Linear Models:
Modeling Crime Rates in 1960s America

MA 576

Wanchen Hong

Abstract

There are many factors that can affect crime rates in a certain area, and many are interested in the effect of punishment on crime rates (Vandaele). Identifying the factors that affect crime rates the most could help us find and fund the right policies against crime to protect citizens. In this project, I used US crime data from the 1960s to explore various factors. I started with a linear regression model and attempted various GLM models, trying to find a model that follows the model assumptions the best. I also fitted a GAM model trying to better capture the variance of the dataset. Overall, the Linear Regression Model shows the best predictability and the Poisson GLM model follows the model assumptions the best. The result of both models shows that punishments, regardless of their length, could help reduce crime rates.

1. Introduction

Crime activity has been an important issue in modern society. People want to live in a place that is safe for their own benefit. Generally, highly developed countries and cities where many people are satisfied living in, have a lower crime rate. Therefore, discovering and studying important factors that can affect the crime rate could help us understand what is needed to improve. In a study by Ehrich in 1973, researchers explored the effect of punishment regimes on crime rates and collected relevant data from individuals in different states over the United States. Today, I aim to revisit the dataset and find the most important factors that may affect the crime rate and build the best-fitting models for this dataset by separating the data into training and validation sets and comparing their results.

2. Data Description and Exploration

The selected dataset covers the USA crime data in 1960, which was originally collected by Ehrlich from the Uniform Crime Report of the FBI and other US government sources. The study includes 16 variables, including *Crime*, the variable of interest, and 15 other possible predictors. They are:
M: percentage of males aged 14–24 in the total state population
So: indicator variable for a Southern state
Ed: mean years of schooling of the population aged 25 years or over
Po1: per capita expenditure on police protection in 1960
*Po*2: per capita expenditure on police protection in 1959
LF: labor force participation rate of civilian urban males in the age group 14-24
MF: number of males per 100 females
Pop: state population in 1960 in hundred thousand
NW: percentage of nonwhites in the population
U1: unemployment rate of urban males 14–24
U2: unemployment rate of urban males 35–39
Wealth: wealth, median value of transferable assets or family income
Ineq: income inequality, percentage of families earning below half the median income
Prob: probability of imprisonment: ratio of number of commitments to number of offenses
Time: average time in months served by offenders in state prisons before their first release
Crime: crime rate, number of offenses per 100,000 population in 1960
All these variables except for *So* are continuous numerical variables. *So* is a binary variable.

I used a scatterplot matrix and correlation matrix (Fig. S1 - S3) first to explore all the variables in the dataset. Some of the variables show a strong dependence (with correlation coefficient >0.7), which are: *So* and *Ed, So* and *NW, So* and *Ineq, Ed* and *Ineq, Po1* and *Po2,*

Po1 and *Wealth, Po2* and *Wealth, U1* and *U2,* and *Wealth* and *Ineq. Po1* and *Po2* have a correlation coefficient of over 0.99, therefore, I decided to only keep *Po1* in the dataset.

3. GLM Model Selection

From the scatterplot matrix (Fig. S1 - S3), all variables, when compared to *Crime,* do not show a clear trend. However, some of the variables do show a vague linear relationship to *Crime*, and some of them show a vague quadratic or log relationship.
The response variable of interest, the crime rate *Crime*, is the number of offenses per 100,000 population, which can be deemed as count data to fit a Poisson GLM model, or put into categories of high/low crime rate and fit a Binomial GLM model.

As the first step, the dataset was separated into training and validation with 20% being validation. I attempted to fit a simple linear model and a Poisson GLM due to the linear pattern shown in the scatter plot and the count nature of the data. For both initial models, I used stepwise regression to find the best parameters to include in the model and verified if the model assumptions hold for them.

For the simple linear regression model, the residuals show a general downward pattern with some potential y-space outliers (Fig.1). The QQ plot (Fig.S4) also shows this non-normality. Then I used studentized residuals to remove y-space outliers in this model to assess it. The resulting model has a similar downward pattern in the residuals (Fig.2), but much more normal in the QQ plot (Fig S5). Compared to the model outputs (Fig. S6), the one with outliers removed shows much better adjusted R^2 values (0.83 vs. 0.77) , therefore, I decided to remove the outliers from the data.

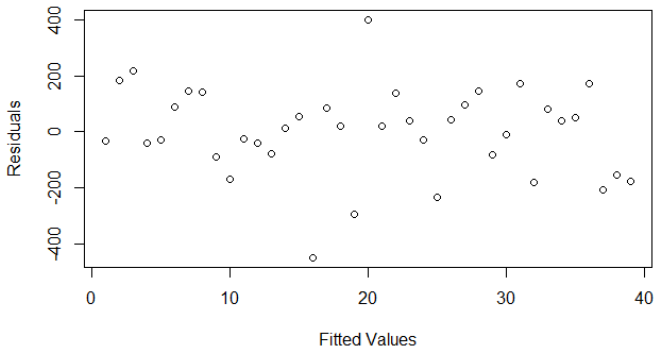

Figure 1. Residual plot of Simple Linear Regression Model.

