EXPLORING SOCIO-ECONOMIC CORRELATES OF CRIME RATES IN USA

A MULTIPLE LINEAR REGRESSION ANALYSIS

SYLWIA JABLONSKA

Table of content

1. Abstract

This study explores the interplay of crime rates and socio-economic factors in 47 U.S. states, using 11 variables. Descriptive statistics reveal diverse patterns in the complex socio-economic landscape. Unexpected correlations emphasise the importance of police expenditure, income inequality, educational levels, the number of males (aged 14-24), and the unemployment rate of urban males (aged 35-39) in the regression model. These factors emerge as significant predictors, offering valuable insights for evidence-based decision-making. ANOVA testing confirms the model's fit.

2. Introduction

Crime rates are complex phenomena influenced by various socio-economic factors. Understanding the dynamics that underlie crime rate is a multifaced challenge that necessitates a nuanced examination of them. This paper aims to show the relationship between crime rates 11 variables each representing potential influence of criminal activity.

3. Initial analysis

The data employed in this analysis includes 47 observations. They consist of crime rate and demographic variables, economic variables, law enforcement variables and geographic variables from 47 states in the United State of America.

3.1 Descriptive Statistics (see: Appendix A)

Crime rate (X1)

- The maximum (199.3) and minimum (32.20) values showcase the range, suggesting the considerable diversity in crime rates between the states.

Educational level (X3)

- The standard deviation in educational level (11.187) indicates some variability, however coupled with mean (105.638) and range (min: 87 max: 122) giving the concentrated distribution we can conclude that educational level is relatively consistent across states.

Police expenditure in 1960 (X4) and 1959 (X5)

- With notable standard deviation (29.719 in 1960 and 27.961 in 1959) and extensive range (1960 – min:45 & max:166; 1959 – min:41 & max:157) it implies significant variability in police expenditure and, by extension, funding levels across states.

Number of non-white people per 1000 (X7)

- The highest standard deviation (102.829) from all the data collected in this study suggests significant variation between states in number of non-white community members.

Unemployment rate of urban males per 1000 in the age group 35-39 years (X9)

- The standard deviation (8.45) suggests a degree of variability around the mean (22.979), signifying that there are differences in the unemployment rates among the states, contributing to a nuanced understanding of the economic landscape for this age

group. The range between the minimum (20.00) and maximum (58.00) values illustrates the spread of unemployment rates, emphasising the diversity in economic conditions for urban males aged 35-39 across the dataset.

Wealth as measured by family income (X10)

- The standard deviation (96.49) suggests a notable degree of variability around the mean (525.383) signifying diverse economic conditions and disparities in wealth among the studied states. The range (min:288 & max:689), further emphasises the significant spread in family income, highlighting the economic heterogeneity experienced by families across the states in the dataset.

The dataset reveals a wide spectrum of socio-economic dynamics across states, evident in key variables. This diversity underscores the complex socio-economic tapestry across the studied states.

3.2 Correlation Hypothesis

Given the considerable variability and diversity uncovered within the dataset, it is reasonable to hypothesize that certain socio-economic factors significantly influence crime rates across the studied states and potentially one another. Specifically:

Hypothesis 1
Crime rate vs. Unemployment rate of urban males per 1000 in the age group 35-39 years
H_0 - there is no correlation between variables.
H_1 - correlation is significant.

Hypothesis 2
Crime rate vs. educational level
H_0 - there is no correlation between variables.
H_1 – correlation is significant.

Hypothesis 3
Educational level vs. wealth
H_0- there is no correlation between variables.
H_1 – correlation is significant.

Hypothesis 4
Income inequality vs. Number of non-white people per 1000.
H_0 – there is no correlation between variables.
H_1 – correlation is significant.

3.3 Correlation Analysis & Hypothesis testing

The Pearson Correlation Coefficients (see: Appendix B) provide insights into the relationships between the variables in the dataset. It allows to test previously stated hypothesis (see: 3.2 Correlation hypothesis) offering statistical evidence of the strength and direction of relationship, aiding in the validation or rejection.

3.3.1 Overview correlation analysis:

Crime rate

- Shows weak negative correlation between crime rate and income inequality, however there is a strong positive correlation between police expenditure (for both 1959 & 1960)

Educational level

- Indicated strong negative correlation between three factors: number of males aged 14-24, number of non-white people and income inequality which is confirmation for previous conclusion (see: 3.1 Descriptive statistics). In contrast, there is strong positive correlation between educational level and wealth, which reject initial hypothesis 3 (see: 3.3.2 Correlation hypothesis testing).

Police expenditure (X4 & X5)

- In addition to previously mention strong positive correlation to crime rate it showcase same correlation to measured wealth, alongside to correlation with each other. On the contrary both factors show strong negative correlation with income inequality.

Number of non-white people

- Alas there is strong positive correlation between number of non-white people and income inequality, alongside with number of males aged 14-24 and Southern states. However, Pearson table shows strong negative correlation to wealth. This data disproves initial hypothesis 4.

Unemployment Rate (X9 & X10)

- Both factors present weak negative correlation to number of males aged 14-24 additionally to strong correlation with respect to each other.

Wealth

- In summary to previously mention correlation wealth displays strong positive correlation to education level and police expenditure (for both 1959 & 1960). Moreover, it shows strong negative correlation to number of males aged 14-24, Southern states, number of non-white people and income inequality.

3.3.2 Correlation hypothesis testing

Hypothesis 1

p-value = 0.7362 > 0.05
Fail to reject H_0.

There is no correlation between crime rate and unemployment rate of urban males aged 35-39

Hypothesis 2

p-value = 0.0269 < 0.05
Some evidence to reject H_0 (at 5% significance level)

p-value = 0.0269 > 0.01
Fail to reject H_0 (at 1% significance level)

There is some correlation between crime rate and educational level at 5% significance level, however there is no correlation at 1% significance level.

Hypothesis 3

p-value=<.0001 < 0.01 < 0.05
Very strong evidence to reject H_0 at both 5% and 1% significance level.

There is very strong correlation between educational level and wealth.

Hypothesis 4

p-value = <.0001 <0.001 < 0.01 < 0.05
Very strong evidence to reject H_0 5%, 1% and 0.1% significance level.

There is very strong correlation between income inequality and number of non-white people per 1000.

Further exploration through multivariate analyses is warranted to deepen understanding of the complex relationships inherent in this dataset.

4. Multiple Linear Regression Analysis

4.1 Selection Method: **Forward Selection** (see: Appendix C)

Initially: **null** model R2 = 0% Regression SS = 0

Step 1: **X4 entered.**
Increase in R^2 = 47.28% - 0% = 47.28%
Increase in Regression SS = 32533 – 0 = 32533

Step 2: **X11 entered.**
Increase in R^2 = 58.03% - 47.28% = 10.75%
Increase in Regression SS = 39931 – 32533 = 7398

Step 3: **X3 entered.**
Increase in R^2: 66.56% - 58.03% = 8.53%
Increase in Regression SS: 45802 – 39931 = 5871

Step 4: **X1 entered.**
Increase in R^2: 70.04% - 66.56% = 3.48%
Increase in Regression SS: 48196 – 45802 = 2394

Step 5: **X9 entered.**
Increase in R^2: 72.96% - 70.04% = 2.92%
Increase in Regression SS: 50206 – 48196 = 2010

Final model

y = -524.37433 + 1.01982(X1) + 2.03077(X3) + 1.23312(X4) + 0.91361(X9) + 0.63493(X11)

4.2 Selection Method: **Backward elimination** (see: Appendix D)
Initially: **Full** model
R^2: 76.47%
Regression SS: 52620

Step 1: **X7 removed**.
Decrease R^2: 76.47% - 76.47% = 0%
Decrease in Regression SS: 52620 – 52619 = 1

Step 2: **X2 removed**.
Decrease R^2: 76.47% - 76.20% = 0.27%
Decrease in Regression SS: 52619 – 52434 = 185

Step 3: **X6 removed**.
Decrease R^2: 76.20% - 75.81% = 0.39%
Decrease in Regression SS: 52434 – 52166 = 268

Step 4: **X5 removed**.
Decrease R^2: 75.81% - 75.38% = 0.43%
Decrease in Regression SS: 52166 – 51867 = 299

Step 5: **X8 removed**.
Decrease R^2: 75.38% - 74.78% = 0.6%
Decrease in Regression SS: 51867 – 51458 = 409

Final model

y= -618.50284 + 1.12518(X1) + 1.81786(X3) + 1.05069(X4) + 0.82817(X9) + 0.15956(X10) + 0.82357(X11)

4.3 Selection Method: **Stepwise selection** (see: Appendix E)

The final model follows Forward Selection model.

