Regression Analysis of Crime-Related and Demographic Statistics for 47 US States in 1960

STAT 425

11.11.11

Luis S. Lin

Contents

1. Introduction

Purpose

The objective of this report was to analyze a data set using regression analysis to determine what factors influence the response. The data are crime-related and demographic statistics for 47 US states in 1960 collected from the FBI's Uniform Crime Report and other government agencies. The data contains 14 variables (see Table 1.) and 47 cases. The goal of the project was to find a model that describes the relationship between crime rate and the demographical data collected in the study.

Outline

An exploratory analysis was conducted first to determine a proper transformation, and identify outliers and high influential points that could affect the subsequent analysis. Next, the model selection was performed using two criterion-based procedures: Akaike Information Criterion (AIC) and Bayes Information Criterion (BIC). Then, a linear regression diagnosis was done to ensure assumptions were satisfied. Finally, other issues pertaining to the analysis were discussed and interpretations of the models were given. The analysis in this report was conducted with R version 2.13.1. For details on the analysis see attached R code.

Variable Names

Variable	Description
R	Crime rate: # of offenses reported to police per million population
Age	The number of males of age 14-24 per 1000 population
S	Indicator variable for Southern states (0 = No, 1 = Yes)
Ed	Mean # of years of schooling x 10 for persons of age 25 or older
Ex0	1960 per capita expenditure on police by state and local government
Ex1	1959 per capita expenditure on police by state and local government
LF	Labor force participation rate per 1000 civilian urban males age 14-24
M	The number of males per 1000 females
N	State population size in hundred thousands
NW	The number of non-whites per 1000 population
U1	Unemployment rate of urban males per 1000 of age 14-24
U2	Unemployment rate of urban males per 1000 of age 35-39
W	Median value of transferable goods and assets or family income in tens of $
X	The number of families per 1000 earning below 1/2 the median income

Table 1. Variable Names

2. Preliminary Analysis

Finding Proper Transformation

A full model was fit and the Box-Cox method was used to determine whether the response variable, R, needed a transformation. The Log-likelihood plot for the Box-Cox transformation is shown in Figure 1. The 95% confidence interval for the transformation runs from -0.256 to 0.757, and does not include 1, meaning that there is strong reason to transform the response. For interpretation purposes, a natural log transformation was chosen, which is supported by the Box-Cox method since zero is included in the confidence interval . This transformation is used for the rest of the analysis.

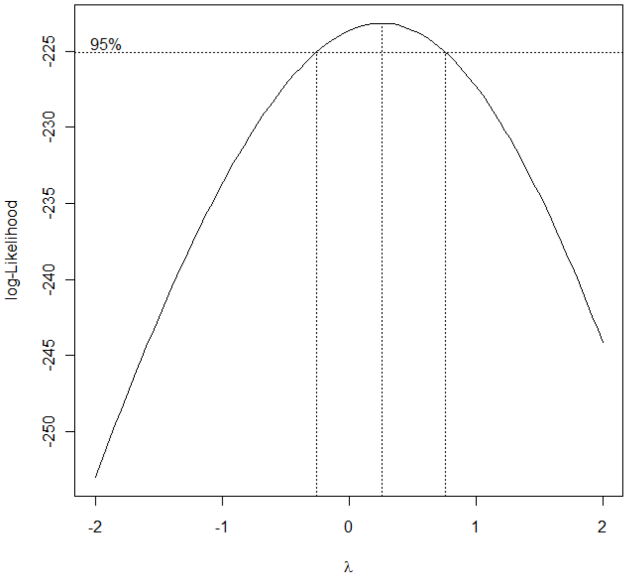


Figure 1. Log-likelihood plot for the Box-Cox transformation

Outlier Test

Variable selection methods are sensitive to outliers and high influential points. Thus an outlier test was conducted by computing the jackknife (externally studentized) residual for all samples with Bonferroni correction. Comparing it with the critical t-value, the test indicates that there were no outliers.

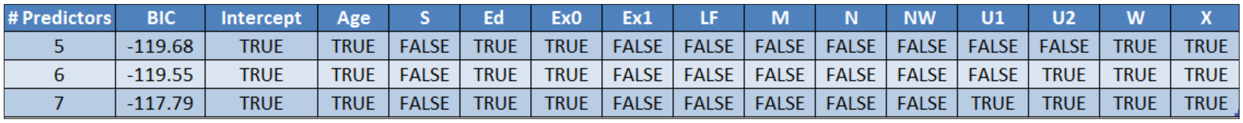
High Influential Points Test

An high influential point test was done by calculating Cook distances. Since all cook distances were less than 1, the test indicates that there were no high influential points

3. Model Selection: BIC

Model Selection

A level-wise searching algorithm was used with BIC. The top three models having the lowest BIC are shown in Table 2. Thus, the BIC criterion suggests using 5 predictors: Age, Ed, Ex0, W, and X. A model with these predictors and log(R) as the response was fitted and a diagnosis was performed.



# Predictors	BIC	Intercept	Age	S	Ed	Ex0	Ex1	LF	M	N	NW	U1	U2	W	X
5	-119.68	TRUE	TRUE	FALSE	TRUE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	TRUE
6	-119.55	TRUE	TRUE	FALSE	TRUE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	TRUE	TRUE
7	-117.79	TRUE	TRUE	FALSE	TRUE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	TRUE	TRUE	TRUE

Table 2. BIC

Check constant variance assumption

Under the assumption, the residuals should vary randomly around zero and the spread of the residuals should be about the same throughout the plot (no systematic patterns, scatter should be symmetric vertically about zero). Figure 2. shows a plot of residuals against the fitted values, with no sign of non-constant variance detected. A quick test was done by regressing the absolute of residual against the fitted values. The p-value for the overall F-test 0.913, indicating no problem with a variance increasing or decreasing relationship (note: this test is not quite right as some weighting should be used and the degrees of freedom should be adjusted). Thus, the assumption of constant variance is reasonable.

