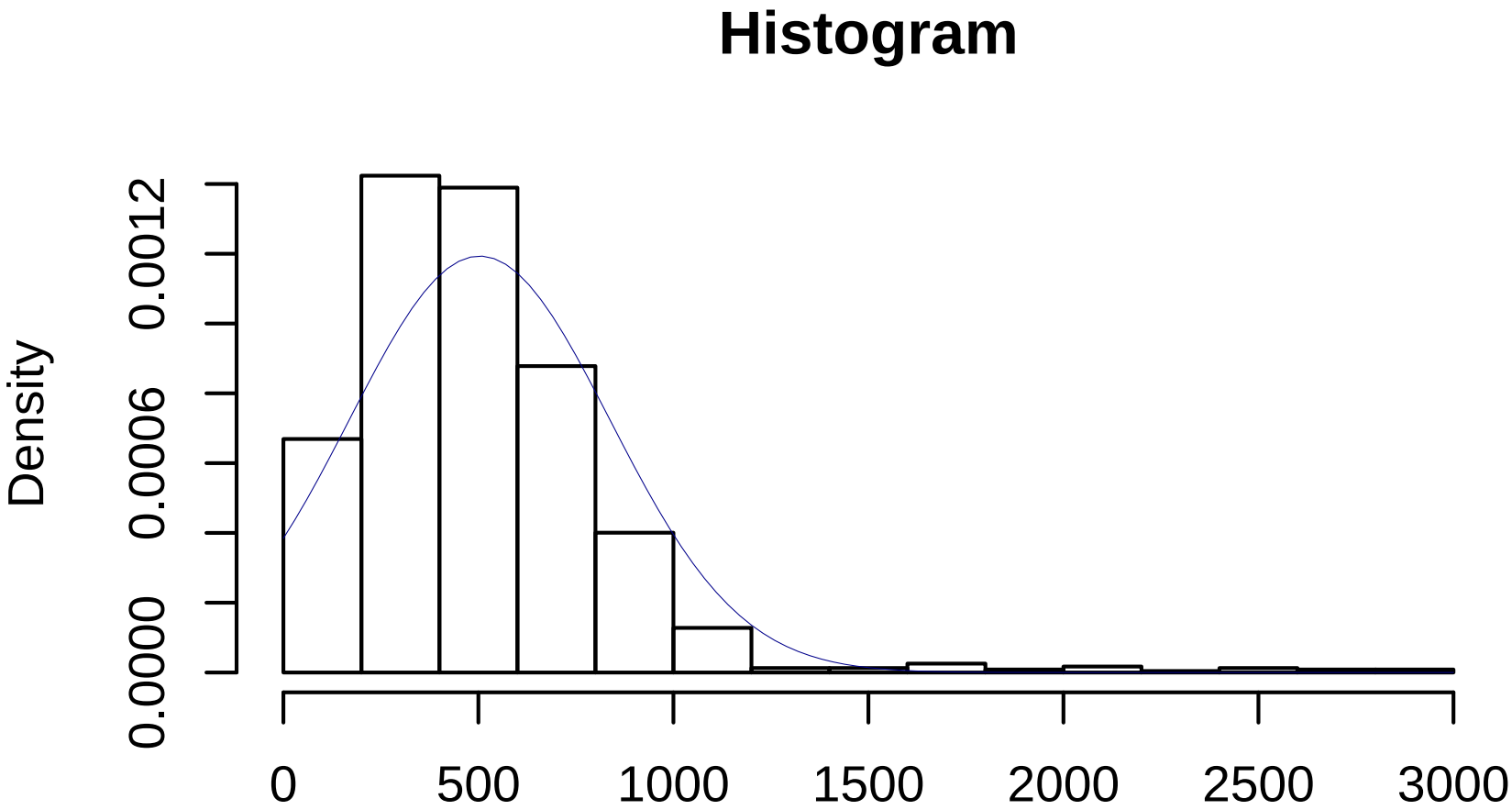
Review the structure of the data

```
## 'data.frame':    1173 obs. of  13 variables:
##  $ year      : Factor w/ 23 levels "1977","1978",..: 1 2
##  $ violent   : num  414 419 413 448 470 ...
##  $ murder    : num  14.2 13.3 13.2 13.2 11.9 10.6 9.2 9.
##  $ robbery   : num  96.8 99.1 109.5 132.1 126.5 ...
##  $ prisoners : int  83 94 144 141 149 183 215 243 256 26
##  $ afam      : num  8.38 8.35 8.33 8.41 8.48 ...
##  $ cauc      : num  55.1 55.1 55.1 54.9 54.9 ...
##  $ male      : num  18.2 18 17.8 17.7 17.7 ...
##  $ population: num  3.78 3.83 3.87 3.9 3.92 ...
##  $ income    : num  9563 9932 9877 9541 9548 ...
##  $ density   : num  0.0746 0.0756 0.0762 0.0768 0.0772 .
##  $ state     : Factor w/ 51 levels "Alabama","Alaska",..
##  $ law       : Factor w/ 2 levels "no","yes": 1 1 1 1 1
```

##Review the Summary of the data

```
##       year          violent            murder           robb
##  1977   : 51   Min.   :  47.0   Min.   : 0.200   Min.
##  1978   : 51   1st Qu.: 283.1   1st Qu.: 3.700   1st Qu.
##  1979   : 51   Median : 443.0   Median : 6.400   Median
##  1980   : 51   Mean   : 503.1   Mean   : 7.665   Mean
##  1981   : 51   3rd Qu.: 650.9   3rd Qu.: 9.800   3rd Qu.
##  1982   : 51   Max.   :2921.8   Max.   :80.600   Max.
##  (Other):867
##    prisoners           afam             cauc
##  Min.   :  19.0   Min.   : 0.2482   Min.   :21.78   Min.
##  1st Qu.: 114.0   1st Qu.: 2.2022   1st Qu.:59.94   1st
##  Median : 187.0   Median : 4.0262   Median :65.06   Medi
##  Mean   : 226.6   Mean   : 5.3362   Mean   :62.95   Mean
##  3rd Qu.: 291.0   3rd Qu.: 6.8507   3rd Qu.:69.20   3rd
##  Max.   :1913.0   Max.   :26.9796   Max.   :76.53   Max.
##
##    population          income          density
##  Min.   : 0.4027   Min.   : 8555   Min.   : 0.000707   M
```