4.4 Conclusion

Both Forward Selection and Backward Elimination highlighted the importance of variables **X1, X3, X4, X9**, and **X11** in predicting the dependent variable. The choice to follow the **Forward Selection model** reflects its systematic identification of influential predictors while maintaining interpretability, offering a balanced and robust approach to model selection.

5. Model Hypothesis Testing and Analysis

ANOVA tables reveal variable significance. Full table assesses impact, while summary identifies best variables in order by importance.

5.1 Full model ANOVA Testing

Hypothesis
H_0 – full regression model is not a good fit to data.
H_1 – model is a good fit to data.

From ANOVA (see: Appendix F)

p-value = <.0001 <0.001 <0.01 <0.05

There is strong evidence to reject H_0 at 5%, 1% and 0.1% significance level.

Full regression model y is a good fit to data.

5.2 Summary of Possible Models (see: Appendix G).

First set of hypotheses:
H_0: **X4** is not required in the regression model.
H_1: **X4** is significant in the model.

F = 78.74 $F_{1,45}(5\%) \approx 4.08$

78.74> $F_{1,45}(5\%)$ – very strong evidence to reject H0 at 5% significance level indicating X4 has strongly significant effect in the regression model.

Second set of hypotheses:
H_0: **X11** is not required in the regression model.
H_1: **X11** is significant in the model.

F = 17.89 $F_{1,45}(5\%) \approx 4.08$

17.89> $F_{1,45}(5\%)$ – strong evidence to reject H_0 at 5% significance level indicating X11 has significant effect in the regression model.

Third set of hypotheses:
H_0: **X3** is not required in the regression model.
H_1: **X3** is significant in the model.

F = 14.19 $F_{1,45}(5\%) \approx 4.08$

14.19> $F_{1,45}(5\%)$ – strong evidence to reject H_0 at 5% significance level indicating X3 has significant effect in the regression model.

Fourth set of hypotheses:
H_0: **X1** is not required in the regression model.
H_1: **X1** is significant in the model.

$F = 5.79$ $F_{1,45}(5\%) \approx 4.08$

$14.19 > F_{1,45}(5\%)$ – moderate evidence to reject H_0 at 5% significance level indicating X1 has moderately significant effect in the regression model.

Fifth set of hypotheses:
H_0: **X9** is not required in the regression model.
H_1: **X9** is significant in the model.

$F = 4.86$ $F_{1,45}(5\%) \approx 4.08$

$4.86 > F_{1,45}(5\%)$ – some evidence to reject H_0 at 5% significance level indicating X9 has moderately significant effect in the regression model.

5.3 Best model

R^2 = 74.78%

Crime rate = -524.37433 + 1.01982(X1) + 2.03077(X3) + 1.23312(X4) + 0.91361(X9) + 0.63493(X11)

5.4 Testing the validity of regression model (see: Appendix C)

As may be seen the data points lie on the diagonal of the normal probability (see Appendix C).
Scatter plot is randomly concentrated.
Only the mean in histogram is not symmetrically bell shaped.

Residual assumptions validated.

6. Conclusions

This study explored crime rates and socio-economic factors across 47 U.S. states, revealing diverse patterns. Unexpected correlations were found, highlighting significant predictors in the regression model. The findings showcase the paramount importance of police expenditure along with income inequality, educational level, number of males (aged 14-24) and unemployment rate of urban male (aged 35-39). Further multivariate analyses are recommended.

7. Recommendations

Future research should focus on local geographic analyses, longitudinal studies, causal inference, cultural influences, technological innovations, social programs, and international datasets. Qualitative research and exploration of the intersectionality of socio-economic factors are also crucial for a comprehensive understanding of crime patterns.

8.1 Appendix A – Simple Statistics

The SAS System

The CORR Procedure

12 Variables: rate X1 X2 X3 X4 X5 X6 X7 X8 X9 X10 X11

Simple Statistics

Variable	N	Mean	Std Dev	Sum	Minimum	Maximum	Label
rate	47	90.50851	38.67627	4254	34.20000	199.30000	Crime Rate defined as the number of offences known to the police per 1,000,000 population
X1	47	138.57447	12.56763	6513	119.00000	177.00000	Number of males aged 14-24 years per 1000 of total state population
X2	47	0.34043	0.47898	16.00000	0	1.00000	Binary variable distinguishing Southern states (S=1) from the rest
X3	47	105.63830	11.18700	4965	87.00000	122.00000	Educational level
X4	47	85.00000	29.71897	3995	45.00000	166.00000	Police expenditure in 1960
X5	47	80.23404	27.96132	3771	41.00000	157.00000	Police expenditure in 1959
X6	47	36.61702	38.07119	1721	3.00000	168.00000	State population size in hundred thousands.
X7	47	101.12766	102.82882	4753	2.00000	423.00000	Number of non-white people per 1000.
X8	47	95.46809	18.02878	4487	70.00000	142.00000	Unemployment rate of urban males per 1000 in the age group 14-24 years

Simple Statistics

Variable	N	Mean	Std Dev	Sum	Minimum	Maximum	Label
X9	47	33.97872	8.44545	1597	20.00000	58.00000	Unemployment rate of urban males per 1000 in the age group 35-39 years
X10	47	525.38298	96.49094	24693	288.00000	689.00000	Wealth as measured by family income (unit: 10 dollars)
X11	47	194.00000	39.89606	9118	126.00000	276.00000	Income inequality expressed as the number of families per 1000 earning below one-half of the median income.

8.2 Appendix B – Pearson Correlation Coefficients Table

Pearson Correlation Coefficients, N = 47
Prob > |r| under H0: Rho=0

	rate	X1	X2	X3	X4	X5	X6	X7	X8	X9	X10	X11
rate Crime Rate defined as the number of offences known to the police per 1,000,000 population	1.00000	-0.08947 0.5498	-0.09064 0.5446	**0.32283** **0.0269**	0.68760 <.0001	0.66671 <.0001	0.33747 0.0204	0.03260 0.8278	**-0.05048** **0.7362**	0.17732 0.2331	0.44132 0.0019	-0.17902 0.2286
X1 Number of males aged 14-24 years per 1000 of total state population	-0.08947 0.5498	1.00000	0.58436 <.0001	-0.53024 0.0001	-0.50574 0.0003	-0.51317 0.0002	-0.28064 0.0560	0.59320 <.0001	-0.22438 0.1295	-0.24484 0.0972	-0.67006 <.0001	0.63921 <.0001
X2 Binary variable distinguishing Southern states (S=1) from the rest	-0.09064 0.5446	0.58436 <.0001	1.00000	-0.70274 <.0001	-0.37264 0.0099	-0.37617 0.0092	-0.04992 0.7390	0.76710 <.0001	-0.17242 0.2465	0.07169 0.6320	-0.63695 <.0001	0.73718 <.0001