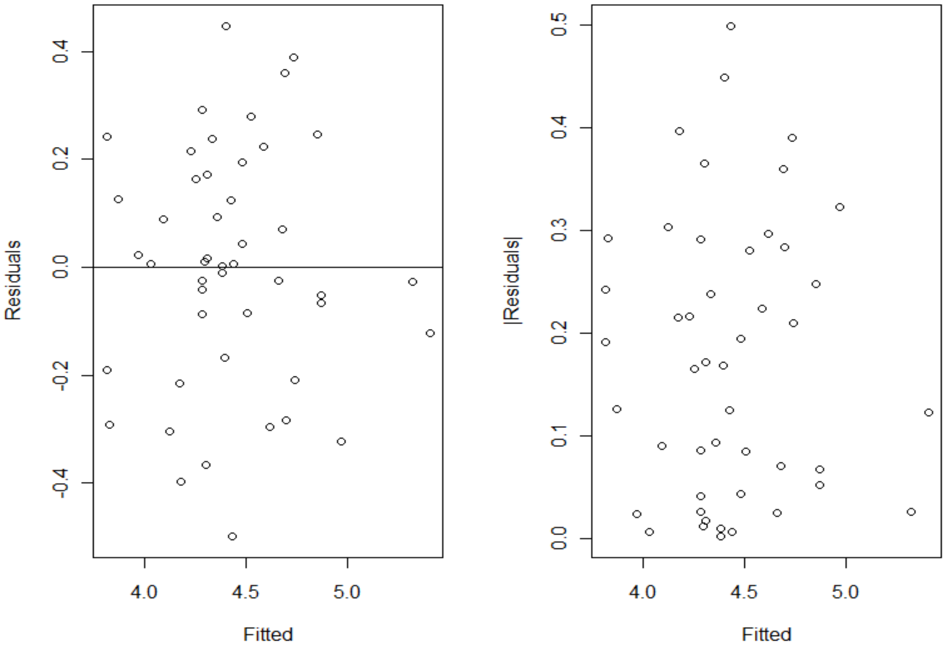


Figure 2: Residuals vs Fitted

Check normality assumption

Under the assumption, errors should be iid normally distributed. The residuals can be assessed for normality using a Q–Q plot, shown in Figure3. This compares the residuals to "ideal" normal observation. Normal residuals should follow the line approximately. The residuals look normal since the residuals don't deviate much from a straight line (for the studentized residuals, they should lie along a 45 degree line since they have been normalized).

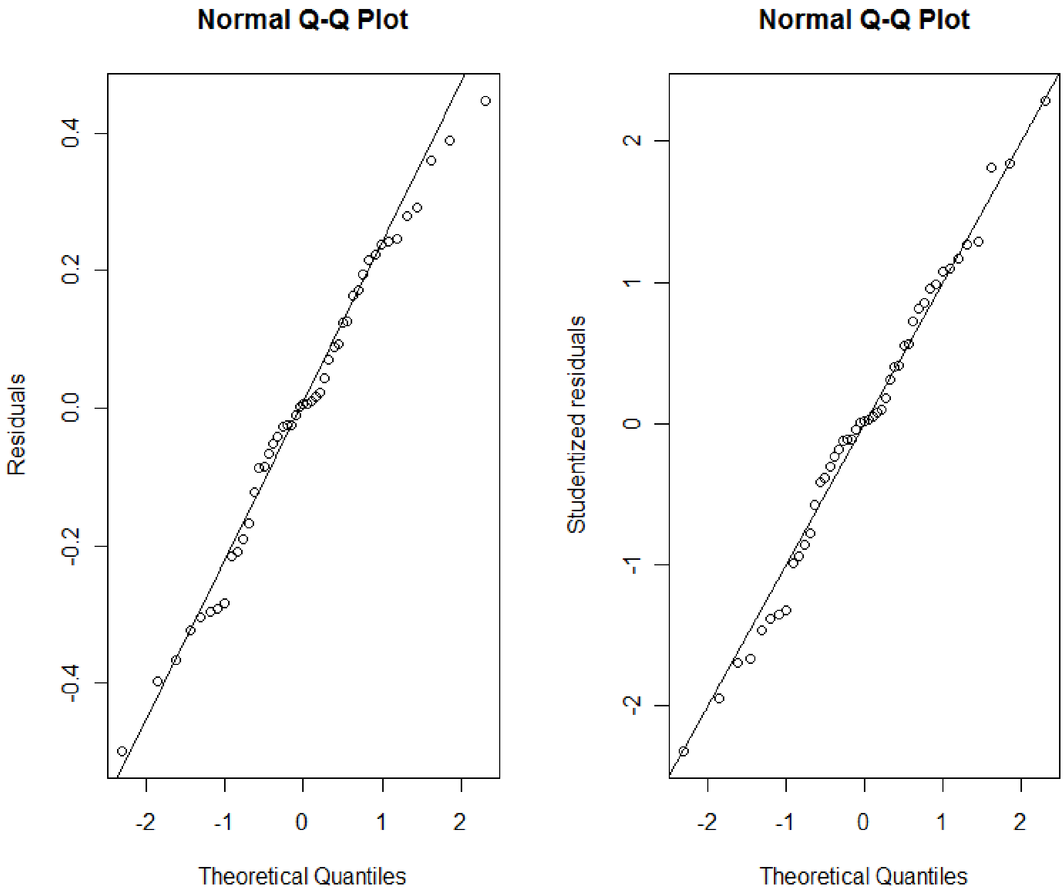


Figure 3: Q-Q Plots

The Shapiro-Wilk test is a formal test for normality. In this case, the p-value (0.791) is large (greater than 0.05), so the null (the residuals are normal) is not rejected. In conclusion, the assumption of normality (observed samples come from normal distribution) is reasonable.