Review data visually



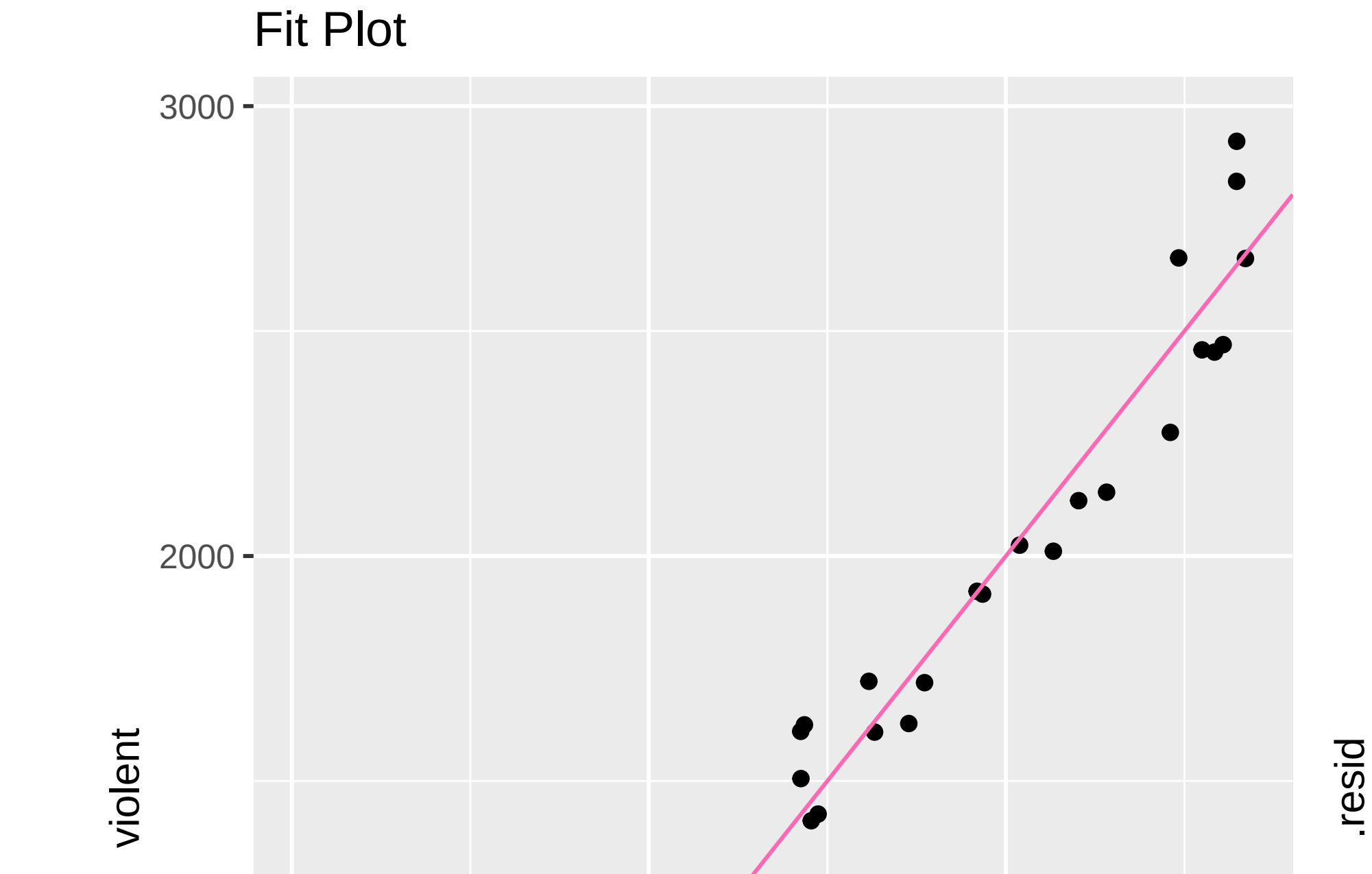
Sort Plot

The violent variable does not appear to be normally distributed

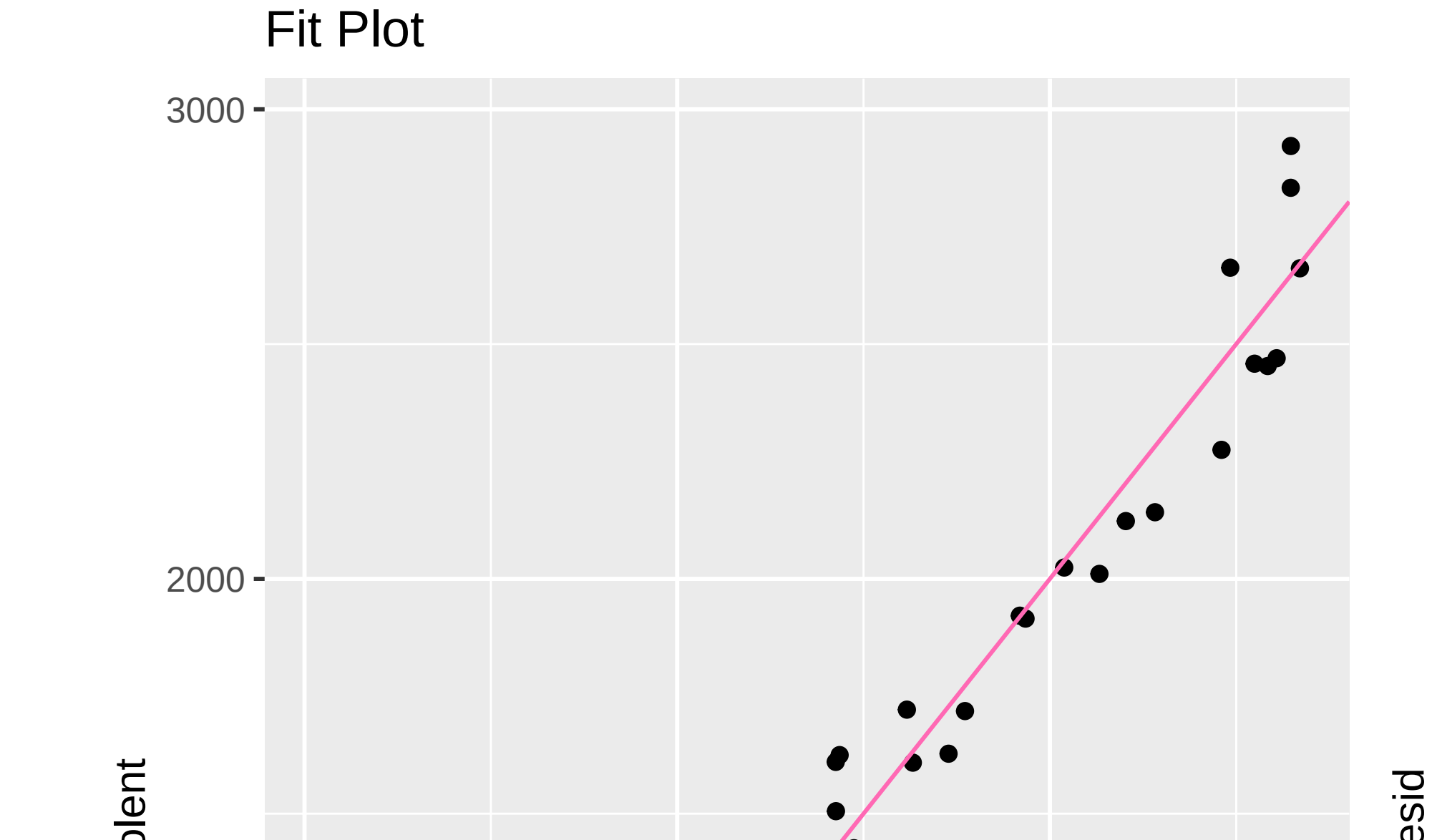
- Histogram, the distribution of the violent variable deviates from the normal distribution
  - Doesn't adhere to the blue QQ-line on the the QQ-Plot.
  - Violent variable is skewed to the right (positively) and the mean (503.074) $>$ median (443)
- Sort Plot there appears to be a few clusters on the right side of the plot, where the violent crime rate is high
- Consider the use a ?log()? transformation on this variable possibly reduce the skewness
- Potentially make it easier to interpret whether or not a relationship exist between variables that might not have otherwise been obvious.