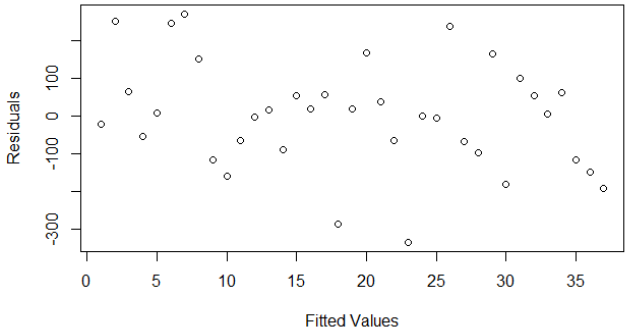

Figure 2. Residual plot of Simple Linear Regression Model with Outliers Removed.

Next for the Poisson GLM with canonical log link, I also fitted a model using stepwise regression. The residuals of this model (Fig. 3) seem slightly more random, which shows a generally better fit. However, there is still a vague downward pattern in the residual plot, with higher values seeming to have a lower variance than higher values, and the residuals are not normal still according to the QQplot (Fig. S7). This could be caused by overdispersion in the data, in which the chi-sq test result is highly significant (Fig. S8). Therefore, I refitted the model using the true dispersion, and the model output eliminates half of the predictors (Fig. S9). Then a reduced Poisson GLM is refitted using only the significant predictors. The resulting residual plot is much more random (Fig. 4), and the QQ plot result also aligns with it (Fig. S10), yet the chi-sq test on deviance residuals shows that the full model is better (Fig. S11). This contradiction may suggest that a Poisson GLM might not be the best model to capture all the relationships in the data set.

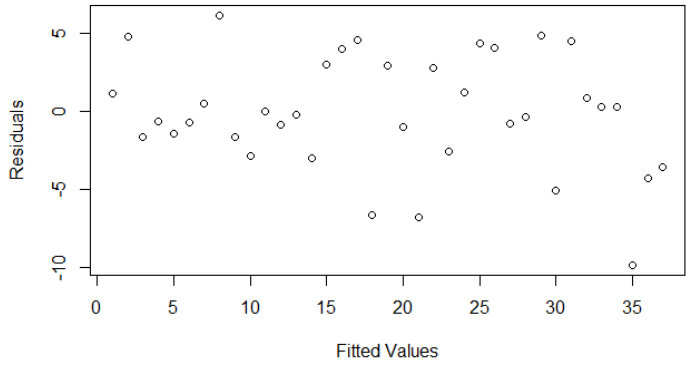

Figure 3. Residual plot of Poisson GLM with Canonical Link.

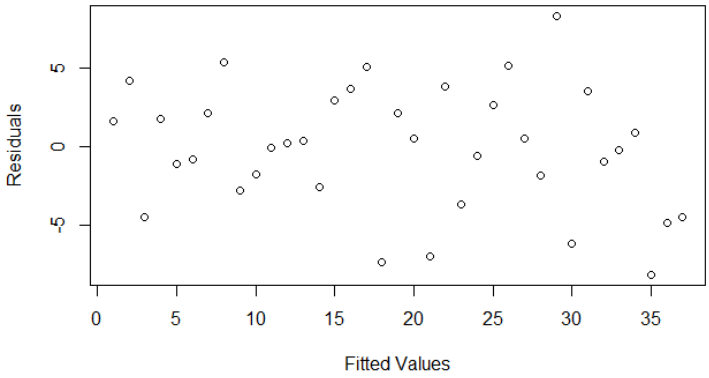

Figure 4. Residual plot of Reduced Poisson GLM.

However, there is still a vague difference in variance at different fitted values (Fig. 4), especially at the two ends, and the QQ plot (Fig. S10) confirms it too. Gamma GLMs with an inverse link and quasi GLMs with variance function equal to μ^2 and with inverse and log link were attempted to address this difference in variance. However, the diagnostic plots show that none of these models can address this issue perfectly (Fig. S12 - S14), all the models' performances are similar to the Poisson GLM above, not capturing the changing variance accurately, and the residuals are not randomly distributed. Therefore, the overall best GLM model is the Reduced Poisson GLM.

4. Other Models: GAM

Another way to model this changing variance in predictors is by using the generalized additive model (GAM), which allows each predictor to have a different nonlinear relationship to the response variable. Therefore, I fitted a Poisson GAM. The output of the model is much more significant than the GLM models with a much higher adjusted R^2 value (0.947) than all previous models (Fig. 5). However, with an adjusted R^2 value this high, it could be a sign of overfitting.

The diagnostic plots show that the residuals are approximately random, with some potential outliers. It seems that the difference in variance is captured well using this model.

```
Parametric coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  6.62898    0.01283  516.87  < 2e-16 ***
So           0.22168    0.02932    7.56 4.04e-14 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Approximate significance of smooth terms:
           edf Ref.df Chi.sq p-value
s(M)     3.467  3.836 258.72  <2e-16 ***
s(Ed)    1.000  1.000 160.43  <2e-16 ***
s(Po1)   1.000  1.000 715.99  <2e-16 ***
s(U2)    3.805  3.960 199.09  <2e-16 ***
s(Ineq)  3.995  4.000 304.79  <2e-16 ***
s(Prob)  3.424  3.796  70.15  <2e-16 ***
s(Time)  3.920  3.991 182.15  <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

R-sq.(adj) =  0.947   Deviance explained = 97.3%
UBRE = 4.0584  Scale est. = 1         n = 37
```

Figure 5. GAM model output.