Check for large leverage points

A leverage point can be defined as an "unusual point in the predictor space—it has the potential to influence the fit", which identifies with unusual or outlying x-vales. Leverages greater than 2p/n should be examined. The leverage for case 29 is greater than this critical value, indicating that case 29 is a large leverage point. This is illustrated in the plots in Figure 4.

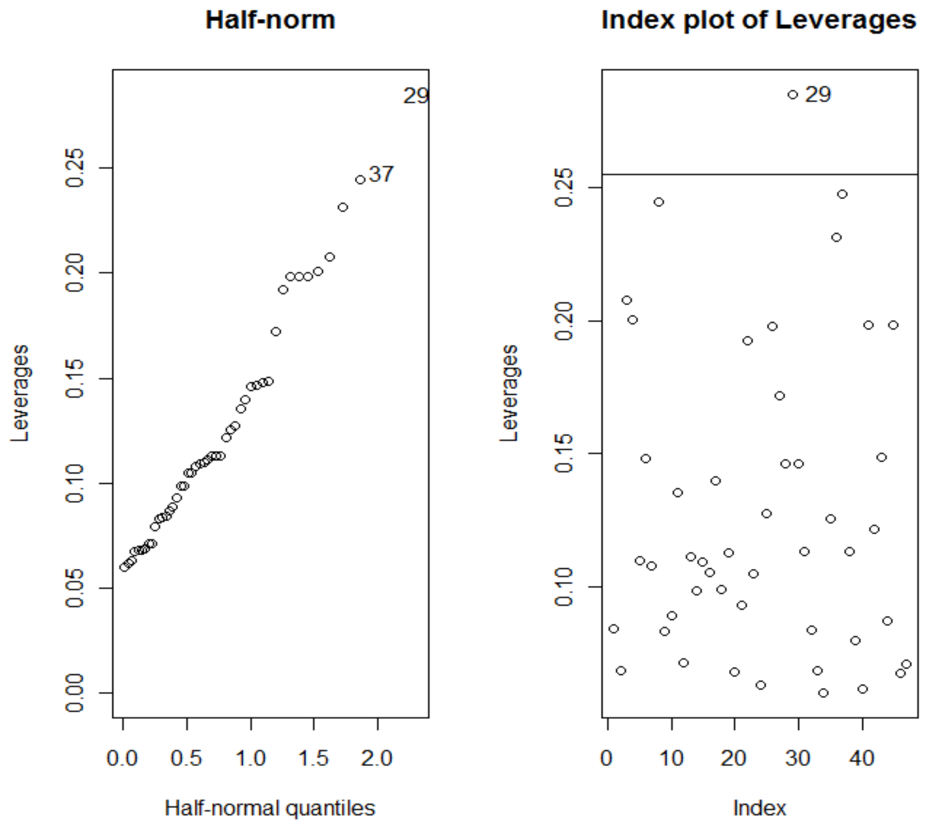


Figure 4: Half-Norm and Leverage Plots

Check for outliers and high influential points

Similar tests as in the preliminary analysis were conducted, resulting in neither outliers nor high influential points detected.

Check Correlated Errors

To check the assumption of uncorrelated errors, a Durbin-Watson test was done. The p-value of the test was 0.182, indicating we cannon reject the null hypothesis that errors are uncorrelated (assumption of uncorrelated errors follows a linear combination of χ^2). Thus, there was no evidence of correlation, so assumption of uncorrelated error holds.

Check the structure of the relationship between the predictors and the response

To detect non-linearity in the model, partial residuals plots were constructed for each predictor. Figure 5. shows these plots and it can be seen that the linear assumption is reasonable, that is the systematic part (Ey=Xb) of the model is correct. Thus, there was no indication of non-linear relationship between response variable and explanatory variables, meaning the linear assumption holds.