Scatterplot

Fit Plot Original Model (mOrig)



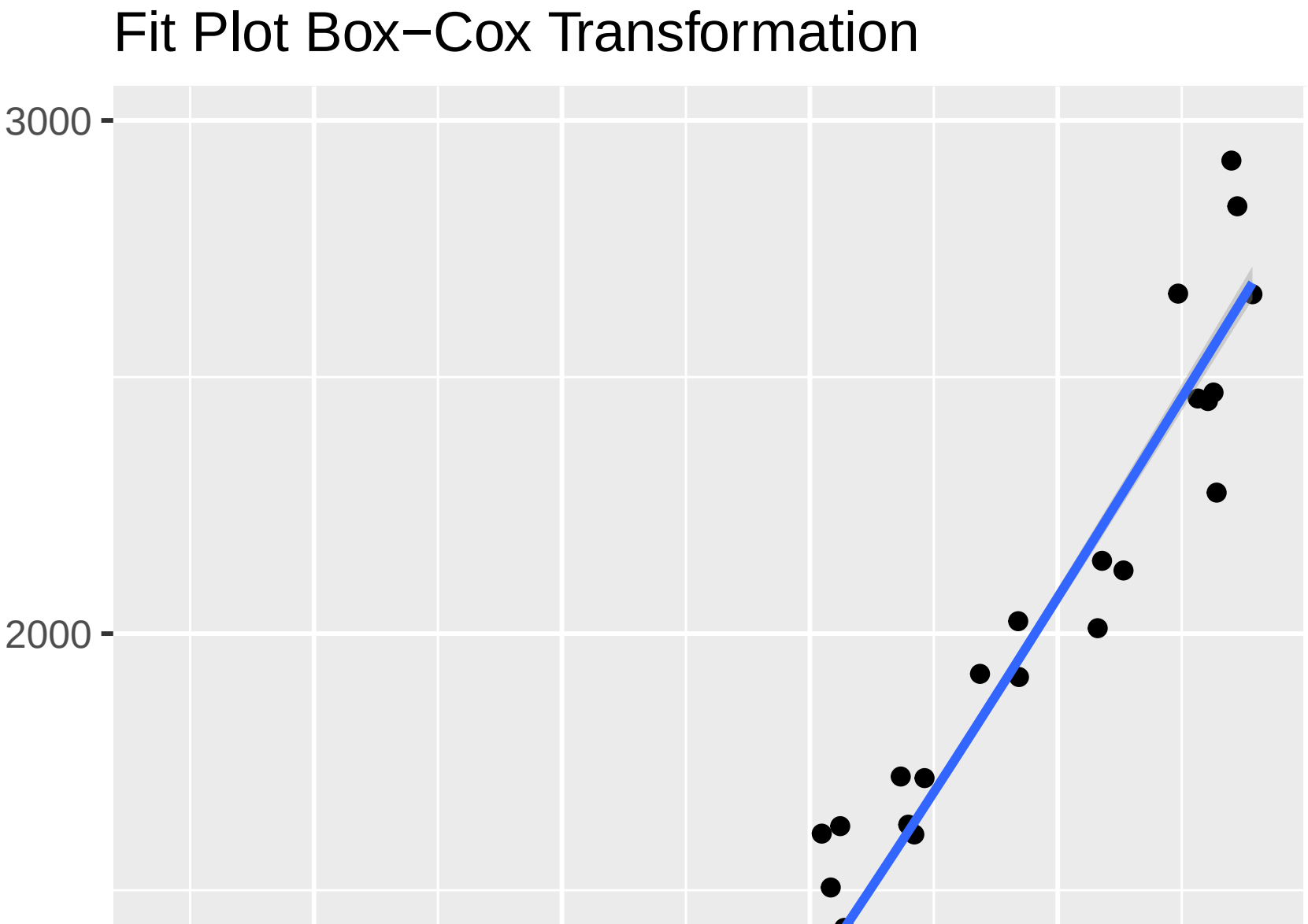
Log Transformation Fitted Plot & Transformation Residuals vs. Fitted (mLog)