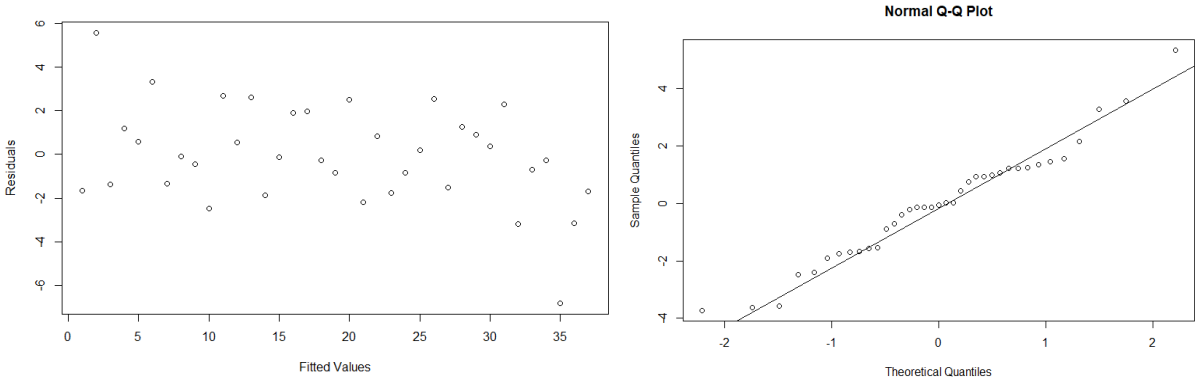

Figure 6. Residual plot and QQ plot of the GAM model.

5. Model Performance

To assess the model performance of the best GLM models (the full and reduced Poisson GLM) and GAM model, I will use them to predict the value of *Crime* in the validation set and evaluate their performance using root mean squared error (RMSE) and mean absolute percentage error (MAPE), and compare the result with a simple linear regression model.

Table 1. Performance of the Poisson GLM and GAM models, Compared to Linear Regression

	Linear Regression	Reduced GLM	Full GLM	Poisson GAM
RMSE	286.0063	346.8865	351.271	302.7024
MAPE	0.1927093	0.2264039	0.2033462	0.2606585

The result shows that the Poisson GLM model performs similarly to the GAM model. This makes sense because the GAM looks almost perfectly random, which implies that there is a potential overfit. The average ~20% error for the full GLM model is still very high for the prediction. At the same time, all advanced models (GLM and GAM) are performing worse than the simple linear regression model even though they follow the model assumptions more closely. This is very likely because the entire dataset is very small, with only 47 observations in total, and only 8 of them are in the validation set, which would very likely cause the validation set to be biased and the model to be inaccurate. Therefore, my best-fit model to predict crime rate using this dataset would be a linear regression model, with the model being (Fig. S6):

$$Crime \sim Po1 + Ineq + Ed + M + Prob + U2$$

In other words, the crime rates would be most affected by *Ed*: average education level, *Po1*: per capita expenditure on police protection, *M*: percentage of males in the total state population, *U2*: the unemployment rate of urban males 35–39, *Ineq*: income inequality, *Prob*: the probability of imprisonment. From this result, we can see that a higher imprisonment probability lowers the crime rate in the state significantly (coefficient = -2441.88), meaning that punishment may be able to help lower the crime rate, and since *Time,* the average time of imprisonment, does not significantly affect the crime rate, we may say that any form of punishment would be able to help with crime rate, to warn the citizens from potential wrongdoings. The results from the Poisson GLM agree with this finding (Fig. S9): the coefficient of *Prob* also has a negative value.

6. Conclusion

The simple linear regression model has the best predictability out of all the models I fitted, yet does not follow the model assumptions well; the GAM model shows the best fit to the model assumptions, but it is very likely overfitted. The models in the middle ground, the Poisson GLM models, do not have the same level of accuracy as the linear regression model, and they also do not follow the model assumptions perfectly like the GAM model. It is a dilemma to find the best model for this dataset, but the result shows one important finding, which also coincides with the interest of Ehrlich and other criminologists, who collected the data. This finding is that crime rates are significantly affected by the probability of punishment given. The higher the probability of punishment, perhaps the higher the probability of people being intimidated by the consequence of committing a crime, and consequently a lower crime rate.

To better model this dataset, there are multiple ways: 1) Gather more data so the training and validation datasets would be larger, and in turn more accurately reflect the fit of the models; 2) Change the response variable and/or the predictor for alternative models, for example, change *Crime* into high/medium/low levels and fit a binomial model instead; 3) Find other

alternative models, like different smoothing function in GAM, or a totally different model like neural networks. These alternative models may be more difficult to explain the fit, but they would fit the dataset with better performance, so one would be able to predict crime rate using other parameters more accurately.

