US Census Crime Rate Modeling
MSG500-MVE190 Linear Statistical Models

Stefan Eng (gussteen@student.gu.se)
Gothenburg University/Chalmers

January 2019

Introduction

We are considering a dataset that provides some demographic information for 440 of the most populous counties in the United States in years 1990-92. Each line of the dataset provides information on 14 variables for a single county. Counties with missing data were deleted from the dataset. We are building models to predict the number of crimes per 1000 people in each county. We first explore linear regression models and then a negative binomial regression model. We found that the best linear regression model performed similarly to the negative binomial regression model on the training and test set using the same variables.

Goals

The goals of the analysis was the find a model that is simple yet explains as much as possible. The model should make sense first and foremost. Automatic methods such as the backwards step algorithm were used as auxiliary methods to supplement a more hand selected model. We decided against using all possible subsets selection as a matter of principle as it can lead to models that lack explainability. We explore interactions between variables as well as standard additive models. Once model selection is done based on the training set the final results are reported against the test set (20% of the dataset). The test set was not looked at or used in the model building process. The models were compared on the training set using 10-fold cross validation and leave one out cross validation (LOOCV).

Data Processing

Some additional variables were created. The variables "beds", "phys", and "area" were all divided by the population to give a per capita total number of hospital beds, cribs and bassinets, per capita number of physicians practicing, and the per capita area. We then remove the total quantities from the data set. We found that working with per capita was more informative than the direct quantities. The data was also transformed almost entirely with natural log transform as it performed better for our model. We arrived at natural log based on the plots, residuals plots, and looking how the model's R^2 changed with regard to the transformations along with cross validation. The data summary can be found in the appendix.

Models

Full Model

```
crm1000 ~ percapitaarea + popul + pop1834 + pop65plus +
          percapitaphys + percapitabeds + higrads + bachelors +
          poors + unemployed + percapitaincome + region
```

The full model was used as a baseline model, and subsequent models were reduced from this model.

Base Model

```
crm1000 ~ percapitaarea + popul + pop1834 +
          percapitabeds + poors + percapitaincome + region
```

These are the variables from which all the other models are based. We arrived at this model by keeping one variable from each correlation cluster that seemed to make the most intuitive sense to keep. We confirmed that this was a good model by starting with all of the variables and using backward selection to arrive at the identicial model that we had selected by hand. Then a partial F-test was performed, see Table 1. The F statistic was extremely small, which means that the residual sum of squares was almost the same after dropping the variables `pop65plus`, `percapitaphys`, `higrads`, `bachelors`, and `unemployed`. From this reduced model, we built up variations that involved transformations and interactions.

Table 1: Partial F-Test for Full model vs final model (without transformations)

Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
342	144262.2				
337	143366.5	5	895.74	0.421	0.834

(Final Model) Log Transformed model with no interactions

This model ended up being our final model selection. The transformation were found through cross validation and analysis of the plots of the data.

```
crm1000 ~ log(percapitaarea) + log(popul) + log(pop1834) +
          log(percapitabeds) + log(poors) + log(percapitaincome) + region
```

Transformations and Interactions with only region.

We included all interactions between the continuous variables and the region. We then backward selected that reduced to (log) population and (log) per capita income against the region. While running the backwards selection we checked that the main effects were not dropped while the interactions were kept.

```
crm1000 ~ log(percapitaarea) + log(popul) + log(pop1834) +
          log(percapitabeds) + log(poors) + log(percapitaincome) +
          region + log(popul):region + log(percapitaincome):region
```

Transformations and Interactions against all variables.

We included the interactions between all variables (including continuous-continuous) then backward selected. Both of the interactions models included `log(popul):region` but the larger interaction model did not include `log(percapitaincome):region`. We suspect this may be from the correlation between per capita income and some of the other variables that are included such as the percentage of poor people.

```
crm1000 ~ log(percapitaarea) + log(popul) + log(pop1834) +
          log(percapitabeds) + log(poors) + log(percapitaincome) +
          region + log(popul):log(percapitabeds) +
          log(popul):log(poors) + log(popul):log(percapitaincome) +
          log(popul):region + log(pop1834):region +
          log(percapitabeds):region + log(poors):log(percapitaincome) +
          log(poors):region
```

Negative Binomial Generalized Linear Model

We used the same variables from the log transformed model with no interactions. Instead of modeling crime rate per 1000 people directly we model the crimes. We use an offset to adjust the parameters according to the population (more is described in Negative Binomial Model section).

```
crimes ~ offset(log(popul / 1000)) + log(percapitaarea) +
        log(popul) + log(pop1834) + log(percapitabeds) +
        log(poors) + log(percapitaincome) + region
```

Other models

For our cross-validation and model selection we included two models which we did not intend to use but only for comparision purposes. They were the full model which was described above. Also, we included a model which we call the *simple model* which only included the natural log of the percentage of poor people in the county.