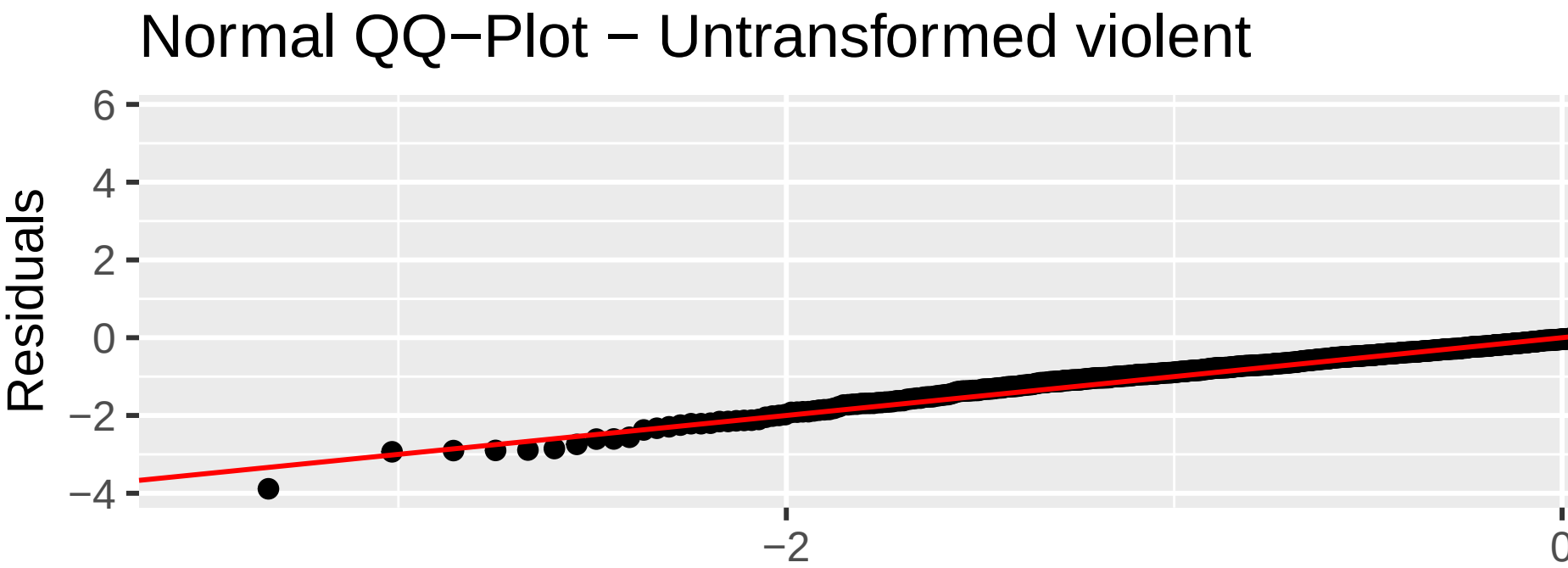
Box-Cox Transformation

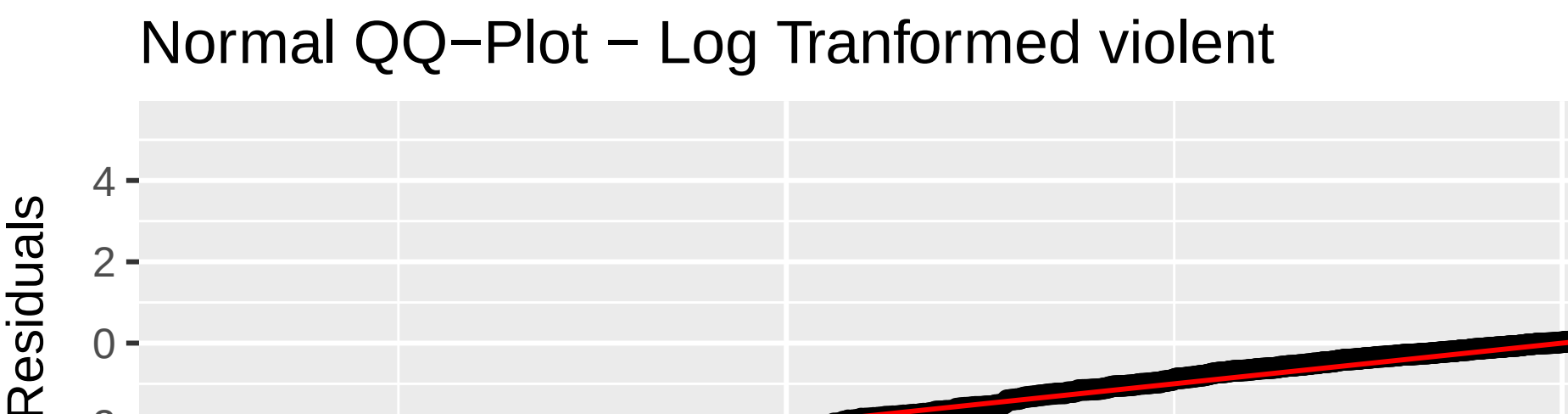
```
## Estimated transformation parameter
##        Y1
## 0.6022324
```

Box-Cox Transformation Fitted Plot & Transformation Residuals vs. Fitted (mBC)