Reference

Ehrlich, I. (1973) Participation in illegitimate activities: a theoretical and empirical investigation. Journal of Political Economy 81, 521–565.
Vandaele, W. (1978) Participation in illegitimate activities: Ehrlich revisited. In Deterrence and Incapacitation, eds A. Blumstein, J. Cohen and D. Nagin, National Academy of Sciences, Washington DC, pp. 270–335.
Venables, W., and Ripley, B. (1998). Modern Applied Statistics with S-Plus, Second Edition. Springer-Verlag.

Appendix

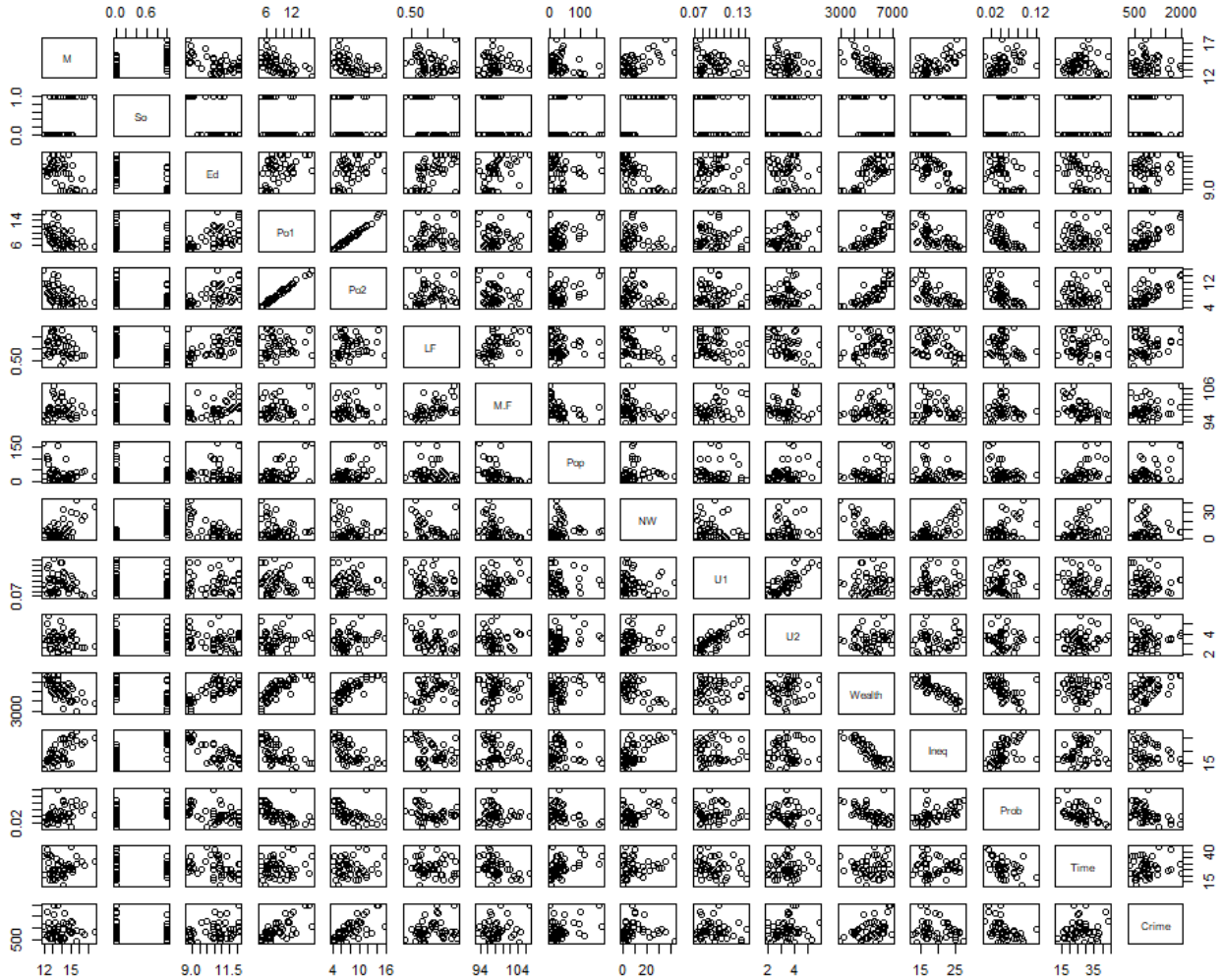

Figure S1. Scatterplot Matrix of all variables

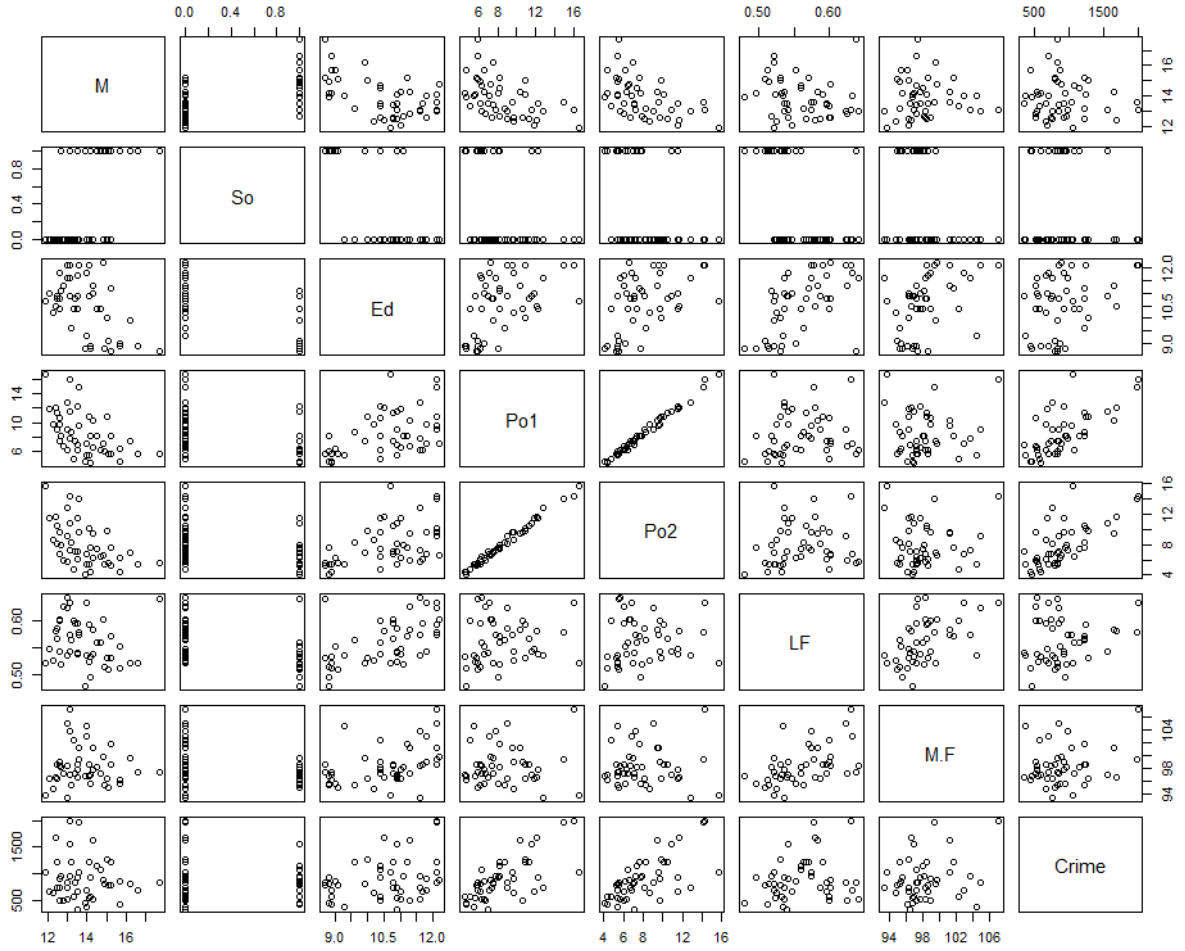

Figure S2. Scatterplot Matrix of *Crime* and first 7 variables: *M*, *So*, *Ed*, *Po1*, *Po*2, *LF* and *M.F*