Multicollinearity

In this dataset there are many variables that are highly correlated (see appendix for correlation plot). We found that the percentage of poor people in a county was highly correlated with the per capita income and the number of high school grads. If we include all of the variables with high multicollinearity the standard errors will be increased and some of the variables may not end up being significant. We used the Variance inflation factor (VIF) to measure the multicollinearity in our model. The VIF gives us a measure of how much the coefficient's variance will increase due to collinearity between the other variables [2]. For each variable i, we calculate the VIF by performing i regressions using all other variables to predict x_i.

$$x_i = \beta_0 + \beta_1 x_1 + \cdots + \beta_{i-1} x_{i-1} + \beta_{i+1} x_{i+1} + \cdots \beta_{p-1} x_{p-1}$$

Then the VIF factor is equal to

$$VIF_i = \frac{1}{1 - R_i^2}$$

Where R_i^2 is the R^2 when using all other variables in a regression on variable i [4]. We removed the variables for per capita income, the number of high school graduates, and the number of bachelor degree graduates based on the VIF results. The last column in Table 2 shows how much the standard error of the variable is increased due to the collinearity. For example, for the bachelors variable the value of 2.834 means that the standard error for $\beta_{bachelors}$ is 2.834 larger than if the variables were not correlated with each other.

Table 2: Variance inflation factor for each variable

	VIF	DF	VIF^(1/(2*Df))
bachelors	8.030	1	2.834
percapitaincome	5.207	1	2.282
higrads	4.899	1	2.213
poors	4.162	1	2.040
percapitabeds	3.309	1	1.819
percapitaphys	3.022	1	1.738
region	2.847	3	1.191
pop1834	2.742	1	1.656
unemployed	2.157	1	1.469
pop65plus	2.123	1	1.457
percapitaarea	1.553	1	1.246
popul	1.290	1	1.136

Outliers

When looking at the training data we noticed one outlier with respect to the target variable in the training dataset (Kings County in New York). It has a crime rate of approximately 296 per 1000 people. The median for the training set was a crime rate of 53.55 per 1000 people. In the training set we found that Kings County has larger leverage and also has high influence. There are also outliers with respect to the other variables but the amount is reduced when we took the log transform of the data. In the appendix, we go into more detailed look at other outliers based on the studentized residuals and influence.

Interactions

We explored and built models that included interactions between the variables. We found that while some of the interactions were significant, when doing cross validation the additive model without any interactions performed the best. Using the partial F-test, we compared the two interaction models with the additive model. For the interactions model with only region, we did not reject the null hypothesis at the $\alpha = 0.05$ level, as the p-value was 0.06. The value was close enough that we decided to continue to compare with this model as well. The results can be see in Table 3. When we run a partial F-test on the interaction models with more interactions, we get a larger test statistic that allows us to reject the null hypothesis at the $\alpha = 0.05$ level. This can be seen in Table 4.

Table 3: Partial F-test for region interactions model

Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
342	124062.3				
336	119758.2	6	4304.037	2.013	0.063

Table 4: Partial F-test for full interactions model

Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
342	124062.3				
326	105258.6	16	18803.67	3.639842	4e-06

In the conditional plot, in Figure 1 we can see the values of the log of population vs the crime rate for each of the regions. The bottom left panel is the West, bottom right is the northeast, top left is the midwest and top right is the south. There seems to be a slightly smaller slope in the West panel, but overall it is not clear that the interactions are needed.

Our original cross validation was performed on data that was not centered (mean subtracted off). Centering can improve the standard errors and thus the p-values of the estimates but the predictions for the new values are the same. Since our final model does not include the interactions, we decided against centering the data. When building the interaction models we included all of the interactions (for region and then for all variables) and then backward selected. We did this twice: once with the data centered and once without the data being centered. While this does not make a difference for the estimate, the standard errors decrease in the centered model. That made the backwards selection algorithm stop earlier when we centered the data, and thus included more of the interaction variables. Again, at this point we checked to make sure that no interaction effects were included when a main effect was dropped. When including these centered models in our cross validation we found that they performed worse than the backward selection models on the original (non-standardized) data so we chose not to include them in our analysis.

