Statistical Sleuthing through Criminal Models

William Chen, Yuan Jiang, Patrick Xu, Alice Zhao

Stat 139: Statistical Sleuthing through Linear Models

Abstract

For our project, we chose to model and analyze crime rates in specific US municipalities. Given the severity of the topic, analyzing the prevalance of crime in the United States could have a potentially large impact on our quality of life. Starting out with over 140 explanatory variables, we hope to discover any differences in the frequency of violent crimes among states and groups of states through one-way ANOVAs and t-tests. In addition, we will regress crime rate on all significant predictors. To ensure linearity in our model, we will only analyze positive crime values, and we will transform the necessary predictors and response variable. We will chooose the best model based on its adjusted R^2 value, and then evaluate and study the diagnostics of the model.

Multiple Regression Models

We will first do some exploratory analysis and look at plots of all the explanatory varaibles vs. the response variable. The scatterplots suggest that transformations are necessary, so we will log transform our response variable, violent crimes, as well as some of the predictors*. Observing the new set of scatter plots, a rough linear relationship can be observed in all plots, so we will now proceed to create our model. We fill run the following regressions in R: main effects model, main effects incuding all pairwise interactions model, constant mean model, backwards stepwise regression model, and finally, a forwards stepwise regression model. Below are the summary statistics for each model.

Model	$\hat{\sigma}$	R^2	Adjusted R^2
1 Intercept (0)	0.5733	0.0000	0.0000
2 All Pred. (12)	0.5471	0.1222	0.0893
3 All Pred. + Inter. (78)	0.5550	**0.2831**	0.0630
4 **Backwards (35)**	**0.5299**	0.2358	**0.1458**
5 Forwards (5)	0.5384	0.1313	0.1180

Table 1 : Below is a table summarizing the performance of our models.

After observing the adjusted R^2 values, we will choose to work with our backwards stepwise regression model. Our model is as follows:

$$\begin{aligned} E(\log(\textit{crime}) \mid \textit{Predictors}) = & -63.618 - 1.623 \cdot \textit{logpopulation} - 0.302 \cdot \textit{logracepctblack} + 0.001 \cdot \textit{racePctWhite} \\ & + 6.963 \cdot \textit{logmedIncome} + 0.195 \cdot \textit{PctPopUnderPov} - 7.112 \cdot \textit{logPctLess9thGrade} \\ & - 1.964 \cdot \textit{logPctUnemployed} + 4.678 \cdot \textit{PctFam2Par} - 3.848 \cdot \textit{PctKids2Par} \\ & - 4.692 \cdot \textit{logPctIlleg} + 3.651 \cdot \textit{logNumImmig} - 6.341 \cdot \textit{logPctWOFullPlumb} \\ & + 0.016 \cdot \textit{logpopulation} \cdot \textit{PctKids2Par} + 0.222 \cdot \textit{logpopulation} \cdot \textit{logPctIlleg} \\ & + 0.252 \cdot \textit{logracepctblack} \cdot \textit{logmedIncome} - 0.023 \cdot \textit{logracepctblack} \cdot \textit{PctFam2Par} \\ & - 0.072 \cdot \textit{logracepctblack} \cdot \textit{logNumImmig} - 0.003 \cdot \textit{racePctWhite} \cdot \textit{PctFam2Par} \\ & + 0.004 \cdot \textit{racePctWhite} \cdot \textit{PctKids2Par} + 0.016 \cdot \textit{racePctWhite} \cdot \textit{logPctIlleg} \\ & - 0.006 \cdot \textit{racePctWhite} \cdot \textit{logNumImmig} + 0.016 \cdot \textit{racePctWhite} \cdot \textit{logPctWOFullPlumb} \\ & + 0.636 \cdot \textit{logmedIncome} \cdot \textit{logPctLess9thGrade} - 0.392 \cdot \textit{logmedIncome} \cdot \textit{PctFam2Par} \\ & + 0.309 \cdot \textit{logmedIncome} \cdot \textit{PctKids2Par} - 0.245 \cdot \textit{logmedIncome} \cdot \textit{logNumImmig} \\ & + 0.676 \cdot \textit{logmedIncome} \cdot \textit{logPctWOFullPlumb} + 0.065 \cdot \textit{PctPopUnderPov} \cdot \textit{logPctUnemployed} \\ & - 0.020 \cdot \textit{PctPopUnderPov} \cdot \textit{PctFam2Par} + 0.016 \cdot \textit{PctPopUnderPov} \cdot \textit{PctKids2Par} \\ & + 0.480 \cdot \textit{logPctLess9thGrade} \cdot \textit{logPctIlleg} + 0.047 \cdot \textit{logPctUnemployed} \cdot \textit{PctKids2Par} \\ & - 0.235 \cdot \textit{logPctUnemployed} \cdot \textit{logNumImmig} - 0.002 \cdot \textit{PctFam2Par} \cdot \textit{PctKids2Par} \\ & - 0.029 \cdot \textit{PctFam2Par} \cdot \textit{logPctWOFullPlumb} \end{aligned}$$