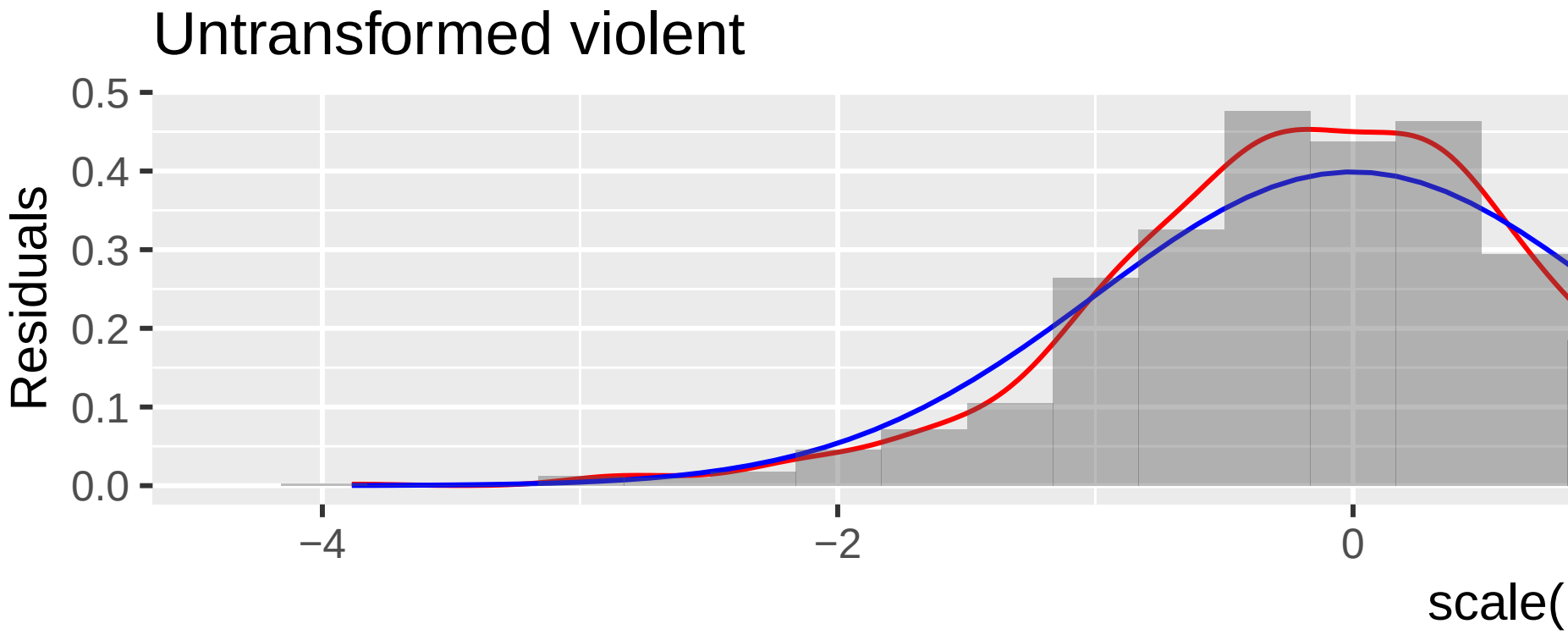


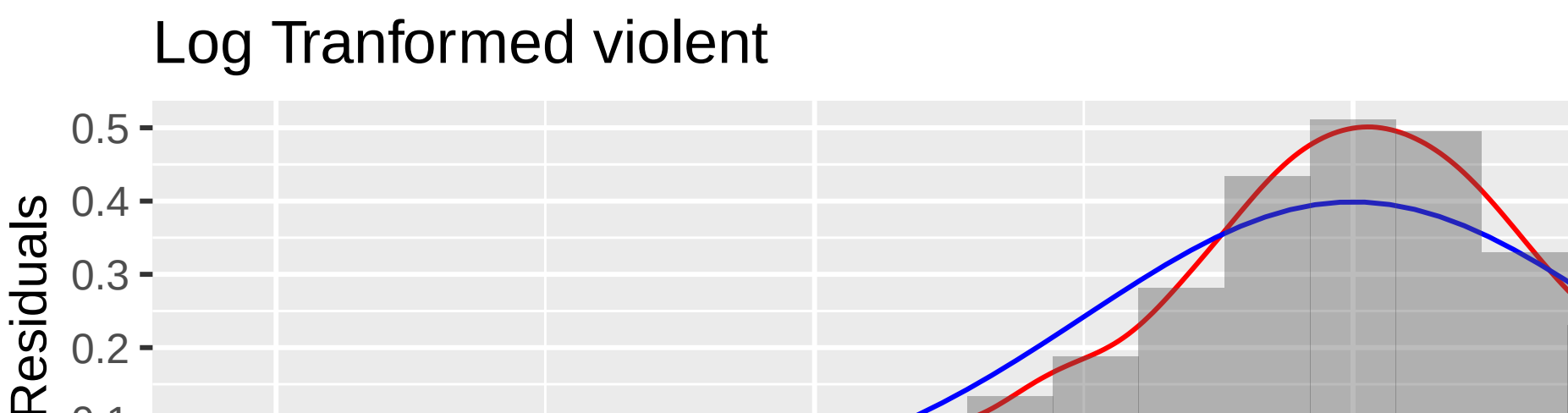
Normal QQ-plots for detecting non normality





Histogram


Untransformed violent
Residuals
0.5
0.4
0.3
0.2
0.1
0.0
−4
−2
0
scale(


Log Tranformed violent
Residuals
0.5
0.4
0.3
0.2

Shapiro test of Normality

```
##
##  Shapiro-Wilk normality test
##
## data:  residuals(mOrig)
## W = 0.97023, p-value = 8.514e-15

##
##  Shapiro-Wilk normality test
##
## data:  residuals(mLog)
## W = 0.97727, p-value = 1.293e-12

##
##  Shapiro-Wilk normality test
##
## data:  residuals(mBC)
## W = 0.98664, p-value = 6.98e-09
```

Stepwise Regression Box Cox Model

```
## Start:  AIC=2239.81
## violent^lam ~ year + murder + robbery + prisoners + afam
##     male + population + income + density + state + law
##
##               Df Sum of Sq   RSS    AIC
## - cauc         1         0  6872 2237.8
## - density      1         0  6873 2237.9
## - afam         1         6  6879 2238.9
## - population   1         7  6879 2239.0
## - income       1         8  6880 2239.2
## <none>                     6872 2239.8
## - law          1        78  6950 2251.0
## - male         1       100  6972 2254.7
## - murder       1       132  7005 2260.2
## - prisoners    1       218  7090 2274.4
## - year        22      1632  8504 2445.7
## - robbery      1      4594 11467 2838.3
## - state       50     33830 40702 4226.3
```

Stepwise Regression Log Model

```
## Start:  AIC=-4785.25
## log(violent) ~ year + murder + robbery + prisoners + afa
##     male + population + income + density + state + law
##
##               Df Sum of Sq     RSS     AIC
## - murder       1     0.000  17.224 -4787.2
## - density      1     0.007  17.231 -4786.8
## - population   1     0.007  17.231 -4786.8
## - income       1     0.008  17.232 -4786.7
## - afam         1     0.018  17.242 -4786.0
## <none>                      17.224 -4785.3
## - cauc         1     0.043  17.267 -4784.3
## - prisoners    1     0.124  17.348 -4778.9
## - law          1     0.151  17.375 -4777.0
## - male         1     0.258  17.482 -4769.8
## - year        22     3.406  20.630 -4617.6
## - robbery      1     3.484  20.708 -4571.2
## - state       50   117.192 134.416 -2475.2
```