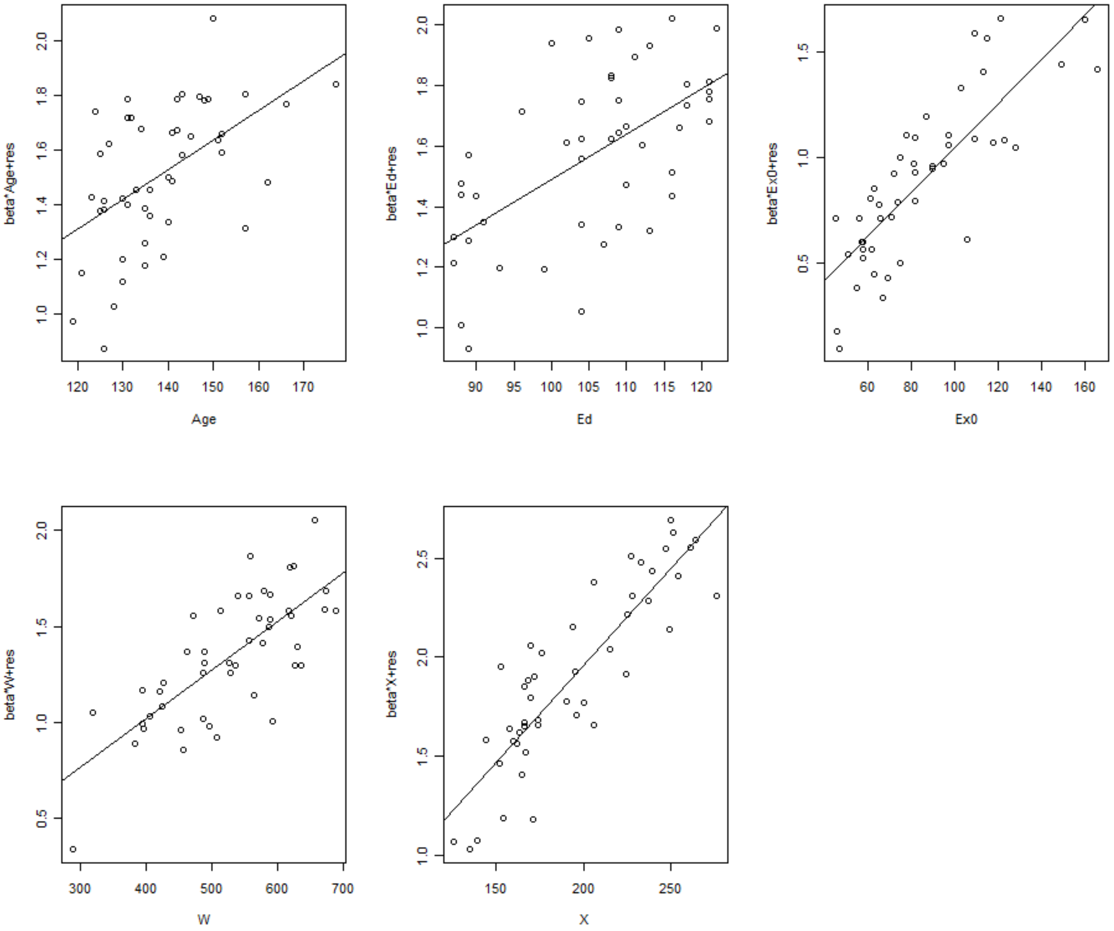


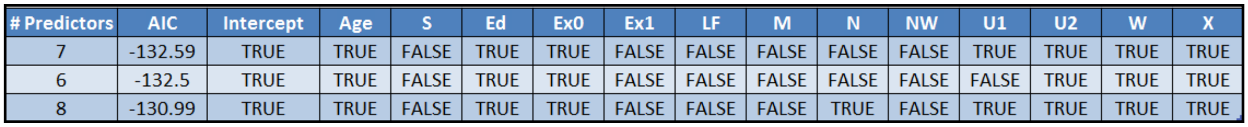
Figure 5: Partial Residual Plots

The diagnosis indicate all assumptions are reasonable and no adjustments are necessary to the model selected by BIC. High leverage point 29 was removed and a new model was fitted to see any changes.

4. Model Selection: AIC

Model Selection

A level-wise searching algorithm was used with AIC. The top three models having the lowest AIC are shown in Table 3. Thus, the AIC criterion suggest using 7 predictors Age, Ed, Ex0, U1, U2, W, and X. A model with these predictors and log(R) as the response was fitted and a diagnosis was performed.



# Predictors	AIC	Intercept	Age	S	Ed	Ex0	Ex1	LF	M	N	NW	U1	U2	W	X
7	-132.59	TRUE	TRUE	FALSE	TRUE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	TRUE	TRUE	TRUE
6	-132.5	TRUE	TRUE	FALSE	TRUE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	TRUE	TRUE
8	-130.99	TRUE	TRUE	FALSE	TRUE	TRUE	FALSE	FALSE	FALSE	TRUE	FALSE	TRUE	TRUE	TRUE	TRUE

Table 3. AIC

Check constant variance assumption

Under the assumption, the residuals should vary randomly around zero and the spread of the residuals should be about the same throughout the plot (no systematic patterns, scatter should be symmetric vertically about zero). Figure 6. shows a plot of residuals against the fitted values, with no sign of non-constant variance detected. A quick test was done by regressing the absolute of residual against the fitted values. The p-value for the overall F-test 0.986, indicating no problem with a variance increasing or decreasing relationship (note: this test is not quite right as some weighting should be used and the degrees of freedom should be adjusted). Thus, the assumption of constant variance is reasonable.

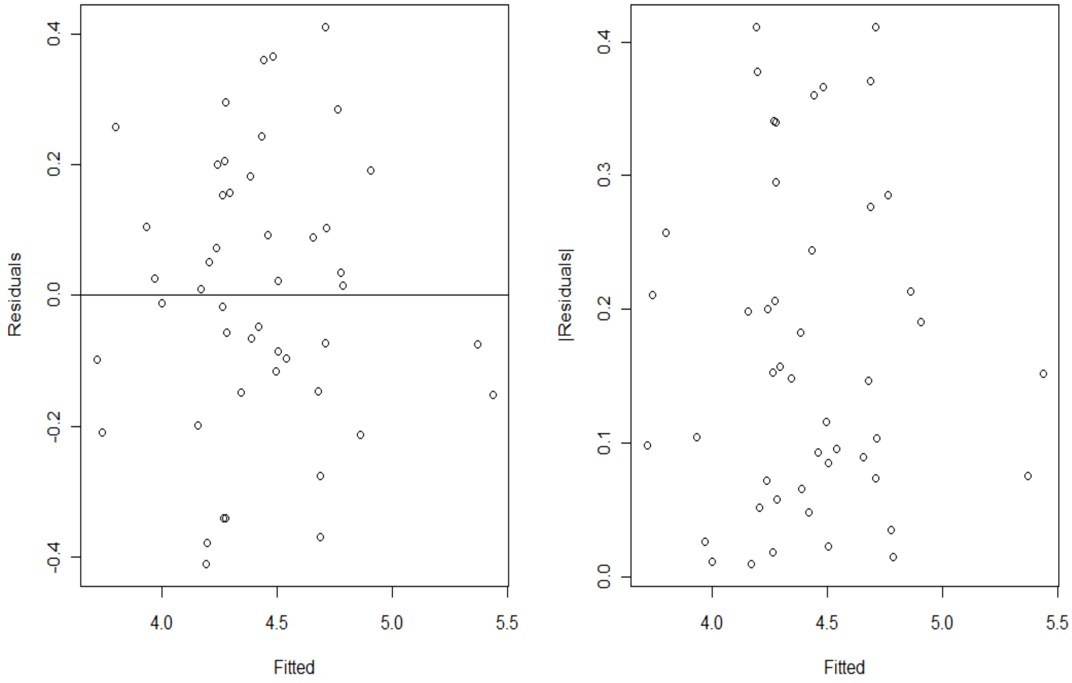


Figure 6: Residuals vs Fitted

Check normality assumption

Under the assumption, errors should be iid normally distributed. The residuals can be assessed for normality using a Q–Q plot, shown in Figure 7. This compares the residuals to "ideal" normal observation. Normal residuals should follow the line approximately. The residuals look normal since the residuals don't deviate much from a straight line (for the studentized residuals, they should lie along a 45 degree line since they have been normalized).