Primary Model Evaluation and Diagnostics

The assumptions of our model are as follows:

Linearity: There should be a linear relationship between log crime rate and each predictor. If this assumption does not hold, we cannot fit a linear model to the data. When we observe the scatter plot matrix of all the observations, we observe a linear relationship among all the transformed data points.

Independence: There should be independence between the residuals. This assumption may be violated if some cities in the dataset border each other since crime may flow into neighboring municipalities.

Equal Spread: Residuals should have equal spread. This can be determined by the residual plot. If we look at the residual plot, we find that the residuals do indeed have (almost) equal variances.

Normality: Residuals should be distributed normally. This can be observed through a QQ plot and histograms of residuals.

ANOVA and t-tests

We will perform various one-way ANOVA tests to determine whether the crime rates within subgroups are similar or different. For our first ANOVA test, we tested for differences in means using State as the categorical factor. Our hypothesis is as follows:

- H_0 (Equal Means Model): $\mu_1 = \mu_2 = \ldots = \mu_{n-1} = \mu_{50}$ where μ is the mean number of violent crimes per state.
- H_a (Separate Means Model): At least one of the means is different.

After running the ANOVA, we obtain an F-statistic of 2.46 and our reference distribution is F distribution$_{47,2167}$. Our test returns a p-value of $1.814 * 10^{-7}$ so at an α significance level of 0.05, we have strong statistical evidence to reject the null hypothesis and conclude that at least one of the state means is different. Intuitively, this is correct because different states should have different crime rates because some states are more dangerous than others.

Visualizatons

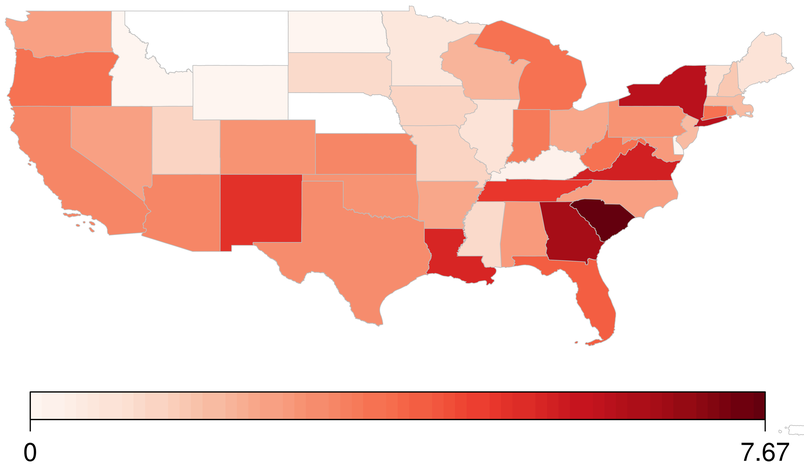

Figure 1 : Population-weighted cartogram of crime rates for each state in the US. South Carolina, New York, and Georgia have the the highest crime rates, as indicated by the intensity of the red.

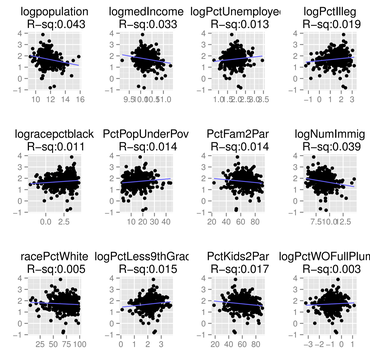

Figure 2 : Scatterplots of each transformed predictor (x-axis) vs. log crime (y-axis). A rough linear relationship can be observed an all plots, satisfying the linearity assumption.