Compare Coefficients of the three models

```
## Calls:
## 1: lm(formula = violent^lam ~ ., data = Guns)
## 2: lm(formula = violent^lam ~ year + murder + robbery +
##   afam + male + state + law, data = Guns)
## 3: lm(formula = log(violent) ~ ., data = Guns)
## 4: lm(formula = log(violent) ~ year + robbery + prisoner
##   male + state + law, data = Guns)
##
##                                Model 1   Model 2   Model 3
## (Intercept)                      11.31     11.56      3.98
## year1978                        0.8856    0.9553    0.0480
## year1979                         2.325     2.389     0.128
## year1980                         2.649     2.671     0.150
## year1981                         2.410     2.450     0.135
## year1982                         2.567     2.602     0.130
## year1983                         2.323     2.391     0.114
## year1984                         3.259     3.418     0.157
## year1985                         4.141     4.346     0.202
```

P-Value of the Partial F-test Box Cox Model and Stepwise Box Cox Model

```
## Analysis of Variance Table
##
## Model 1: violent^lam ~ year + murder + robbery + prisone
##     state + law
## Model 2: violent^lam ~ year + murder + robbery + prisone
##     male + population + income + density + state + law
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1   1094 6889.0
## 2   1090 6872.4  4    16.557 0.6565 0.6224
```

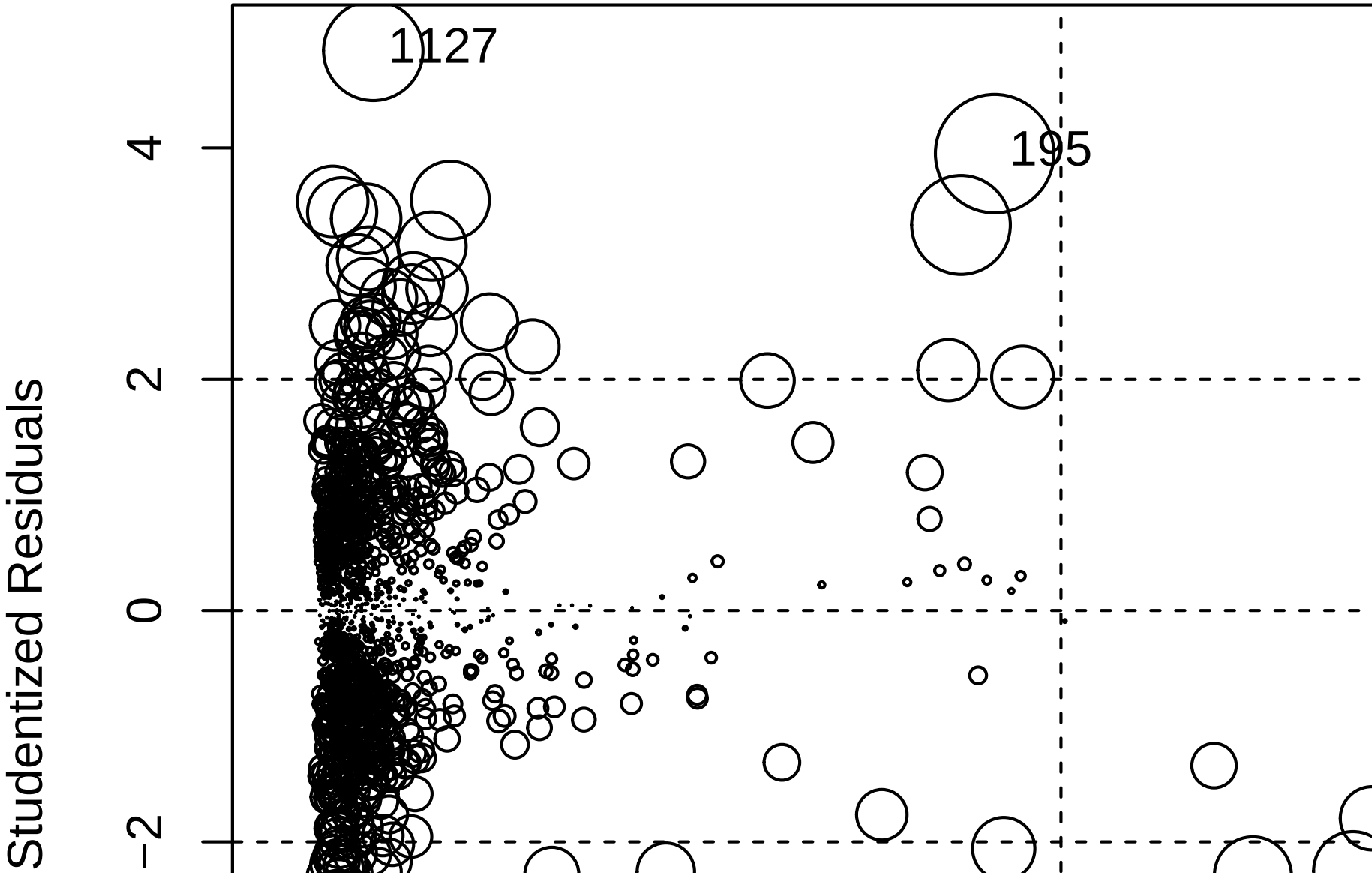
The results produced by the anova function show a non-significant result (p-value = 0.6224). Therefore we should reject the Box-Cox transformed larger model (mBC) and move forward with the Box-Cox transformed smaller model (mBCStep).