Pearson Correlation Coefficients, N = 47
Prob > |r| under H0: Rho=0

	rate	X1	X2	X3	X4	X5	X6	X7	X8	X9	X10	X11
X3 Educational level	**0.32283** **0.02069**	-0.53024 0.00001	-0.70274 <.0001	1.00000	0.48295 0.0006	0.49941 0.0004	-0.01723 0.9085	-0.66488 <.0001	0.01810 0.9039	-0.21568 0.1454	**0.73600** **<.0001**	-0.76866 <.0001
X4 Police expenditure in 1960	0.68760 <.0001	-0.50574 0.0003	-0.37264 0.0099	0.48295 0.0006	1.00000	0.99359 <.0001	0.52628 0.0001	-0.21371 0.1492	-0.04370 0.7706	0.18509 0.2129	0.78723 <.0001	-0.63050 <.0001
X5 Police expenditure in 1959	0.66671 <.0001	-0.51317 0.0002	-0.37617 0.0092	0.49941 0.0004	0.99359 <.0001	1.00000	0.51379 0.0002	-0.21877 0.1396	-0.05171 0.7299	0.16922 0.2555	0.79426 <.0001	-0.64815 <.0001
X6 State population size in hundred thousands.	0.33747 0.0204	-0.28064 0.0560	-0.04992 0.7390	-0.01723 0.9085	0.52628 0.0001	0.51379 0.0002	1.00000	0.09515 0.5246	-0.03812 0.7992	0.27042 0.0660	0.30826 0.0350	-0.12629 0.3976
X7 Number of non-white people per 1000.	0.03260 0.8278	0.59320 <.0001	0.76710 <.0001	-0.66488 <.0001	-0.21371 0.1492	-0.21877 0.1396	0.09515 0.5246	1.00000	-0.15645 0.2936	0.08091 0.5888	-0.59011 <.0001	**0.67731** **<.0001**
X8 Unemployment rate of urban males per 1000 in the age group	-0.05048 0.7362	-0.22438 0.1295	-0.17242 0.2465	0.01810 0.9039	-0.04370 0.7706	-0.05171 0.7299	-0.03812 0.7992	-0.15645 0.2936	1.00000	0.74592 <.0001	0.04486 0.7646	-0.06383 0.6699

Pearson Correlation Coefficients, N = 47
Prob > |r| under H0: Rho=0

	rate	X1	X2	X3	X4	X5	X6	X7	X8	X9	X10	X11
14-24 years												
X9 Unemploymen t rate of urban males per 1000 in the age group 35-39 years	0.17732 0.2331	-0.24484 0.0972	0.07169 0.6320	-0.21568 0.1454	0.18509 0.2129	0.16922 0.2555	0.27042 0.0660	0.08091 0.5888	0.74592 <.0001	1.00000	0.09207 0.5382	0.01568 0.9167
X10 Wealth as measured by family income (unit: 10 dollars)	0.44132 0.0019	-0.67006 <.0001	-0.63695 <.0001	**0.73600 <.0001**	0.78723 <.0001	0.79426 <.0001	0.30826 0.0350	-0.59011 <.0001	0.04486 0.7646	0.09207 0.5382	1.00000	-0.88400 <.0001
X11 Income inequality expressed as the number of families per 1000 earning below one-half of the median income.	-0.17902 0.2286	0.63921 <.0001	0.73718 <.0001	-0.76866 <.0001	-0.63050 <.0001	-0.64815 <.0001	-0.12629 0.3976	0.67731 <.0001	-0.06383 0.6699	0.01568 0.9167	-0.88400 <.0001	1.00000

8.3 Appendix C – Forward Selection

The SAS System

The REG Procedure
Model: MODEL1
Dependent Variable: rate Crime Rate defined as the number of offences known to the police per 1,000,000 population

Number of Observations Read	47
Number of Observations Used	47

Forward Selection: Step 1

Variable X4 Entered: R-Square = 0.4728 and C(p) = 35.4276

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	32533	32533	40.36	<.0001
Error	45	36276	806.13907		
Corrected Total	46	68809			

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	14.44640	12.66926	1048.15843	1.30	0.2602
X4	0.89485	0.14086	32533	40.36	<.0001

Bounds on condition number: 1, 1

Forward Selection: Step 2

Variable X11 Entered: R-Square = 0.5803 and C(p) = 21.4331

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	39931	19966	30.42	<.0001
Error	44	28878	656.31982		
Corrected Total	46	68809			

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	-94.46616	34.39470	4950.90882	7.54	0.0087
X4	1.24148	0.16375	37726	57.48	<.0001
X11	0.40953	0.12198	7398.18643	11.27	0.0016

Bounds on condition number: 1.6598, 6.6393

Forward Selection: Step 3

Variable X3 Entered: R-Square = 0.6656 and C(p) = 10.7413

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	45802	15267	28.53	<.0001
Error	43	23008	535.05987		
Corrected Total	46	68809			

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	-327.54088	76.91367	9703.45462	18.14	0.0001
X3	1.57869	0.47661	5870.49757	10.97	0.0019
X4	1.24314	0.14785	37827	70.70	<.0001
X11	0.75058	0.15077	13261	24.78	<.0001

Bounds on condition number: 3.1105, 21.643

Forward Selection: Step 4

Variable X1 Entered: R-Square = 0.7004 and C(p) = 7.5655

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	48196	12049	24.55	<.0001
Error	42	20614	490.79828		
Corrected Total	46	68809			

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	-424.92222	85.85140	12023	24.50	<.0001
X1	0.76022	0.34421	2394.04639	4.88	0.0327
X3	1.66050	0.45797	6452.18670	13.15	0.0008
X4	1.29804	0.14377	40008	81.52	<.0001
X11	0.64091	0.15270	8646.71205	17.62	0.0001

Bounds on condition number: 3.4783, 37.613

Forward Selection: Step 5

Variable X9 Entered: R-Square = 0.7296 and C(p) = 5.2202

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	50206	10041	22.13	<.0001
Error	41	18604	453.74745		
Corrected Total	46	68809			

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	-524.37433	95.11557	13791	30.39	<.0001
X1	1.01982	0.35320	3782.81167	8.34	0.0062
X3	2.03077	0.47419	8322.11780	18.34	0.0001

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
X4	1.23312	0.14163	34394	75.80	<.0001
X9	0.91361	0.43409	2009.88258	4.43	0.0415
X11	0.63493	0.14685	8482.73176	18.69	<.0001

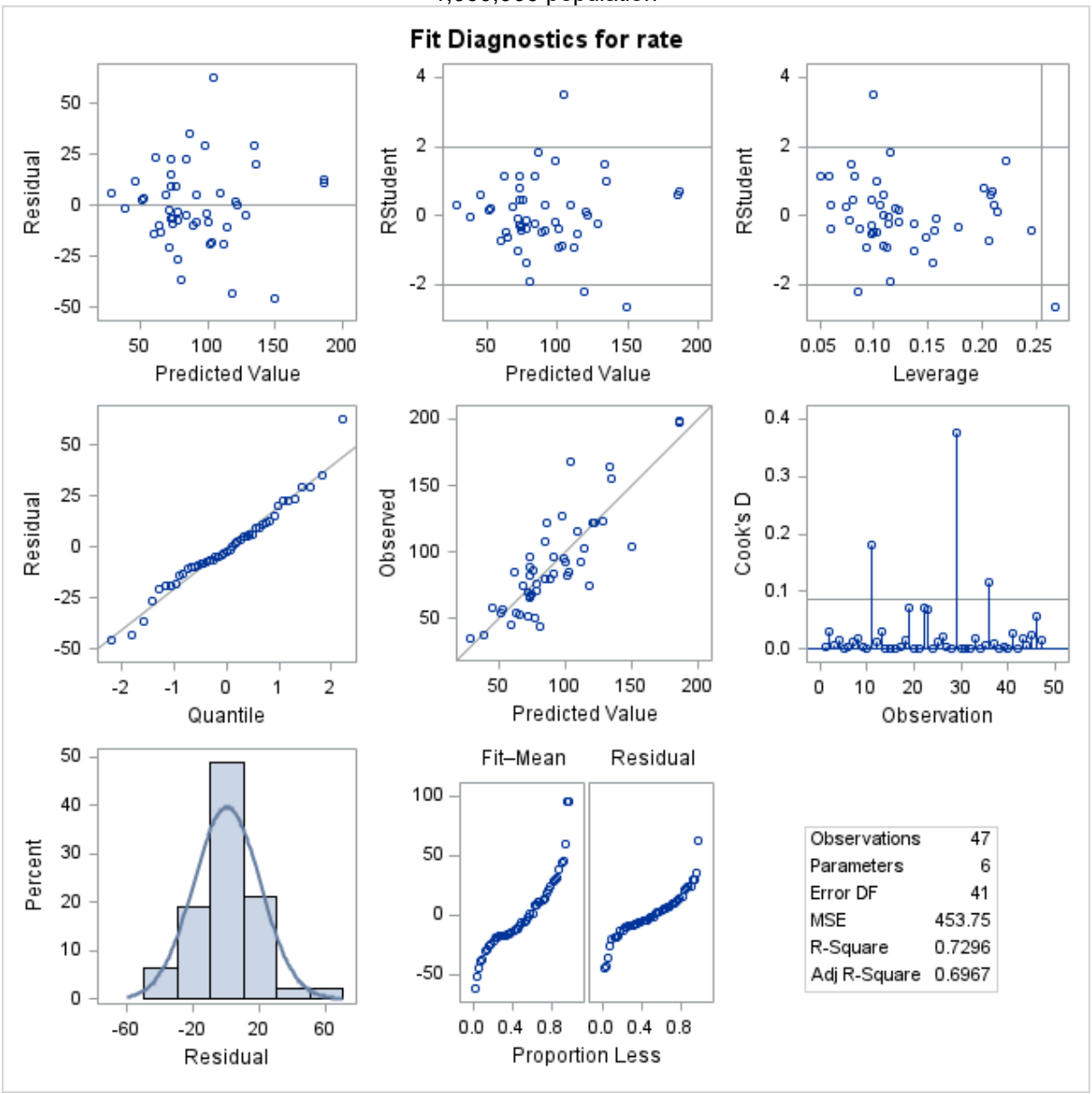
Bounds on condition number: 3.4796, 57.443