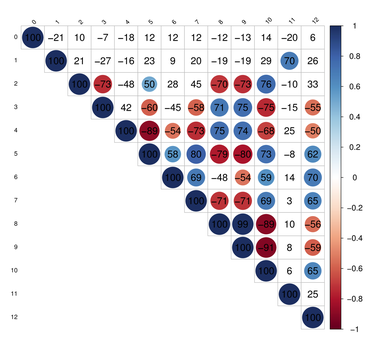

Figure 3 : Correlation matrix for all tranformed predictors. Due to high correlations between multiple predictors, we will include all pairwise interaction terms for our stepwise regressions. *

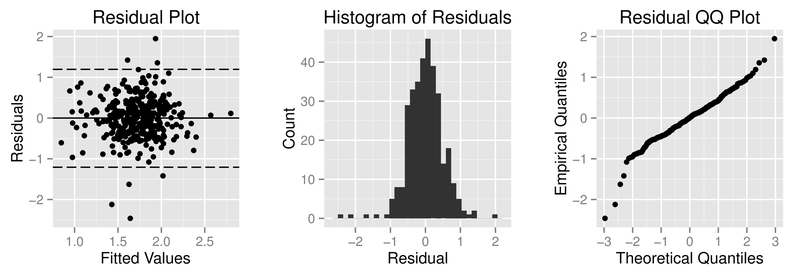

Figure 4 : Residual diagnostics. From left to right: Residual plot exhibits no patterns; while outliers do exist in our data, they do not demonstrate high leverage (as confirmed by Cook's distance plot). Our histogram of residuals demonstrate that residuals are roughly normally distributed. Our QQ plot further supports that our residuals are roughly normally distributed.

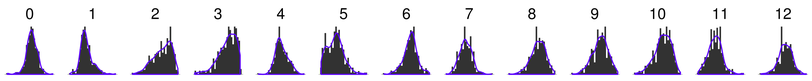

Figure 5 : Histograms of each predictor. Figures show that predictor is roughly normally distributed, though not perfectly so. This is ok because we have many data points. *

* Legend: 0 logviolentcrime 1 logpopulation 2 logracepctblack 3 racePctWhite 4 logmedIncome 5 PctPopUnderPov 6 logPctLess9thGrade 7 logPctUnemployed 8 PctFam2Par 9 PctKids2Par 10 logPctIlleg 11 logNumImmig 12 logPctWOFullPlumb

ANOVA and t-tests (Continued)

We will now test the intermediate vs. separate means model. We are particularly interested in factors like population, population density, wealth, education levels, and percentage of single parents (%SP). We will run ANOVA tests taking the top 10 states with those factors to test if that subgroup has a different mean. Below are all the hypotheses for these 5 ANOVA tests:

Top 10 vs Bottom 40 by Population

$H_0 : \mu_1 = \mu_2 = \ldots = \mu_{10}; \mu_{11} = \mu_{12} = \ldots = \mu_{50}$.

$H_a : \mu_1, \mu_2, \ldots, \mu_{50}$

Top 10 vs Bottom 40 by Population Density

$H_0 : \mu_1 = \mu_2 = \ldots = \mu_{10}; \mu_{11} = \mu_{12} = \ldots = \mu_{50}$.

$H_a : \mu_1, \mu_2, \ldots, \mu_{50}$

Top 10 vs Bottom 40 by Wealth

$H_0 : \mu_1 = \mu_2 = \ldots = \mu_{10}; \mu_{11} = \mu_{12} = \ldots = \mu_{50}$.

$H_a : \mu_1, \mu_2 \ldots \mu_{50}$

Top 10 vs Bottom 40 by Education

$H_0 : \mu_1 = \mu_2 = \ldots = \mu_{10}; \mu_{11} = \mu_{12} = \ldots = \mu_{50}$.