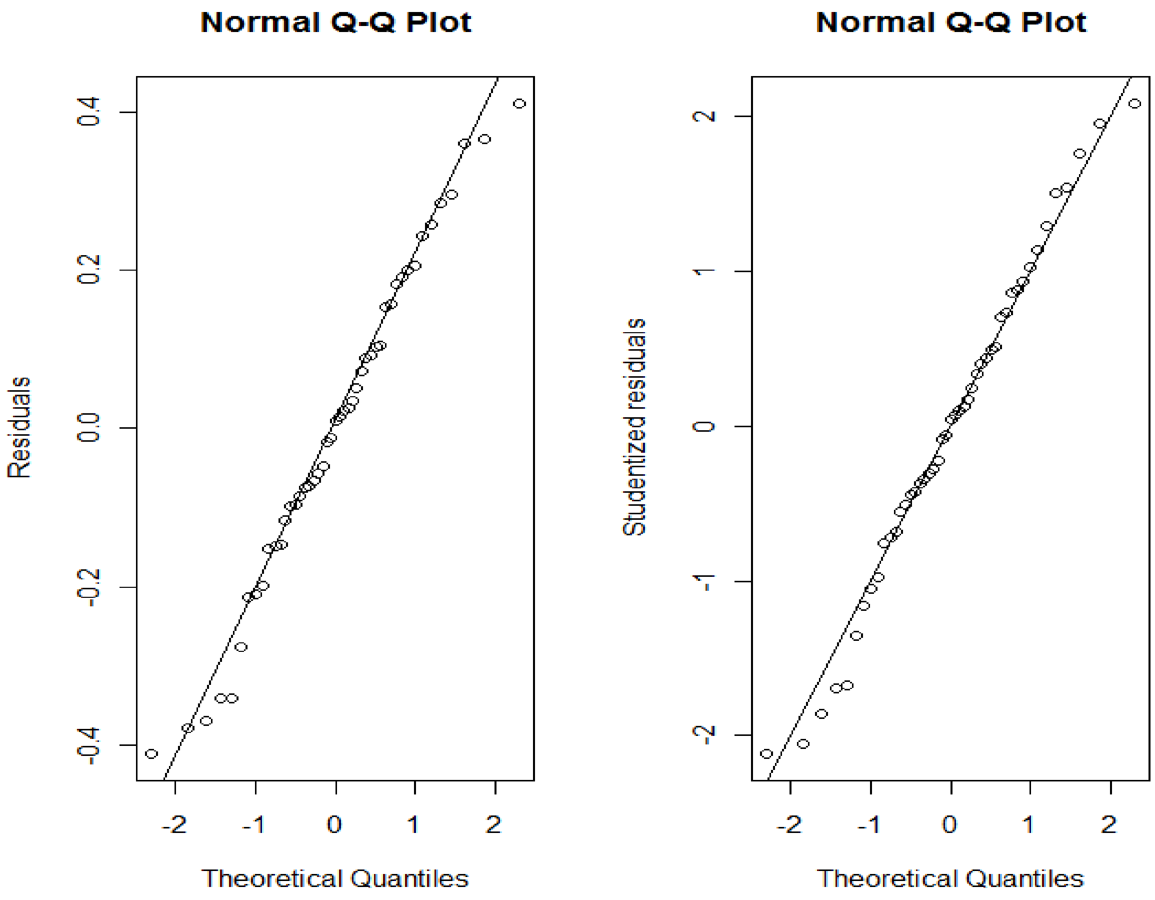


Figure 7: Q-Q Plots

The Shapiro-Wilk test is a formal test for normality. In this case, the p-value (0.714) is large (greater than 0.05), so the null (the residuals are normal) is not rejected. In conclusion, the assumption of normality (observed samples come from normal distribution) is reasonable.

Check for large leverage points

A leverage point can be defined as an "unusual point in the predictor space—it has the potential to influence the fit", which identifies with unusual or outlying x-vales. Leverages greater than 2p/n should be examined. No large leverages points detected. This is illustrated in the plots in Figure 8.