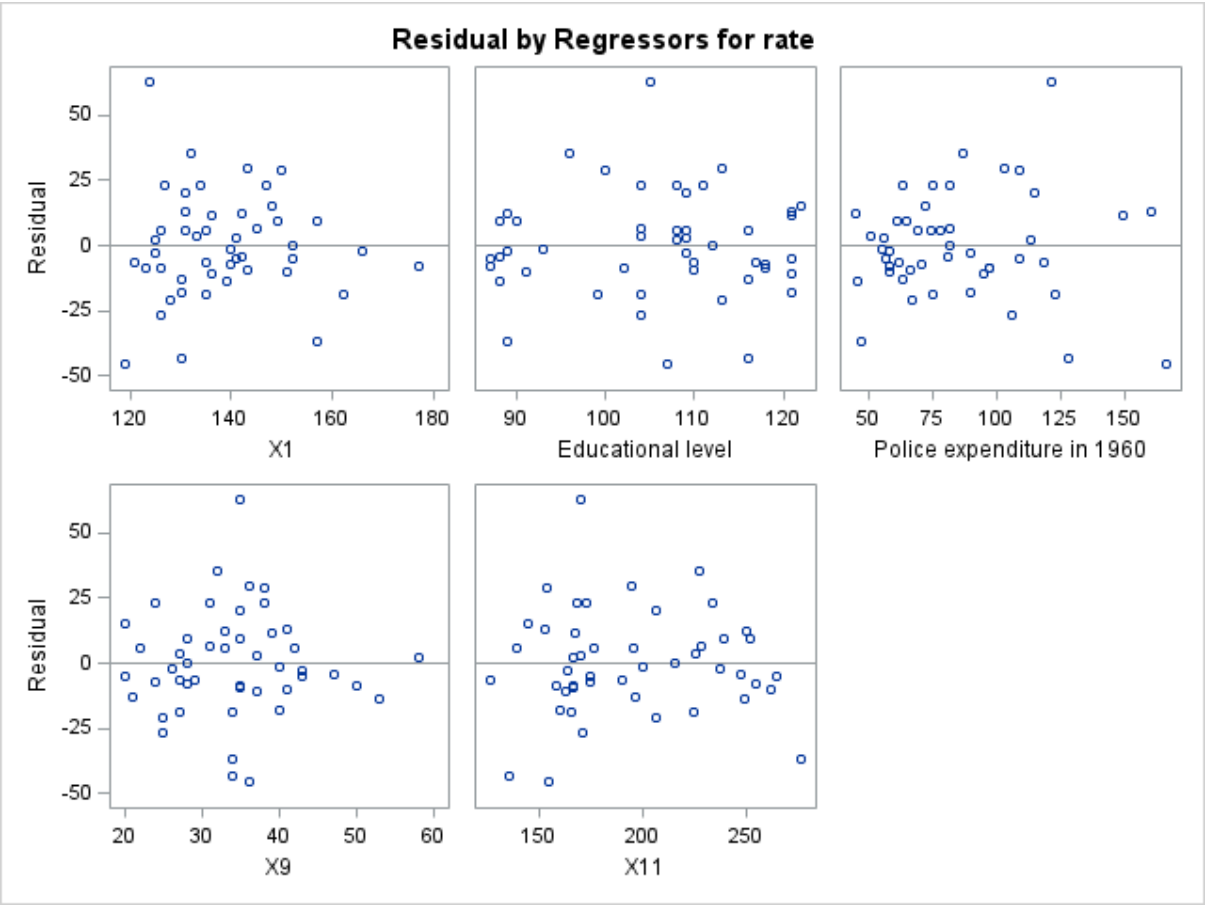
No other variable met the 0.0500 significance level for entry into the model.

Summary of Forward Selection

Step	Variable Entered	Label	Number Vars In	Partial R-Square	Model R-Square	C(p)	F Value	Pr > F
1	X4	Police expenditure in 1960	1	0.4728	0.4728	35.4276	40.36	<.0001
2	X11	Income inequality expressed as the number of families per 1000 earning below one-half of the median income.	2	0.1075	0.5803	21.4331	11.27	0.0016
3	X3	Educational level	3	0.0853	0.6656	10.7413	10.97	0.0019
4	X1	Number of males aged 14-24 years per 1000 of total state population	4	0.0348	0.7004	7.5655	4.88	0.0327
5	X9	Unemployment rate of urban males per 1000 in the age group 35-39 years	5	0.0292	0.7296	5.2202	4.43	0.0415

The REG Procedure
Model: MODEL1
Dependent Variable: rate Crime Rate defined as the number of offences known to the police per 1,000,000 population




Residual by Regressors for rate
Residual
50
25
0
-25
-50
120
140
160
180
X1
90
100
110
120
Educational level
50
75
100
125
150
Police expenditure in 1960
Residual
50
25
0
-25
-50
20
30
40
50
60
X9
150
200
250
X11

8.4 Appendix D – Backwards Elimination

The SAS System

The REG Procedure
Model: MODEL2
Dependent Variable: rate Crime Rate defined as the number of offences known to the police per 1,000,000 population

Number of Observations Read 47

Number of Observations Used 47

Backward Elimination: Step 0

All Variables Entered: R-Square = 0.7647 and C(p) = 12.0000

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	11	52620	4783.65615	10.34	<.0001
Error	35	16189	462.54454		
Corrected Total	46	68809			

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	-608.12646	113.50101	13278	28.71	<.0001
X1	1.13644	0.39643	3801.05775	8.22	0.0070
X2	-7.08044	12.59250	146.23475	0.32	0.5775
X3	1.90376	0.56648	5223.98811	11.29	0.0019
X4	1.71473	0.99535	1372.77075	2.97	0.0938
X5	-0.71326	1.07249	204.57901	0.44	0.5104
X6	-0.09121	0.11137	310.26713	0.67	0.4183
X7	-0.00293	0.05890	1.14863	0.00	0.9605
X8	-0.41862	0.34601	677.05440	1.46	0.2344
X9	1.64398	0.81934	1862.14342	4.03	0.0526
X10	0.14799	0.10096	993.89141	2.15	0.1516

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
X11	0.84495	0.20956	7519.41608	16.26	0.0003

Bounds on condition number: 89.435, 2386.9

Backward Elimination: Step 1

Variable X7 Removed: R-Square = 0.7647 and C(p) = 10.0025

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	10	52619	5261.90690	11.70	<.0001
Error	36	16190	449.72799		
Corrected Total	46	68809			