Figure 1: Conditional Plot of crime rate, for log(popul) given region
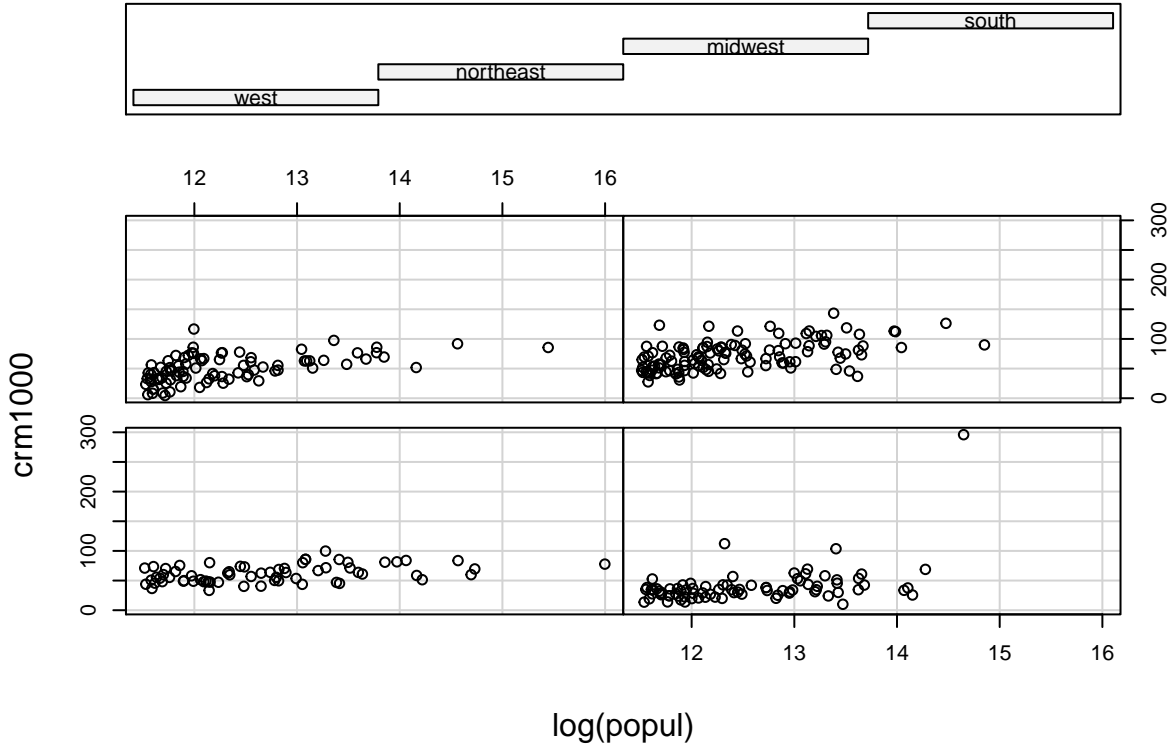
Model Interpretation

On the training data, we had an R^2 of 0.543 which means that 54.3% of the error is explained by our model. The R^2_{adj} was 0.531.

Table 5: Linear regression model coefficients and p-values (Training Data)

	Estimate	Std. Error	Pr(>|t|)
(Intercept)	-372.847	108.328	0.001
log(percapitaarea)	-5.127	1.494	0.001
log(popul)	5.725	1.970	0.004
log(pop1834)	17.534	7.813	0.025
log(percapitabeds)	3.625	2.454	0.141
log(poors)	25.475	3.932	0.000
log(percapitaincome)	24.893	9.948	0.013
regionnortheast	-18.548	3.830	0.000
regionmidwest	-10.724	3.761	0.005
regionsouth	4.974	3.453	0.151

Let's look at what these estimates actually mean. Since we are working on a log scale we can interpret the estimate β_{area} for log(percapitaarea) as the change in crime rate per 1000, crm_{1000}, when $log(percapitaarea)$ increases by 1. That is, $\ln x_{area} + 1 = \ln(e \cdot x_{area})$. So if x_{area} is *multiplied* by $e \approx 2.718$, then crm_{1000} increases by β_{area}. It can be easier to interpret if we look at percentage increase instead of multiplying by e. If the per captia area increases by 10%, then the crime rate per 1000 people will *decrease* by 0.49. $\beta_{area} \cdot ln(1.10) = -5.127 \cdot ln(1.10) = -0.49$. This is summarized in Table 6.