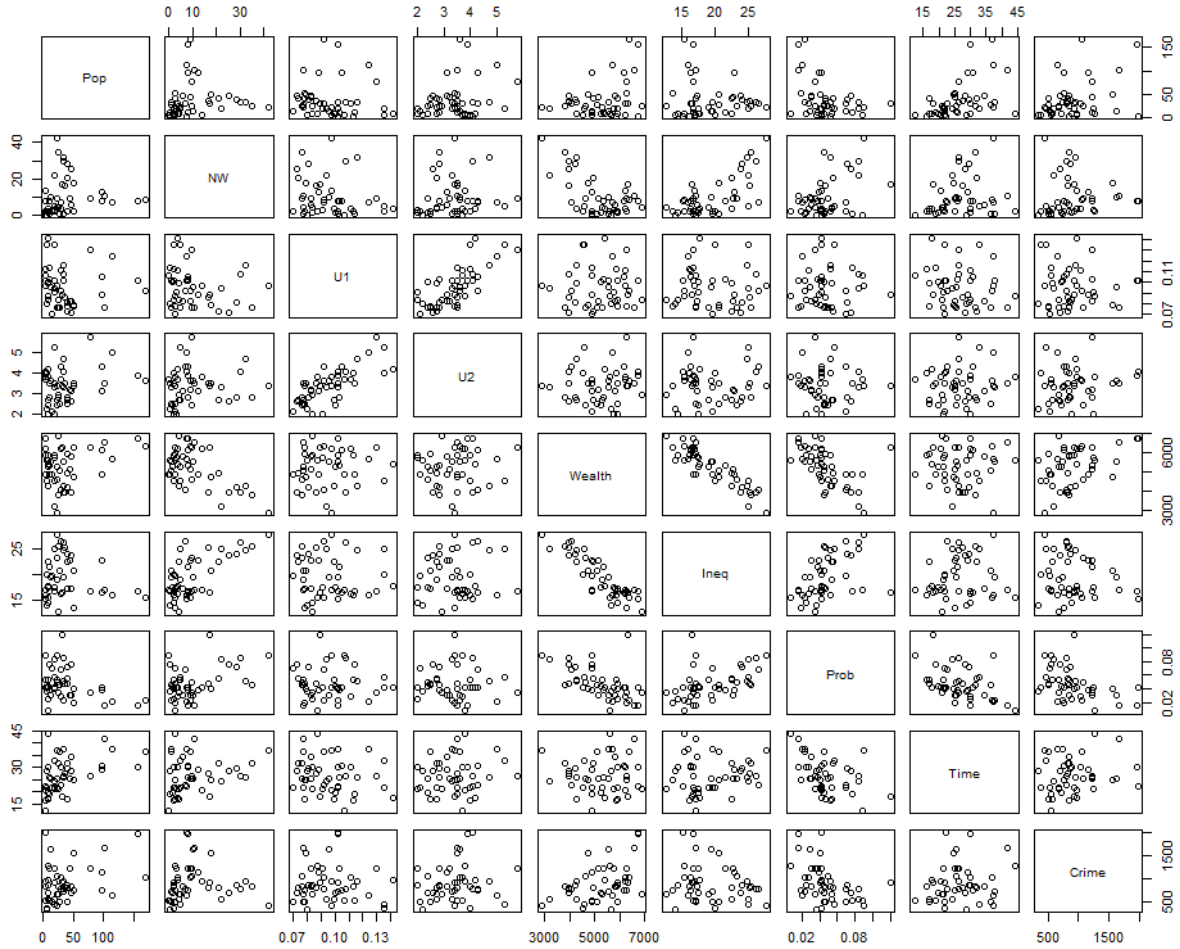

Figure S3. Scatterplot Matrix of *Crime* and last 8 variables: *Pop*, *NW*, *U1*, *U2*, *Wealth*, *Ineq*, *Prob* and *Time*

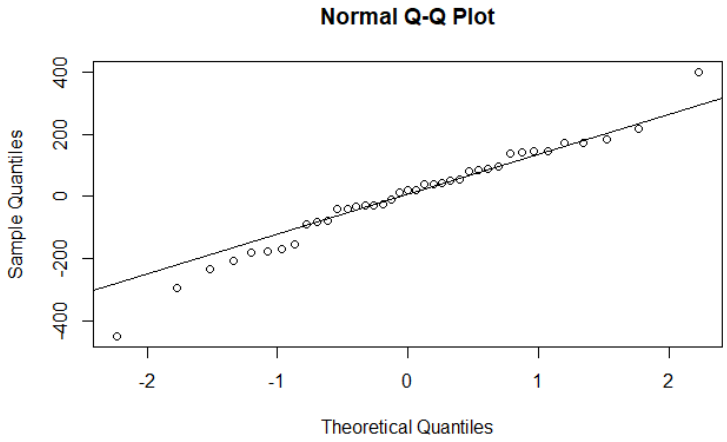

Figure S4. QQ plot of Simple Linear Regression Model

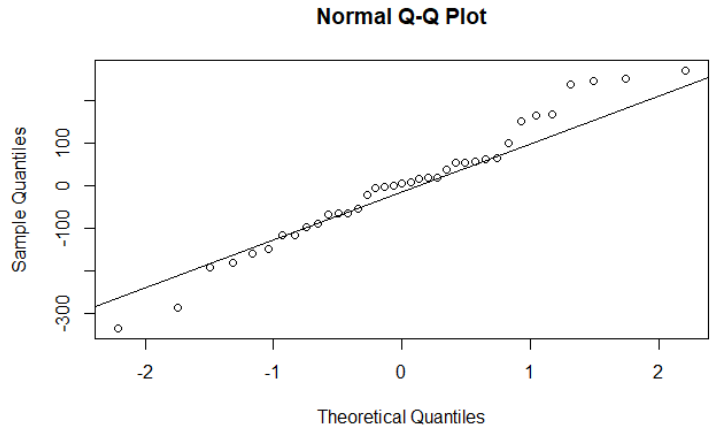

Figure S5. QQ plot of Simple Linear Regression Model with Outliers Removed

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -5480.9052   901.5130  -6.080 1.12e-06 ***
Po1           116.3456    15.5364   7.489 2.39e-08 ***
Ineq           63.5674    14.2207   4.470 0.000104 ***
Ed            228.4524    47.2426   4.836 3.70e-05 ***
M             112.8504    33.3334   3.385 0.001998 **
U2            106.3220    38.5863   2.755 0.009869 **
Prob        -3949.2111  1588.7328  -2.486 0.018724 *
Pop            -1.3712     0.9666  -1.419 0.166337
So            131.4225   103.4077   1.271 0.213526
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 178.5 on 30 degrees of freedom
Multiple R-squared:  0.8194,    Adjusted R-squared:  0.7712
F-statistic: 17.01 on 8 and 30 DF,  p-value: 3.324e-09
```

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -6060.29     800.11  -7.574 1.91e-08 ***
Po1           110.98      11.34   9.784 7.59e-11 ***
Ineq           57.61      11.56   4.984 2.44e-05 ***
Ed            236.28      39.38   6.000 1.39e-06 ***
M             151.62      29.33   5.169 1.45e-05 ***
Prob        -2441.88    1287.64  -1.896  0.06757 .
U2            119.82      34.24   3.499  0.00148 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 155.7 on 30 degrees of freedom
Multiple R-squared:  0.8593,    Adjusted R-squared:  0.8311
F-statistic: 30.52 on 6 and 30 DF,  p-value: 1.727e-11
```

Figure S6. Left: LM output with all data vs. Right: LM output with outliers removed.