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	-607.85711	111.79050	13297	29.57	<.0001
X1	1.13082	0.37473	4095.33529	9.11	0.0047
X2	-7.32813	11.40864	185.55332	0.41	0.5247
X3	1.90690	0.55511	5307.02847	11.80	0.0015
X4	1.71751	0.97993	1381.52708	3.07	0.0882
X5	-0.72062	1.04744	212.86576	0.47	0.4959
X6	-0.09187	0.10904	319.22657	0.71	0.4051
X8	-0.41846	0.34117	676.60173	1.50	0.2279
X9	1.64277	0.80756	1861.03788	4.14	0.0493
X10	0.14905	0.09730	1055.23242	2.35	0.1343
X11	0.84399	0.20576	7566.87341	16.83	0.0002

Bounds on condition number: 87.738, 2100.5

Backward Elimination: Step 2

Variable X2 Removed: R-Square = 0.7620 and C(p) = 8.4036

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	9	52434	5825.94619	13.16	<.0001
Error	37	16376	442.58813		
Corrected Total	46	68809			

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	-599.93385	110.22238	13112	29.63	<.0001
X1	1.07284	0.36080	3913.19405	8.84	0.0052
X3	1.97508	0.54052	5909.34199	13.35	0.0008
X4	1.82174	0.95870	1598.11560	3.61	0.0652
X5	-0.84627	1.02082	304.17667	0.69	0.4124
X6	-0.08352	0.10740	267.65763	0.60	0.4417
X8	-0.35671	0.32473	534.04890	1.21	0.2791
X9	1.51067	0.77471	1682.91611	3.80	0.0588
X10	0.14728	0.09649	1031.20290	2.33	0.1354
X11	0.79683	0.19068	7728.98994	17.46	0.0002

Bounds on condition number: 84.678, 1797.8

Backward Elimination: Step 3

Variable X6 Removed: R-Square = 0.7581 and C(p) = 6.9823

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	8	52166	6520.73226	14.89	<.0001
Error	38	16643	437.98470		
Corrected Total	46	68809			

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	-605.25439	109.43622	13397	30.59	<.0001
X1	1.12472	0.35273	4453.12332	10.17	0.0029
X3	2.04750	0.52967	6544.87560	14.94	0.0004
X4	1.75564	0.94994	1496.01605	3.42	0.0724
X5	-0.83905	1.01545	299.03395	0.68	0.4138
X8	-0.32755	0.32088	456.40153	1.04	0.3138
X9	1.44963	0.76670	1565.72744	3.57	0.0663
X10	0.14088	0.09564	950.42496	2.17	0.1490
X11	0.77165	0.18693	7463.44401	17.04	0.0002

Bounds on condition number: 84.671, 1574.7

Backward Elimination: Step 4

Variable X5 Removed: R-Square = 0.7538 and C(p) = 5.6288

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	7	51867	7409.54631	17.06	<.0001
Error	39	16942	434.42186		
Corrected Total	46	68809			

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	-614.66543	108.39833	13968	32.15	<.0001
X1	1.13175	0.35119	4511.60480	10.39	0.0026
X3	2.02739	0.52695	6430.55889	14.80	0.0004
X4	0.98629	0.18751	12020	27.67	<.0001
X8	-0.30919	0.31880	408.60846	0.94	0.3381
X9	1.44156	0.76352	1548.59980	3.56	0.0665
X10	0.14548	0.09509	1016.86094	2.34	0.1341

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
X11	0.79609	0.18382	8147.68549	18.76	0.0001

Bounds on condition number: 8.9136, 220.78

Backward Elimination: Step 5

Variable X8 Removed: R-Square = 0.7478 and C(p) = 4.5122

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	6	51458	8576.36928	19.77	<.0001
Error	40	17351	433.77652		
Corrected Total	46	68809			

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	-618.50284	108.24560	14162	32.65	<.0001
X1	1.12518	0.35086	4461.00468	10.28	0.0026
X3	1.81786	0.48027	6214.72948	14.33	0.0005
X4	1.05069	0.17522	15597	35.96	<.0001
X9	0.82817	0.42740	1628.68477	3.75	0.0597
X10	0.15956	0.09390	1252.58447	2.89	0.0970
X11	0.82357	0.18149	8932.27708	20.59	<.0001

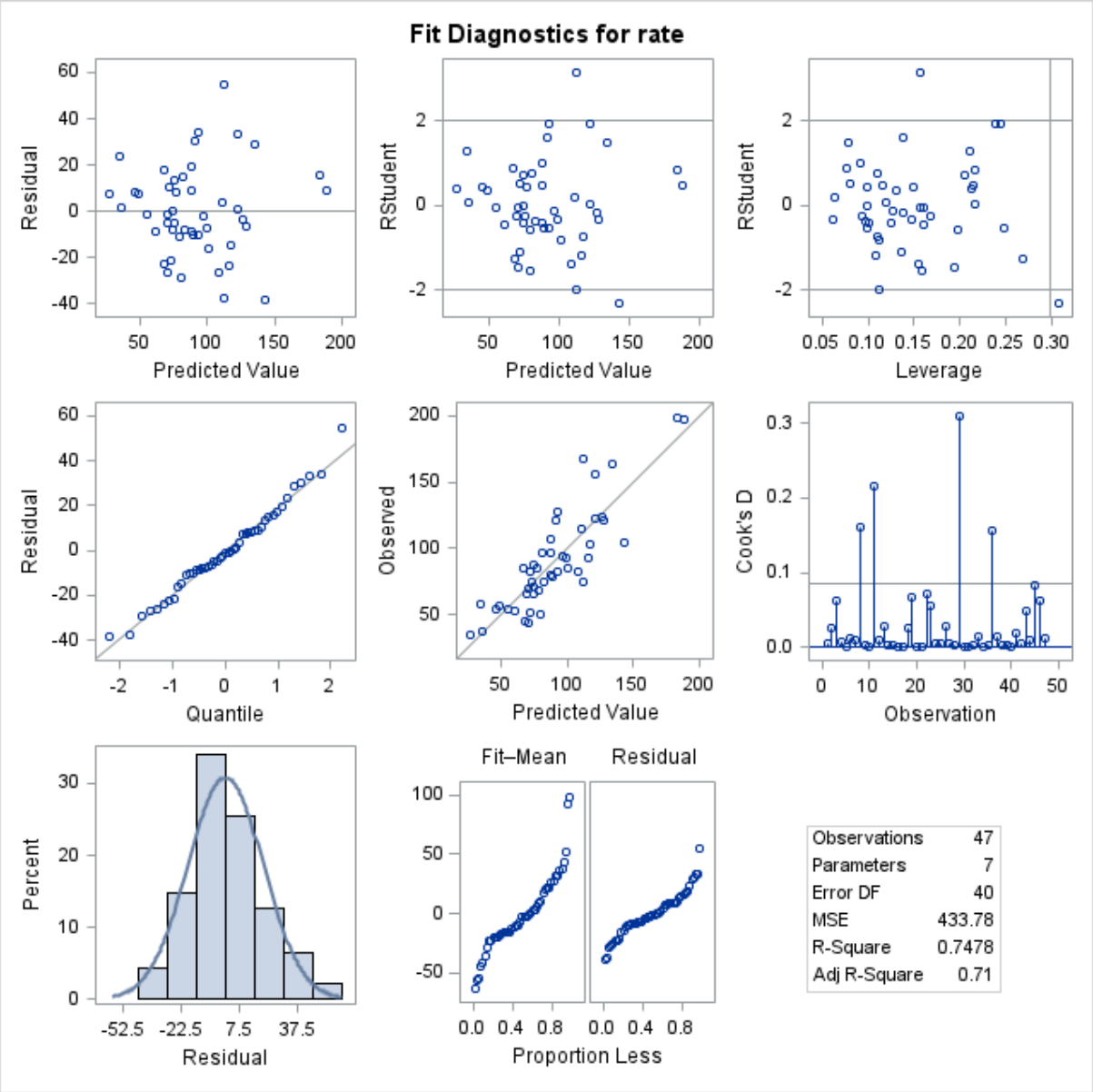
Bounds on condition number: 8.7056, 141.88