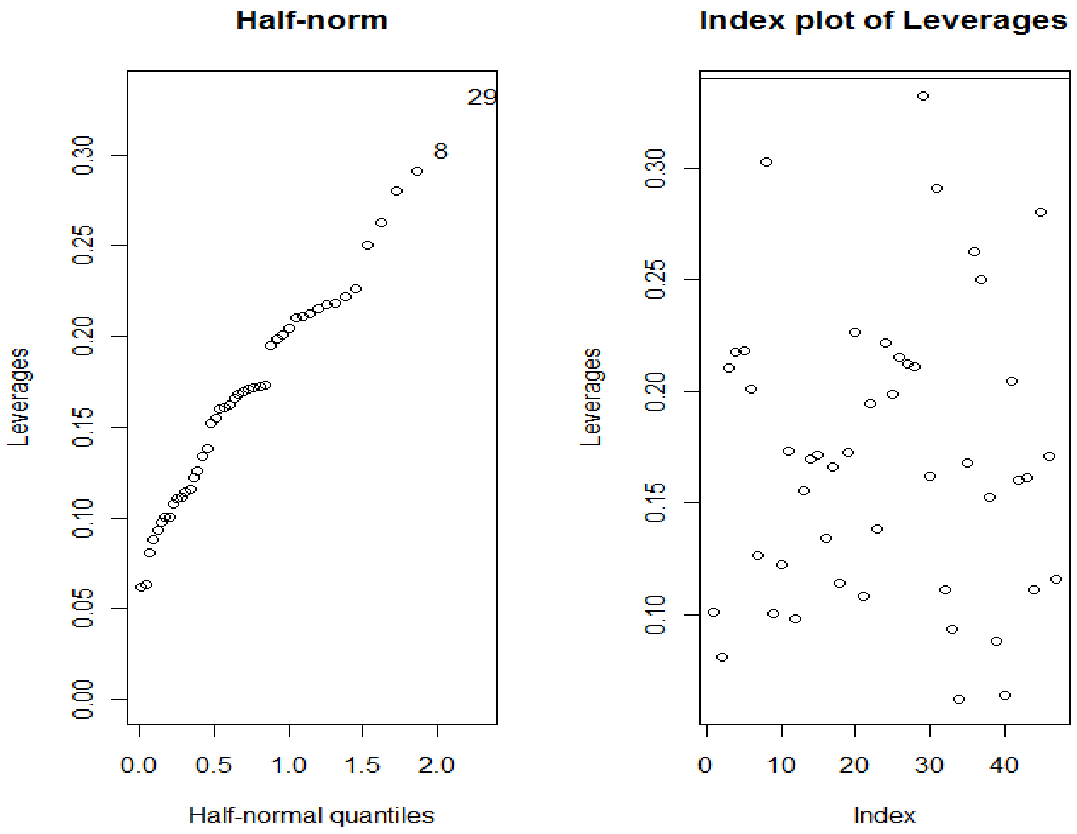


Figure 8: Half-Norm and Leverage Plots

Check for outliers and high influential points

Similar tests as in the preliminary analysis were conducted, resulting in neither outliers nor high influential points detected.

Check Correlated Errors

To check the assumption of uncorrelated errors, a Durbin-Watson test was done. The p-value of the test was 0.138, indicating we cannon reject the null hypothesis that errors are uncorrelated (assumption of uncorrelated errors follows a linear combination of χ^2). Thus, there was no evidence of correlation, so assumption of uncorrelated error holds.

Check the structure of the relationship between the predictors and the response

To detect non-linearity in the model, partial residuals plots were constructed for each predictor. Figure 9. shows these plots and it can be seen that the linear assumption is reasonable, that is the systematic part (Ey=Xb) of the model is correct, for all predictors except U1. Thus, there is an indication of anon-linear relationship between response variable (R) and explanatory variable U1, meaning the linear assumption does not hold for this predictor.