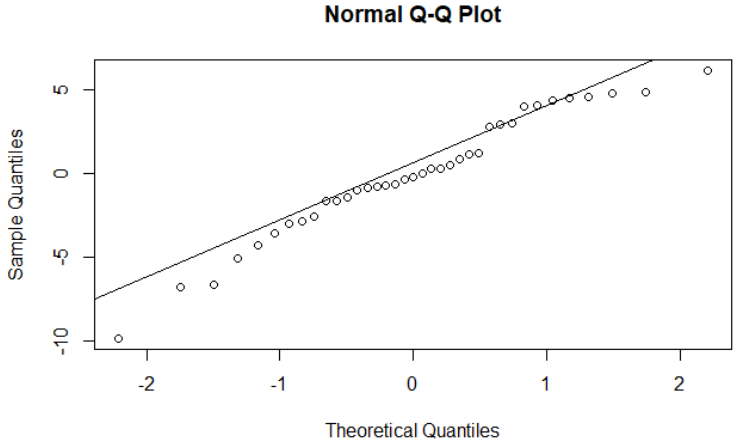

Figure S7. QQ plot of Best Poisson GLM with Canonical Link

```
> 1-pchisq(sigma2*pois_step1$df.residual,pois_step1$df.residual)
[1] 0
```

Figure S8. Chi-squared Test for Overdispersion, p << 0.05

```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  1.0876855  0.2779606   3.913 9.11e-05 ***
M            0.1664306  0.0075859  21.940  < 2e-16 ***
So           0.3337658  0.0224812  14.846  < 2e-16 ***
Ed           0.2694327  0.0125064  21.544  < 2e-16 ***
Po1          0.1116881  0.0032503  34.363  < 2e-16 ***
MF          -0.0150163  0.0034735  -4.323 1.54e-05 ***
Pop         -0.0014271  0.0002001  -7.131 9.94e-13 ***
NW          -0.0033217  0.0010086  -3.293  0.00099 ***
U2           0.1485963  0.0086209  17.237  < 2e-16 ***
Ineq         0.0464790  0.0034552  13.452  < 2e-16 ***
Prob        -5.9041932  0.4018928 -14.691  < 2e-16 ***
Time        -0.0057682  0.0013534  -4.262 2.03e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 5310.78  on 36  degrees of freedom
Residual deviance:  484.09  on 25  degrees of freedom
AIC: 824.34
```

```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  1.0876855  1.2069498   0.901 0.367490
M            0.1664306  0.0329391   5.053 4.36e-07 ***
So           0.3337658  0.0976169   3.419 0.000628 ***
Ed           0.2694327  0.0543046   4.962 6.99e-07 ***
Po1          0.1116881  0.0141133   7.914 2.50e-15 ***
MF          -0.0150163  0.0150824  -0.996 0.319438
Pop         -0.0014271  0.0008689  -1.642 0.100518
NW          -0.0033217  0.0043795  -0.758 0.448177
U2           0.1485963  0.0374334   3.970 7.20e-05 ***
Ineq         0.0464790  0.0150029   3.098 0.001948 **
Prob        -5.9041932  1.7450834  -3.383 0.000716 ***
Time        -0.0057682  0.0058766  -0.982 0.326319
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 18.85436)

    Null deviance: 5310.78  on 36  degrees of freedom
Residual deviance:  484.09  on 25  degrees of freedom
AIC: 824.34
```

Figure S9. Left: Poisson GLM output assuming dispersion = 1 vs. Right: Poisson GLM output with true dispersion.