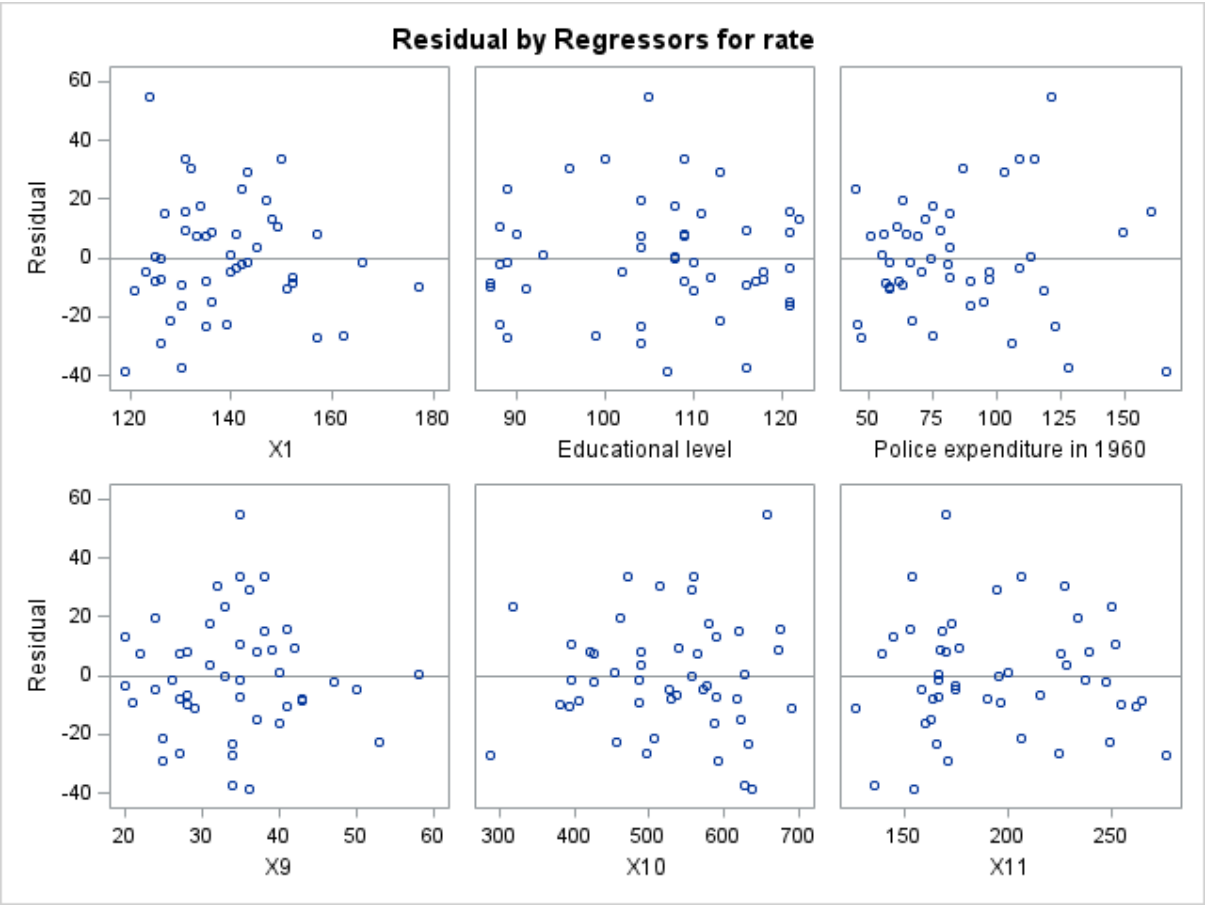
All variables left in the model are significant at the 0.1000 level.

Summary of Backward Elimination

Step	Variable Removed	Label	Number Vars In	Partial R-Square	Model R-Square	C(p)	F Value	Pr > F
1	X7	Number of non-white people per 1000.	10	0.0000	0.7647	10.0025	0.00	0.9605
2	X2	Binary variable distinguishing Southern states (S=1) from the rest	9	0.0027	0.7620	8.4036	0.41	0.5247
3	X6	State population size in hundred thousands.	8	0.0039	0.7581	6.9823	0.60	0.4417
4	X5	Police expenditure in 1959	7	0.0043	0.7538	5.6288	0.68	0.4138
5	X8	Unemployment rate of urban males per 1000 in the age group 14-24 years	6	0.0059	0.7478	4.5122	0.94	0.3381

The REG Procedure
Model: MODEL2
Dependent Variable: rate Crime Rate defined as the number of offences known to the police per 1,000,000 population




Residual by Regressors for rate
Residual
60
40
20
0
-20
-40
120
140
160
180
X1
90
100
110
120
Educational level
50
75
100
125
150
Police expenditure in 1960
20
30
40
50
60
X9
300
400
500
600
700
X10
150
200
250
X11

8.5 Appendix E – Stepwise Selection

The SAS System

The REG Procedure
Model: MODEL3
Dependent Variable: rate Crime Rate defined as the number of offences known to the police per 1,000,000 population

Number of Observations Read 47

Number of Observations Used 47

Stepwise Selection: Step 1

Variable X4 Entered: R-Square = 0.4728 and C(p) = 35.4276

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	32533	32533	40.36	<.0001
Error	45	36276	806.13907		
Corrected Total	46	68809			

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	14.44640	12.66926	1048.15843	1.30	0.2602
X4	0.89485	0.14086	32533	40.36	<.0001

Bounds on condition number: 1, 1

Stepwise Selection: Step 2

Variable X11 Entered: R-Square = 0.5803 and C(p) = 21.4331

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	39931	19966	30.42	<.0001
Error	44	28878	656.31982		
Corrected Total	46	68809			

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	-94.46616	34.39470	4950.90882	7.54	0.0087
X4	1.24148	0.16375	37726	57.48	<.0001
X11	0.40953	0.12198	7398.18643	11.27	0.0016

Bounds on condition number: 1.6598, 6.6393

Stepwise Selection: Step 3

Variable X3 Entered: R-Square = 0.6656 and C(p) = 10.7413

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	45802	15267	28.53	<.0001
Error	43	23008	535.05987		
Corrected Total	46	68809			

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	-327.54088	76.91367	9703.45462	18.14	0.0001
X3	1.57869	0.47661	5870.49757	10.97	0.0019
X4	1.24314	0.14785	37827	70.70	<.0001
X11	0.75058	0.15077	13261	24.78	<.0001

Bounds on condition number: 3.1105, 21.643

Stepwise Selection: Step 4

Variable X1 Entered: R-Square = 0.7004 and C(p) = 7.5655

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	48196	12049	24.55	<.0001
Error	42	20614	490.79828		
Corrected Total	46	68809			

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	-424.92222	85.85140	12023	24.50	<.0001
X1	0.76022	0.34421	2394.04639	4.88	0.0327
X3	1.66050	0.45797	6452.18670	13.15	0.0008
X4	1.29804	0.14377	40008	81.52	<.0001
X11	0.64091	0.15270	8646.71205	17.62	0.0001

Bounds on condition number: 3.4783, 37.613

Stepwise Selection: Step 5

Variable X9 Entered: R-Square = 0.7296 and C(p) = 5.2202

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	50206	10041	22.13	<.0001
Error	41	18604	453.74745		
Corrected Total	46	68809			

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	-524.37433	95.11557	13791	30.39	<.0001
X1	1.01982	0.35320	3782.81167	8.34	0.0062
X3	2.03077	0.47419	8322.11780	18.34	0.0001

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
X4	1.23312	0.14163	34394	75.80	<.0001
X9	0.91361	0.43409	2009.88258	4.43	0.0415
X11	0.63493	0.14685	8482.73176	18.69	<.0001

Bounds on condition number: 3.4796, 57.443

All variables left in the model are significant at the 0.1500 level.

No other variable met the 0.0500 significance level for entry into the model.