$H_a : \mu_1, \mu_2, \ldots, \mu_{50}$

Top 10 vs Bottom 40 by %SP

$H_0 : \mu_1 = \mu_2 = \ldots = \mu_{10}; \mu_{11} = \mu_{12} = \ldots = \mu_{50}$.

$H_a : \mu_1, \mu_2, \ldots, \mu_{50}$

Identifying Factor	F-Stat	P-value	Conclusion
Population	2.49	$1.6 * 10^{-7}$	Separate Means
Pop. Density	2.51	$1.175 * 10^{-7}$	Separate Means
Wealth	2.51	$1.183 * 10^{-7}$	Separate Means
Education	2.50	$1.20 * 10^{-7}$	Separate Means
Single Parents	2.34	$1.21 * 10^{-6}$	Separate Means

Above are the results of our ANOVA tests. We observe in all cases that the p-value < 0.05, so we reject the null hypothesis based on each category and conclude that at least one mean in the top 10 subgroup differs.

With our t-tests, we decided to split our data into two different subsets to compare the means of these two subsets. The comparisons we chose were CA vs. WI, Northern States vs. Southern States, Western States vs. Eastern States, and Democratic States vs. Republican States (As chosen by the last presidential election). In all four comparisons, we deemed the variances of the samples in question to be close enough so that we could use the pooled t-test. The results and conclusions of our tests are provided below.

ANOVA and t-test Results

ANOVA	df	Sum of Squares	Mean Square	F-stat	Pr ($> F$)
factor(state)	47	55.49	1.18055	2.4602	1.814×10^{-7}
Residuals	2167	1039.84	0.47985		

Table 2 : One-way ANOVA Table

Sample 1	Sample 2	Var 1	Var 2	Pooled?	P-value	Conclusion
CA	WI	0.563	0.356	Yes	0.1218	Fail to Reject
North	South	0.404	0.596	Yes	0.0447	Reject (South $>$ North)
West	East	0.477	0.505	Yes	0.9253	Fail to Reject
Red	Blue	0.589	0.441	Yes	0.0161	Reject (Red $>$ Blue)

Table 3 : t-test Results and Conclusions

State	Var
CA	0.56
NY	0.62
MI	0.59
TN	0.41
TX	0.48
RI	0.60
WI	0.36
...	...

Table 4 : Select state variances

Conclusions and Future Work

Throughout our study, we noticed numerous improvements that could be undertaken in future studies to develop more holistic and accurate conclusions about our data. For our multiple linear regression model, we found that the most significant single predictors (in terms of lowest p-value) were log population, log med income, percent families with 2 parents, and log percent illegal immigrants. These predictors especially are worth further study. Our adjusted R^2 value was seemingly low, which indicates that there is a lot of unexplained noise. To improve the accuracy of our model, we could include more interaction terms, try other transformations, or examine the leverage of our outliers (if high leverage, remove).

With respect to our t-tests, we realized that some of our sample sizes were drastically different. Additionally, we acknowledge that a vast majority of our data points had a value of 0 for its violent crimes per population rate, which can skew our conclusions. For our actual results, we noted that there was no association in crime rates between California and Wisconsin and between Western and Eastern states. However, we did see differences between Red and Blue states, as well as North and South states. Due to these findings, an interesting future study might be focused on analyzing crime rates between conservatives and liberals, which could lead to interesting conclusions in not only statistics, but other fields, such as psychology, as well.

For our ANOVA tests, we only assumed equal variances among the states, because it would be incredibly difficult to test for equality among all the states. We did, however, run a check through some of the variances, and they seemed relatively similar. Unfortunately, with such a large dataset, it was extremely easy to reject the null hypothesis and say that the separate means model would be preferable. If possible, future studies could look into generating more accurate and revealing models and tests.

One biohazard that we do not want to fall victim to is ecological fallacy. Just because we have developed conclusions for populations does not mean that we can also make conclusions for individuals or subsets of individuals in these populations. Thus, we cannot comment on individual likelihoods of committing crimes, although our study does serve as potential background and foundation for further studies.

We would like to acknowledge and thank Victoria, Jiexing, and Lo-Hua for the help and guidance throughout the class.

Data set obtained from University of California: Irvine, Machine Learning Repository, access it here: http://archive.ics.uci.edu/ml/datasets/Communities+and+Crime