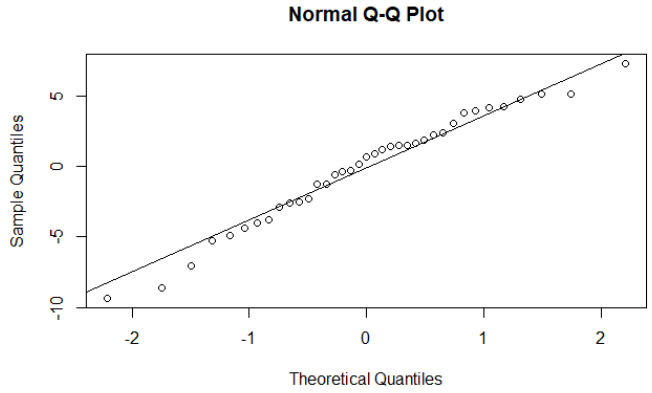

Figure S10. QQ plot of Reduced Poisson GLM.

```
Analysis of Deviance Table

Model 1: Crime ~ M + So + Ed + Po1 + U2 + Ineq + Prob
Model 2: Crime ~ M + So + Ed + Po1 + MF + Pop + NW + U2 + Ineq + Prob +
    Time
  Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
1        29     574.71
2        25     484.09  4   90.617 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure S11. Chi-Squared Test Result between Full and Reduced Poisson GLM. The full model is a statistically better fit.

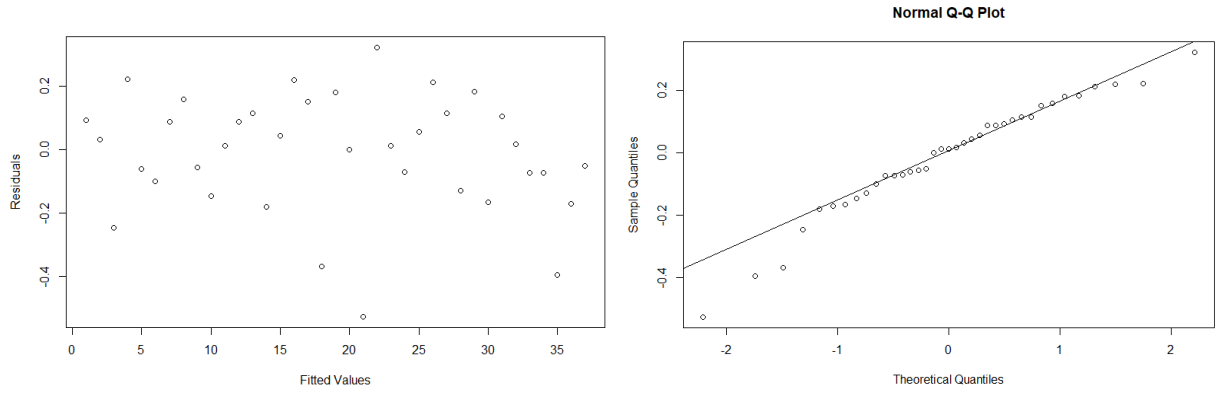

Figure S12. Residual plot and QQ plot of Best Gamma GLM with Canonical Link.

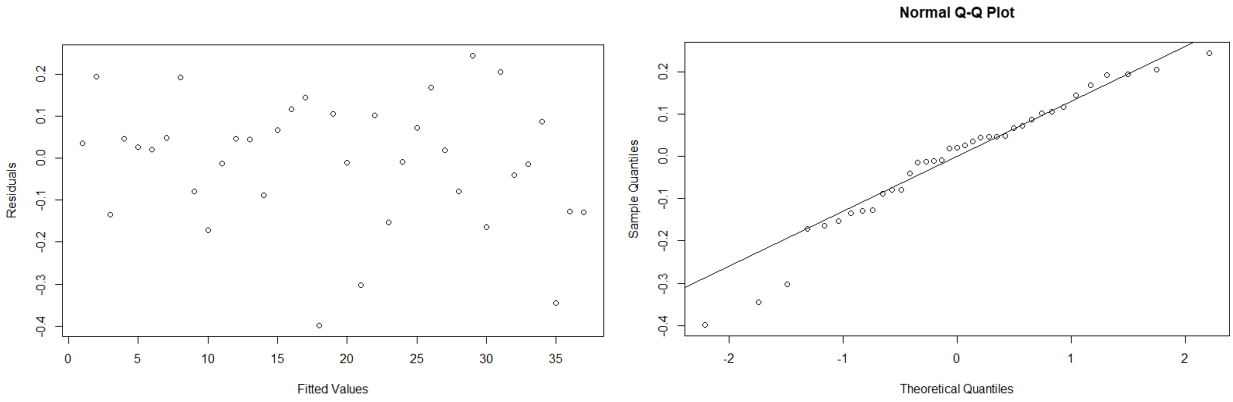

Figure S13. Residual plot and QQ plot of Quasi GLM with Log Link and Square Variance.

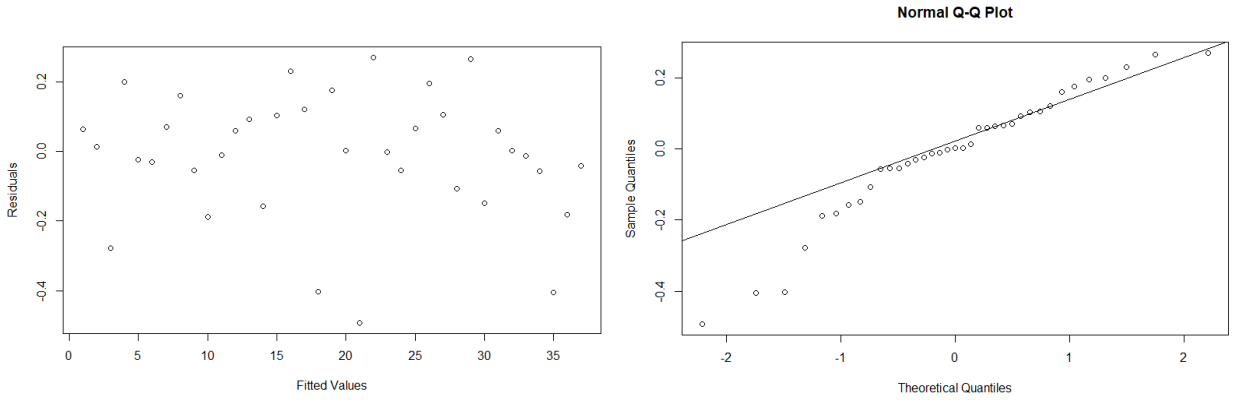

Figure S14. Residual plot and QQ plot of Quasi GLM with Inverse Link and Square Variance.