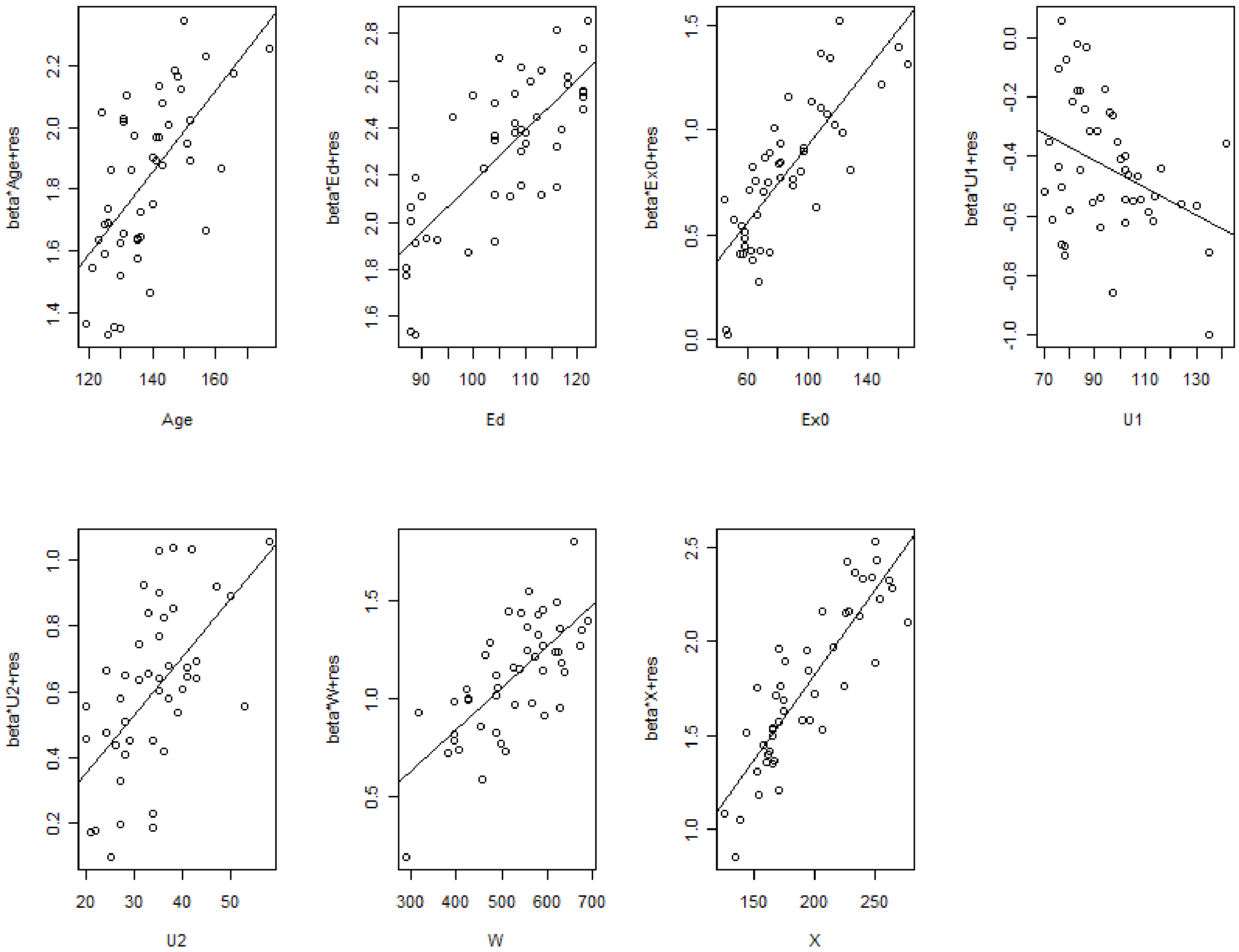


Figure 9: Partial Residual Plots

Check the structure of the relationship between the predictors and the response

The diagnosis indicate that all assumptions are reasonable except the linear structure between the response R and the predictor U1. Thus, adjustments to the AIC model were needed.

Adjustments

U1 and U2 are highly correlated (0.746), which suggests that only one of them can be included in the model. Looking at Table 3., the second best AIC model was the same as the best AIC model but without U1, and its AIC (-132.50) was very close to the chosen AIC model (-152.59). After doing an F-test comparing these 2-models, the p-value (0.191) suggests that these models are not significantly different. In other words, the addition of U1 does not have a significant effect in reducing the residual sum of squares. Thus, it was decided to drop U1 from the model. A model with log(R) as the response and with the 6 predictors Age, Ed, Ex0, U2, W, and X was selected.

5. Analysis

BIC

The final model by BIC was:

$$\mathrm{Log}(R) = -2.792 + 0.011 \cdot Age + 0.015 \cdot Ed + 0.010 \cdot Ex0 + 0.003 \cdot W + 0.010 \cdot X$$

Predictor	Coefficient	Standard Error	P-value	Effect on R
Intercept	-2.792	1.135	0.018	
Age	0.011	0.004	0.006	↑ 1.1 %
Ed	0.015	0.005	0.005	↑ 1.5 %
Ex0	0.010	0.002	0.000	↑ 1.0 %
W	0.003	0.001	0.020	↑ 0.3 %
X	0.010	0.002	0.000	↑ 1.0 %

Table 4. Summary of BIC model

The summary in Table 4. shows that all predictors are significant at a 0.05 level. The model explains 71.02% $(R^2 = 0.7102)$ variation of the data. The coefficients can be interpreted as follows: A unit increase in the predictor is associated with an average of 100xCoefficient percent increase in the response while all other variables in the model are held constant .So for instance, holding all other predictors constant, an increase of 1 unit of Age (The number of males of age 14-24 per 1000 population) increases the crime rate (R) by 1.1 % because R is multiplied by 1.011 (1+0.011).

All variables selected in this model were expected to be selected as predictors of the response, since age, education (Ed), police expenditure (Ex0), wealth (W) and poverty (X) are usually associated with crime rate. All coefficients are similar in magnitude (increase R by 1.0% to 1.5%), that is all have about the same effect on the response, with the exception of W, which only increases R by 0.3%.

All variables positively contribute to crime rate. This is expected from age and poverty (X), but it seems counterintuitive for education and police expenditure (Ex0), since one would expect lower crime rate for higher education and higher police expenditure. One possible reason is that in larger cities, where crime rates are usually higher, there is a higher concentration of educated people and crime-related expenditures tend to be higher. Another reason might be that a lurking variable might be causing the relationship of these predictors with the crime rate and was not accounted in the study. There is a strong collinearity in the data, which might be the cause of the counterintuitive effects of the predictors on the response.

Note: after removing case 29, the regression coefficients remain about the same but W is no longer significant at the 0.05 level. The model explains 72.71% $(R^2 = 0.7271)$ variation of the data. Table 5. shows the output of this regression model. The changes of the coefficient of the model (without the high leverage point) with respect to the coefficients of the model with all cases are shown in the last column. We see that the magnitude of all coefficients decrease, with W having the largest decrease, except Ex0, which increases by about 15%.

Predictor	Coefficient	Standard Error	P-value	Effect on R	Change of coefficient
Intercept	-2.339	1.144	0.048		
Age	0.010	0.004	0.013	↑ 0.97 %	-11.10 %
Ed	0.014	0.005	0.007	↑ 1.39 %	-6.56 %
Ex0	0.012	0.002	0.000	↑ 1.21 %	15.30 %
W	0.002	0.001	0.055	↑ 0.21 %	-17.34 %
X	0.009	0.002	0.000	↑ 0.94 %	-4.10 %

Table 5. Summary of BIC model with case 29 removed

AIC (adjusted)

The final model by AIC was:

$$\mathrm{Log}(R) = -3.575 + 0.013 \cdot Age + 0.019 \cdot Ed + 0.010 \cdot Ex0 + 0.008 \cdot U2 + 0.002 \cdot W + 0.009 \cdot X$$

Predictor	Coefficient	Standard Error	P-value	Effect on R
Intercept	--3.575	1.186	0.005	
Age	0.013	0.004	0.001	↑ 1.3 %
Ed	0.019	0.004	0.001	↑ 1.9 %
Ex0	0.010	0.002	0.000	↑ 1.0 %
U2	0.008	0.005	0.077	↑ 0.8 %
W	0.002	0.001	0.030	↑ 0.2 %
X	0.009	0.002	0.000	↑ 0.9 %

Table 6. Summary for AIC (adjusted) model

The summary in Table 6. shows that all predictors are significant at a 0.05 level, except U2. The model explains 73.22% ($R^2 = 0.7322$) variation of the data. The coefficients can be interpreted as follows: A unit increase in the predictor is associated with an average of 100xCoefficient percent increase in the response while all other variables in the model are held constant .So for instance, holding all other predictors constant, an increase of 1 unit of Age (The number of males of age 14-24 per 1000 population) increases the crime rate (R) by 1.3 % because R is multiplied by 1.013 (1+0.013).