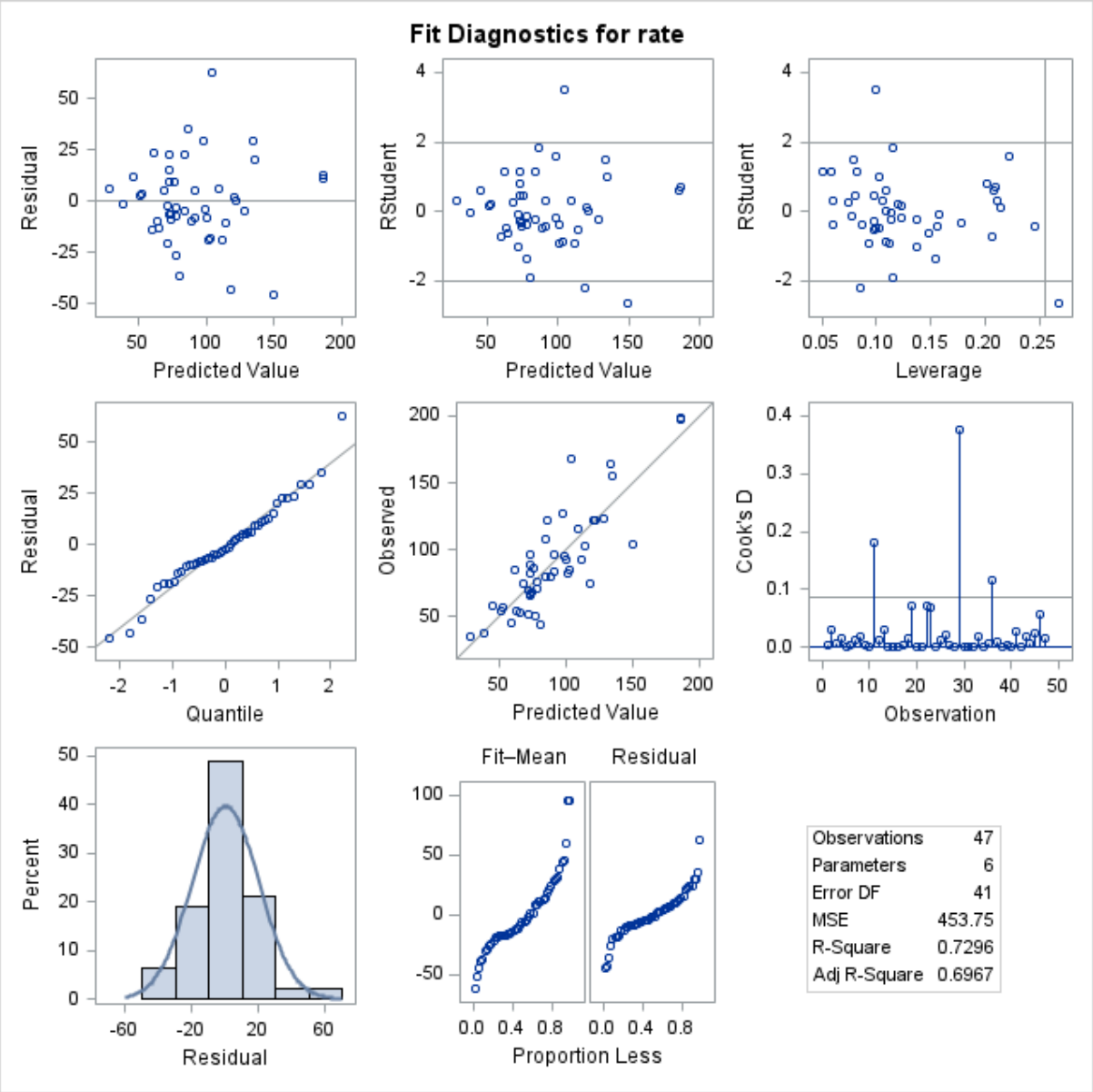
Summary of Stepwise Selection

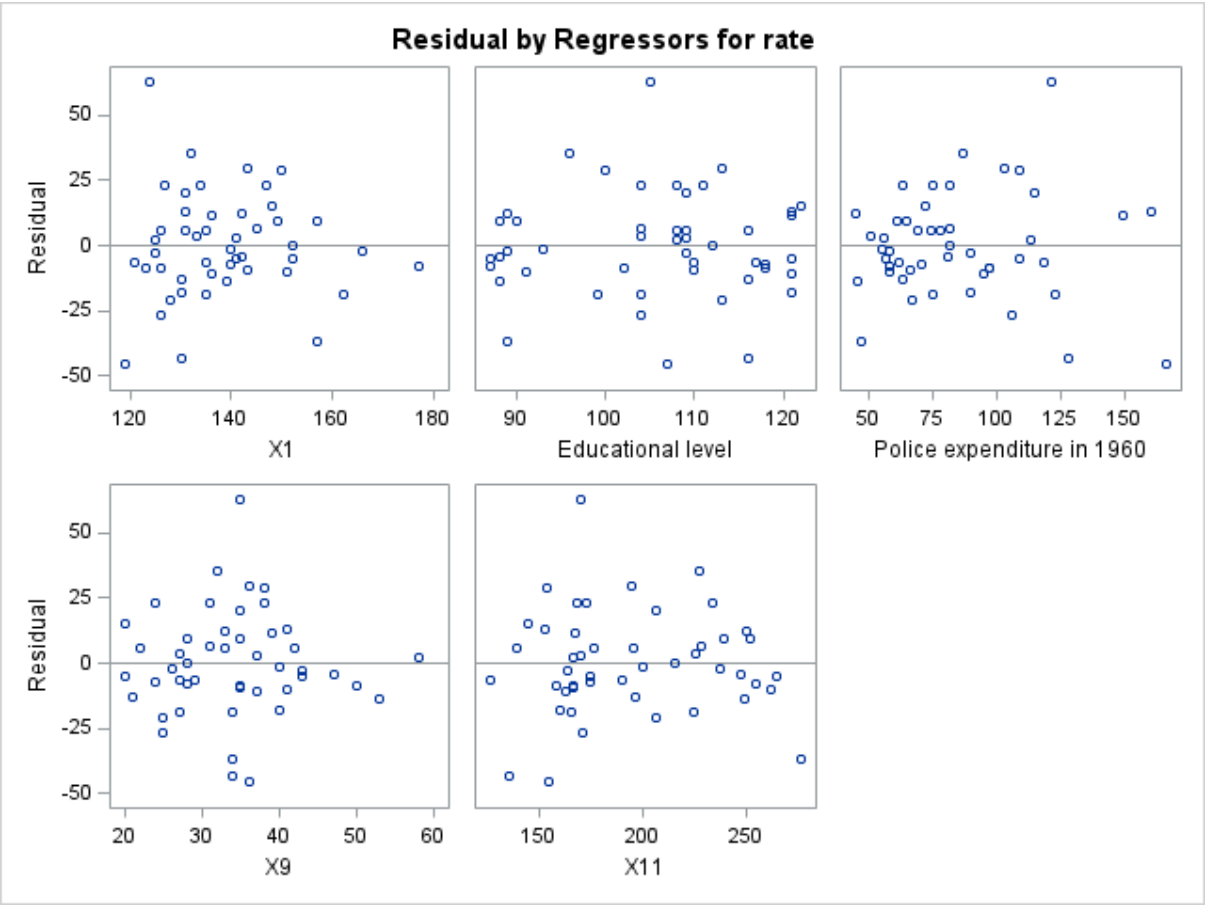
Step	Variable Entered	Variable Removed	Label	Number Vars In	Partial R-Square	Model R-Square	C(p)	F Value	Pr > F
1	X4		Police expenditure in 1960	1	0.4728	0.4728	35.4276	40.36	<.0001
2	X11		Income inequality expressed as the number of families per 1000 earning below one-half of the median income.	2	0.1075	0.5803	21.4331	11.27	0.0016
3	X3		Educational level	3	0.0853	0.6656	10.7413	10.97	0.0019
4	X1		Number of males aged 14-24 years per 1000 of total state population	4	0.0348	0.7004	7.5655	4.88	0.0327

Summary of Stepwise Selection

Step	Variable Entered	Variable Removed	Label	Number Vars In	Partial R-Square	Model R-Square	C(p)	F Value	Pr > F
5	X9		Unemployment rate of urban males per 1000 in the age group 35-39 years	5	0.0292	0.7296	5.2202	4.43	0.0415

The REG Procedure
Model: MODEL3
Dependent Variable: rate Crime Rate defined as the number of offences known to the police per 1,000,000 population