P-Value of the Partial F-test Log Model and Stepwise Log Model

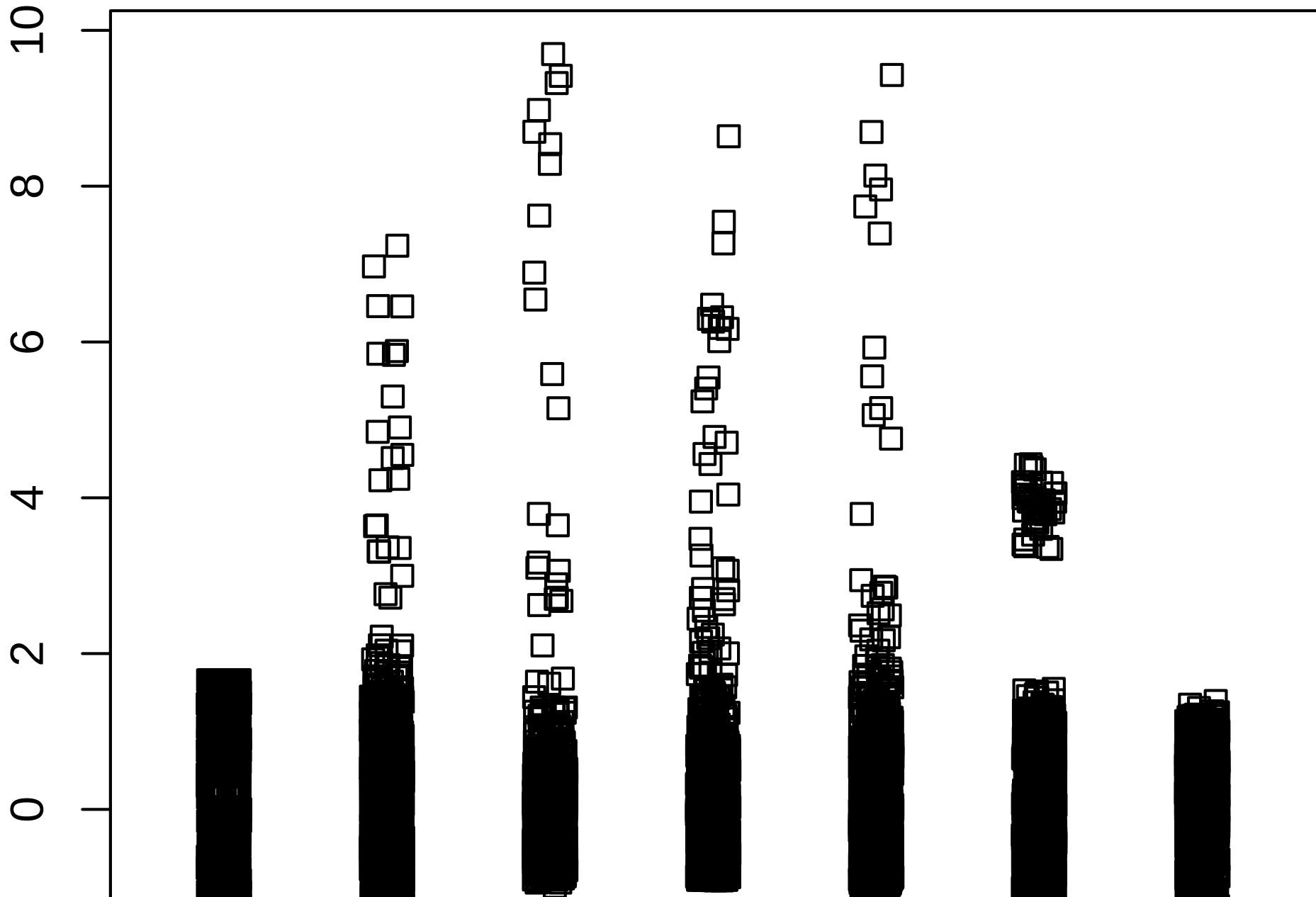
```
## Analysis of Variance Table
##
## Model 1: log(violent) ~ year + robbery + prisoners + cau
##     law
## Model 2: log(violent) ~ year + murder + robbery + prison
##     male + population + income + density + state + law
##   Res.Df    RSS Df Sum of Sq     F Pr(>F)
## 1   1095 17.266
## 2   1090 17.224  5  0.042034 0.532 0.7522
```

WHAT DO WE DO HERE????