All variables selected in this model were expected to be selected as predictors of the response, since age, education (Ed), police expenditure (Ex0), unemployment (U2) wealth (W) and poverty (X) are usually associated with crime rate. All coefficients are similar in magnitude (increase R by 0.9% to 1.9%), that is all have about the same effect on the response, with the exception of W, which only increases R by 0.2%.

All variables positively contribute to crime rate. This is expected from age, unemployment (U2) and poverty (X), but it seems counterintuitive for education and police expenditure (Ex0), since one would expect lower crime rate for higher education and higher police expenditure. One possible reason is that in larger cities, where crime rates are usually higher, there is a higher concentration of educated people and crime-related expenditures tend to be higher. Another reason might be that a lurking variable might be causing the relationship of these predictors with the crime rate and was not accounted in the study. There is a strong collinearity in the data, which might be the cause of the counterintuitive effects of the predictors on the response.

Collinearity

When variables are highly correlated, the variances of regression coefficients tend to be large. High standard errors mean low t-statistics, which in turn lead to not rejecting the null hypothesis (coefficients are zero). In other words, collinearity makes the detection of an effect more difficult if one exists. Because the standard errors are high, the coefficients are not reliable and may change erratically with small changes to the model, making the interpretation more difficult. The square root of the variance inflator factor (VIF) for the full model is shown in Table 7. One can interpret these values as follows: the standard error of the coefficient of Ex0 and Ex1 are approximately 10 times larger than it would have been without collinearity.

Age	S	Ed	Ex0	Ex1	LF	M	N	NW	U1	U2	W	X
2	2	2	10	10	2	2	2	2	2	2	3	3

Table 7. sqrt(VIF)

Table 8. Shows the correlation between all variables, with highly correlated variables (greater in magnitude than 0.7) in bold. We see that Ex1 and Ex0, and U1 and U2 are highly correlated. Also W is highly correlated with Ed, Ex0,Ex2 and X.

	R	Age	S	Ed	Ex0	Ex1	LF	M	N	NW	U1	U2	W	X
R	**1**	-0.09	-0.09	0.32	0.69	0.67	0.19	0.21	0.34	0.03	-0.05	0.18	0.44	-0.18
Age	-0.09	**1**	0.58	-0.53	-0.51	-0.51	-0.16	-0.03	-0.28	0.59	-0.22	-0.24	-0.67	0.64
S	-0.09	0.58	**1**	-0.7	-0.37	-0.38	-0.51	-0.31	-0.05	**0.77**	-0.17	0.07	-0.64	**0.74**
Ed	0.32	-0.53	-0.7	**1**	0.48	0.5	0.56	0.44	-0.02	-0.66	0.02	-0.22	**0.74**	**-0.77**
Ex0	0.69	-0.51	-0.37	0.48	**1**	**0.99**	0.12	0.03	0.53	-0.21	-0.04	0.19	**0.79**	-0.63
Ex1	0.67	-0.51	-0.38	0.5	**0.99**	**1**	0.11	0.02	0.51	-0.22	-0.05	0.17	**0.79**	-0.65
LF	0.19	-0.16	-0.51	0.56	0.12	0.11	**1**	0.51	-0.12	-0.34	-0.23	-0.42	0.29	-0.27
M	0.21	-0.03	-0.31	0.44	0.03	0.02	0.51	**1**	-0.41	-0.33	0.35	-0.02	0.18	-0.17
N	0.34	-0.28	-0.05	-0.02	0.53	0.51	-0.12	-0.41	**1**	0.1	-0.04	0.27	0.31	-0.13
NW	0.03	0.59	**0.77**	-0.66	-0.21	-0.22	-0.34	-0.33	0.1	**1**	-0.16	0.08	-0.59	0.68
U1	-0.05	-0.22	-0.17	0.02	-0.04	-0.05	-0.23	0.35	-0.04	-0.16	**1**	**0.75**	0.04	-0.06
U2	0.18	-0.24	0.07	-0.22	0.19	0.17	-0.42	-0.02	0.27	0.08	**0.75**	**1**	0.09	0.02
W	0.44	-0.67	-0.64	**0.74**	**0.79**	**0.79**	0.29	0.18	0.31	-0.59	0.04	0.09	**1**	**-0.88**
X	-0.18	0.64	**0.74**	**-0.77**	-0.63	-0.65	-0.27	-0.17	-0.13	0.68	-0.06	0.02	**-0.88**	**1**

Table 8. Correlation Matrix

6. Summary

- The preliminary analysis did not find any outliers or high influential points and suggested a log transformation of the response R (crime rate).

- Model selection by BIC suggested the predictors: Age, Ed, Ex0, W, and X

- Model selection by AIC suggested the predictors: Age, Ed, Ex0, U1, U2, W, and

- Diagnosis resulted in:
  - BIC model: high leverage point at case 29
  - AIC model: non-linearity in the model between R and U1. U1 was removed from model.

- Analysis:
  - Predictors included in both models (Age, Ed, Ex0, U2, W and X) all have positively contribute to crime rate.
  - W was determined to be insignificant at the 0.05 level after removing high leverage point for the BIC model.

- Collinearity:
  - There is a strong collinearity in the data, which might be the cause of the counterintuitive effects of the predictors on the response.

- For future analysis:
  - A study on the interactions (since S is a categorical variable) should be included.
  - Address the collinearity issue. When variables are highly correlated, they convey the same information. One suggestion is to combine highly collinear variables. Note that collinearity does not affect the accuracy of the prediction, but is does affect the interpretation of the predictors and their relationship with the response