Residual by Regressors for rate
Residual
50
25
0
-25
-50
120
140
160
180
X1
90
100
110
120
Educational level
50
75
100
125
150
Police expenditure in 1960
Residual
50
25
0
-25
-50
20
30
40
50
60
X9
150
200
250
X11

8.6 Appendix F – ANOVA for full model

Source	DF	Sum of Squares	Mean Square	F Value	P-value
(Full)Model	11	52620	4783.65615	10.34	<.0001
Error	35	16189	462.54454		
Total	46	68809			

8.7 Appendix G – ANOVA for summarised model

Source	DF	SS	MS	F
Fit X4	1	32533	32533	78.74
Additional benefit of X11	1	7398	7398	17.89
Additional benefit of X3	1	5871	5871	14.19
Additional benefit of X1	1	2394	2394	5.79
Additional benefit of X9	1	2010	2010	4.86
Residual	45	18603	413.4	
Total	46	68809		

Calculation of SS (Sum of Squares):

Fit X4: SS(X4) = Regression SS(X4) = 32533
Additional benefit of X11: SS(X11) = Regression SS(X11) - Regression SS(X4) = 39931 - 32533 = 7398
Additional benefit of X3: SS(X3) = Regression SS(X3) - Regression SS(X11) = 45802 - 39931 = 5871
Additional benefit of X1: SS(X1) = Regression SS(X1) - Regression SS(X3) = 48196 - 45802 = 2394
Additional benefit of X9: SS(X9) = Regression SS(X9) - Regression SS(X1) = 50206 - 48196 = 2010
Residual: SS(Residual) = Total SS - Regression SS(X9) = 68809 - 50206 = 18603

Calculation of MS (Mean Square):

Fit X4: MS(X4) = SS(X4) / df(X4) = 32533 / 1 = 32533
Additional benefit of X11: MS(X11) = SS(X11) / df(X11) = 7398 / 1 = 7398
Additional benefit of X3: MS(X3) = SS(X3) / df(X3) = 5871 / 1 = 5871
Additional benefit of X1: MS(X1) = SS(X1) / df(X1) = 2394 / 1 = 2394
Additional benefit of X9: MS(X9) = SS(X9) / df(X9) = 2010 / 1 = 2010
Residual: MS(Residual) = SS(Residual) / df(Residual) = 18603 / 45 = 413.4

Calculation of F:

Fit X4: F(X4) = MS(X4) / MS(Residual) = 32533 / 413.4 = 78.75 (approximately)
Additional benefit of X11: F(X11) = MS(X11) / MS(Residual) = 7398 / 413.4 = 17.89 (approximately)
Additional benefit of X3: F(X3) = MS(X3) / MS(Residual) = 5871 / 413.4 = 14.19 (approximately)
Additional benefit of X1: F(X1) = MS(X1) / MS(Residual) = 2394 / 413.4 = 5.79 (approximately)
Additional benefit of X9: F(X9) = MS(X9) / MS(Residual) = 2010 / 413.4 = 4.86 (approximately)

8.8 Appendix H – SAS code

```
DATA AmericaHasAProblem;
  INPUT rate X1-X11; /*Define variables for crime rate and socio-economic
factors. In this version: without LF and M - does not affect the regression
model*/
  LABEL rate = 'Crime Rate defined as the number of offences known to the
police per 1,000,000 population'
            X1 = 'Number of males aged 14-24 years per 1000 of total state
population'
            X2 = 'Binary variable distinguishing Southern states (S=1) from
the rest'
            X3 = 'Educational level'
            X4 = 'Police expenditure in 1960'
            X5 = 'Police expenditure in 1959'
            X6 = 'State population size in hundred thousand.'
            X7 = 'Number of non-white people per 1000.'
            X8 = 'Unemployment rate of urban males per 1000 in the age
group 14-24 years'
            X9 ='Unemployment rate of urban males per 1000 in the age group
35-39 years'
            X10 ='Wealth as measured by family income (unit: 10 dollars)'
            X11 = 'Income inequality expressed as the number of families
per 1000 earning below one-half of the median income.';
  DATALINES;
79.1  151   1     91    58    56                    33    301   108   41
      394   261
163.5 143   0     113   103   95                    13    102   96    36
      557   194
57.8  142   1     89    45    44                    18    219   94    33
      318   250
196.9 136   0     121   149   141                   157   80    102   39
      673   167
123.4 141   0     121   109   101                   18    30    91    20
      578   174
68.2  121   0     110   118   115                   25    44    84    29
      689   126
96.3  127   1     111   82    79                    4     139   97    38
      620   168
155.5 131   1     109   115   109                   50    179   79    35
      472   206
85.6  157   1     90    65    62                    39    286   81    28
      421   239
70.5  140   0     118   71    68                    7     15    100   24
      526   174
167.4 124   0     105   121   116                   101   106   77    35
      657   170
84.9  134   0     108   75    71                    47    59    83    31
      580   172
51.1  128   0     113   67    60                    28    10    77    25
      507   206
66.4  135   0     117   62    61                    22    46    77    27
      529   190
79.8  152   1     87    57    53                    30    72    92    43
      405   264
94.6  142   1     88    81    77                    33    321   116   47
      427   247
53.9  143   0     110   66    63                    10    6     114   35
      487   166
```

```
92.9  135   1      104   123   115                    31    170   89    34
      631   165
75.0  130   0      116   128   128                    51    24    78    34
      627   135
122.5 125   0      108   113   105                    78    94    130   58
      626   166
74.2  126   0      108   74    67                     34    12    102   33
      557   195
43.9  157   1      89    47    44                     22    423   97    34
      288   276
121.6 132   0      96    87    83                     43    92    83    32
      513   227
96.8  131   0      116   78    73                     7     36    142   42
      540   176
52.3  130   0      116   63    57                     14    26    70    21
      486   196
199.3 131 0 121    160   143                    3      77    102   41    674
      152
34.2  135   0      109   69    71                     6     4     80    22
      564   139
121.6 152   0      112   82    76                     10    79    103   28
      537   215
104.3 119   0      107   166   157                    168   89    92    36
      637   154
69.6  166   1      89    58    54                     46    254   72    26
      396   237
37.3  140   0      93    55    54                     6     20    135   40
      453   200
75.4  125   0      109   90    81                     97    82    105   43
      617   163
107.2 147 1 104    63    64                     23     95    76    24    462
      233
92.3  126   0      118   97    97                     18    21    102   35
      589   166
65.3  123   0      102   97    87                     113   76    124   50
      572   158
127.2 150   0      100   109   98                     9     24    87    38
      559   153
83.1  177   1      87    58    56                     24    349   76    28
      382   254
56.6  133   0      104   51    47                     7     40    99    27
      425   225
82.6  149   1      88    61    54                     36    165   86    35
      395   251
115.1 145   1      104   82    74                     96    126   88    31
      488   228
88.0  148   0      122   72    66                     9     19    84    20
      590   144
54.2  141   0      109   56    54                     4     2     107   37
      489   170
82.3  162   1      99    75    70                     40    208   73    27
      496   224
103.0 136   0      121   95    96                     29    36    111   37
      622   162
45.5  139   1      88    46    41                     19    49    135   53
      457   249
50.8  126 0 104    106   97                     40     24    78    25    593
      171
84.9  130   0      121   90    91                     3     22    113   40
      588   160
RUN;
```

```
PROC GPLOT DATA=AmericaHasAProblem; /* Generate scatter diagrams with the
response variable y*/
PLOT rate * (X1-X11);
RUN;
PROC CORR DATA=AmericaHasAProblem; /* Compute correlation coefficients and
their respective p-values */
VAR rate X1-X11;
RUN;
PROC REG DATA=AmericaHasAProblem; /* Obtain the best final regression
models from each of the Variable Selection methods */
MODEL rate = X1-X11/SELECTION = FORWARD SLENTRY = 0.05;
MODEL rate = X1-X11/SELECTION = BACKWARD SLENTRY = 0.05;
MODEL rate = X1-X11/SELECTION = STEPWISE SLENTRY = 0.05;
RUN;
```