Checking for any influential points that are controlling the results



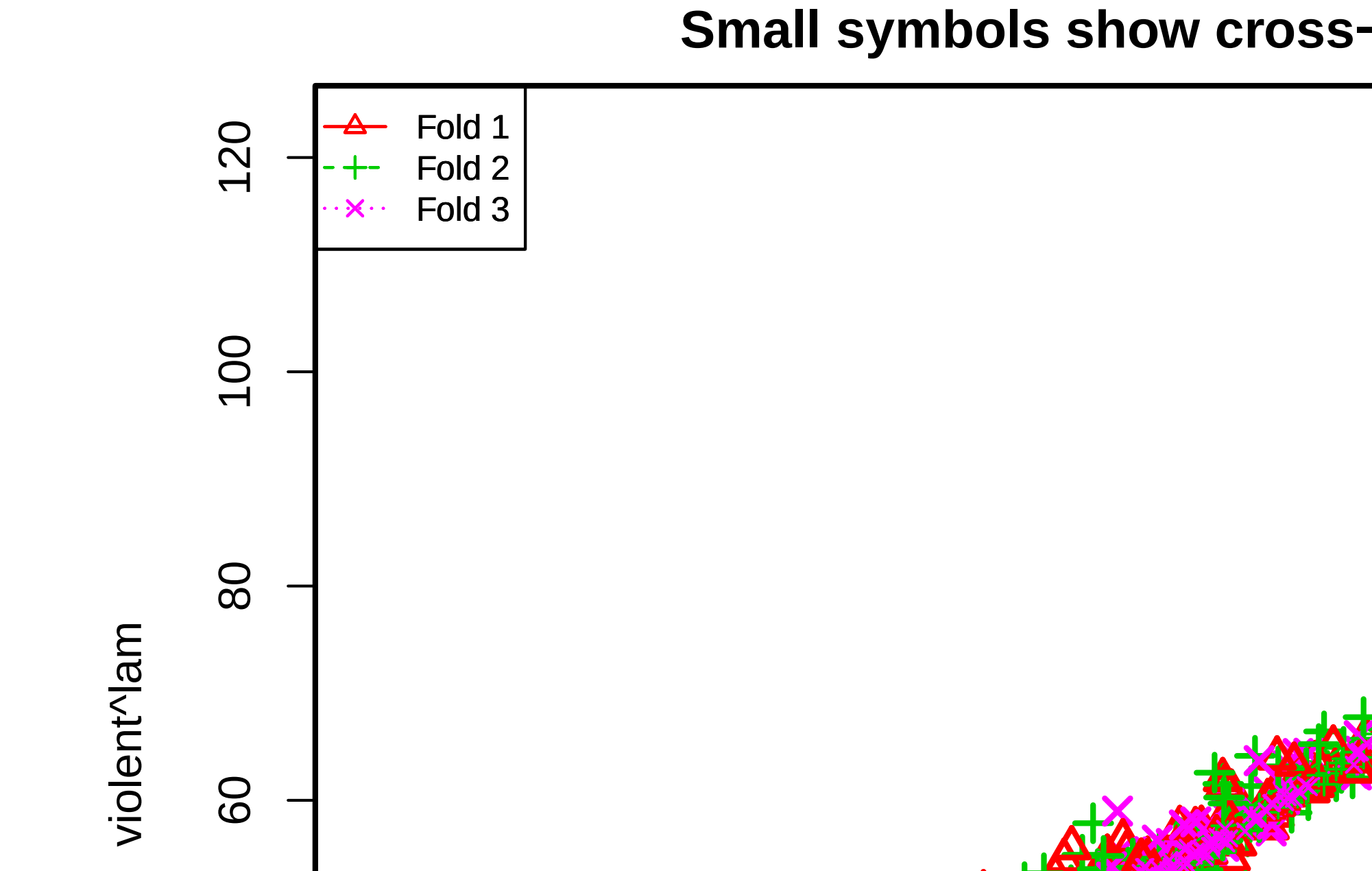
Transformation Check



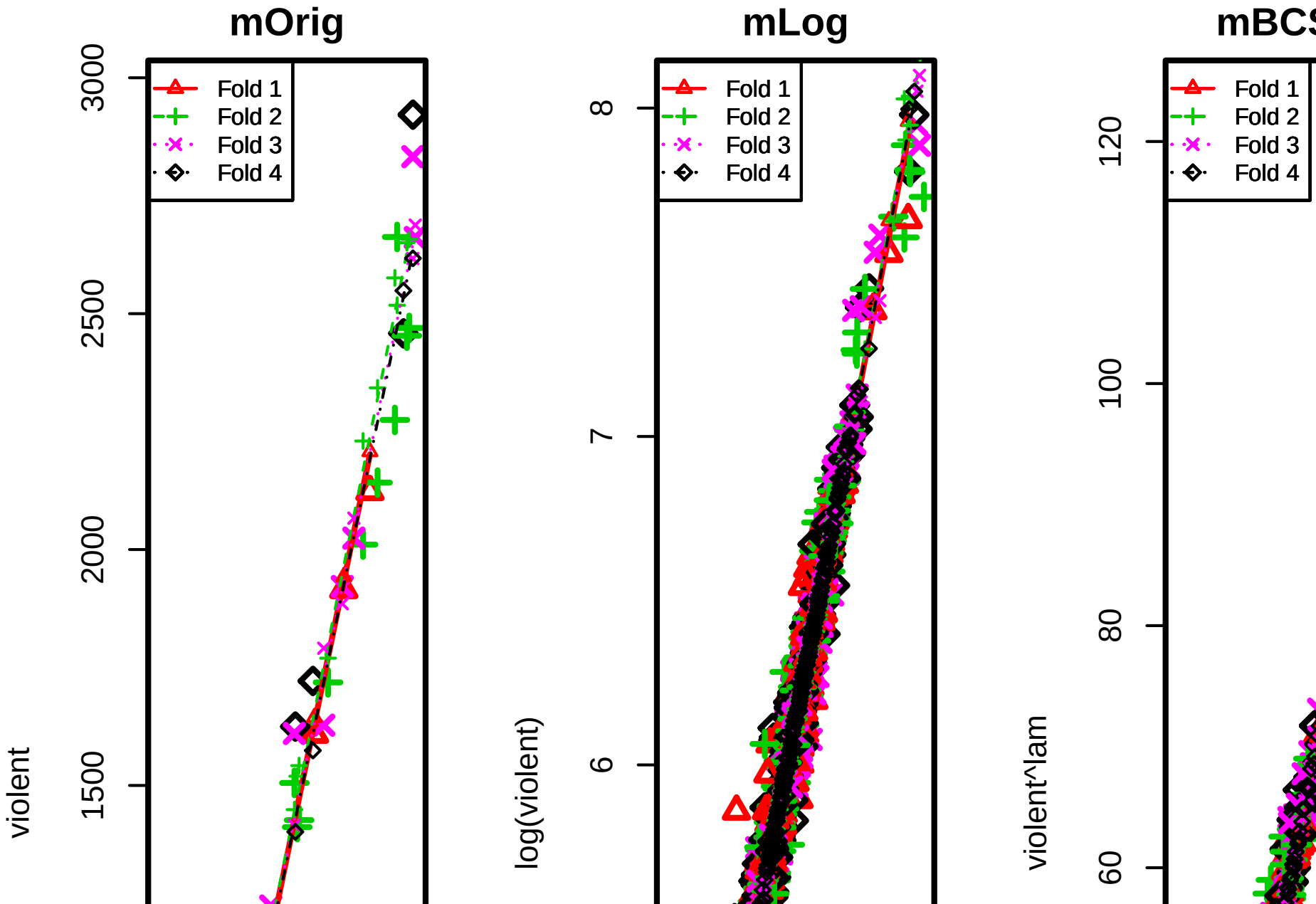
Stepwise regression with cauc normalized with log

```
## Start:  AIC=9240.41
## violent ~ year + murder + robbery + prisoners + afam + l
##     male + population + income + density + state + law
##
##                Df Sum of Sq      RSS     AIC
## - density       1       236  2685757  9238.5
## - log(cauc)     1       335  2685857  9238.6
## <none>                       2685522  9240.4
## - income        1      5124  2690646  9240.6
## - afam          1      6355  2691877  9241.2
## - male          1     18702  2704223  9246.5
## - population    1     24164  2709685  9248.9
## - law           1     26090  2711611  9249.7
## - prisoners     1    177230  2862751  9313.4
## - murder        1    332563  3018084  9375.3
## - year         22    502422  3187943  9397.6
## - robbery       1   3052545  5738066 10129.0
## - state        50   8884729 11570250 10853.6
```

Cross-Validation for Linear Models



Cross-Validation for Linear Models against 10 seeds each



Conclusion

Based on the cross-validation model and the Coefficients comparison above we determine that the Log transformation optimized by the Stepwise function is the optimal model