Table 6: Change in crime rate by percentage increase

	5%	10%	20%	30%
percapitaarea	-0.250	-0.489	-0.935	-1.345
popul	0.279	0.546	1.044	1.502
pop1834	0.855	1.671	3.197	4.600
percapitabeds	0.177	0.346	0.661	0.951
poors	1.243	2.428	4.645	6.684
percapitaincome	1.215	2.373	4.538	6.531

From this table we can see that if the the percentage of poor people increased from 20 to 22, we would expect to see an increase of 2.428 in the number of crimes per 1000 people. We can also see that if the per captia area increases from 5×10^{-3} to 5.5×10^{-3} we would expect the number of crimes per 1000 people to *drop* by 0.489.

Region

We used "West" as our reference category so all the parameter estimates are in relation to the west region. We can interpret the estimate $\beta_{region_{NE}} = -18.548$ as the estimate change in crimes per 1000 people between the west and the north east. That is, holding all else constant, the north-east region is estimated to have 18.548 less crimes per 1000 people. Both the north-east and the midwest are statistically significant at the $\alpha = 0.05$ level. The difference between the south and the west's rate of crimes is not statistically significant. That means that the there is not enough evidence from the data to show there is a difference in crime rates between the south and the west (assuming all else is held constant).

Negative Binomial Model

We also explored generalized linear models for predicting the crime rate. Since we are dealing with the rate $1000 * crimes/popul$, we can formulate our model as

$$log(crimes) = log(popul/1000) + X\beta$$

Since we are using the (log) population in our model, having an offset is equivalent to the value of the coefficient for log(population) will increased by 1, and the intercept decreased by log(1000). We first tested a Poisson regression which assumed that the variance and expected value are the same for the crimes. When performing a dispersion test at $\alpha = 0.05$, we rejected the null that dispersion is equal to one. (Estimated dispersion was 2622.163). When dispersion is greater than one, we say that the data is overdispersed. The negative binomial model works better in the presence of overdispersion.

Interpreting Results

Table 7: Negative Binomial Coefficient and p-values

	Estimate	p-value
(Intercept)	-3.569	0.052
log(percapitaarea)	-0.072	0.004
log(popul)	0.096	0.004
log(pop1834)	0.336	0.011

Table 7: Negative Binomial Coefficient and p-values

	Estimate	p-value
log(percapitabeds)	0.097	0.020
log(poors)	0.414	0.000
log(percapitaincome)	0.477	0.005
regionnortheast	-0.468	0.000
regionmidwest	-0.243	0.000
regionsouth	0.043	0.462

We can interpret the results from negative binomial similarly to Poisson regression [1] [3]. Let's look at how the percentage of poor people in a county influences the number of crimes. We have that

$$\log(crimes) = \beta_0 + \beta_{poors} \log x_{poors} + \cdots$$

or

$$crimes = \exp(\beta_0 + \beta_{poors} \log x_{poors} + \cdots) = e^{\beta_0} x_{poors}^{\beta_{poors}} \cdots$$

Now if we want to see what happens as x_{poors} changes, say to x'_{poors}, then we have that

$$\frac{crimes'}{crimes} = \left(\frac{x'_{poors}}{x_{poors}}\right)^{\beta_{poors}}$$

That is, the percentage change in the crimes is equal to the percentage change in percentage of poor people to the power of the coefficient for poors. Equivalently we pose this additively as

$$\log crimes' - \log crimes = \beta_{poors} \left(\log x'_{poors} - \log x_{poors}\right)$$

So if we kept all other variables constant, and increased x_{poors} from 10 to 12 we would expect:

$$\frac{crimes'}{crimes} = \left(\frac{12}{10}\right)^{0.414} = 1.078$$

That is, we would expect the crimes to increase by 7.8%. We can confirm this for the predicted crimes per 1000 people while holding all other variables at the means. When $x_{poors} = 10$, then the predicted crime rate is 71.68. The predicted crime from for $x_{poors} = 12$ is 77.3. The ratio of these values is 1.078 which is what we found from the coefficient interpretation.

It is also useful to look at how the predicts change as one of the variables varies. We set all of the variables involved in the regression to the means and then only varied the percentage of poor from 1 to 37, which is approximately the range found in the training set. It is easier to see how the crime rate changes from this graph due to the poor percentage changing than just the estimated coefficients. This can be seen in Figure 2

Cross Validation

Once we had built our models we used 10-fold cross validation to compare them against each other. We also used LOOCV which resulted in very similar results to the 10-fold cross validation. Once we had included the negative binomial model we switched to only using the 10-fold cross validation. The interactions models had a better adjusted R^2 value and were found with the partial F-test to have at least one parameters that should not be set to 0 (at alpha = 0.05 significance level). We included a model called "Full", which had all of the dependent variables included (without transformation) and a "simple" model which is only using the natural log of the percentage of poor people for each county.

Figure 2: Varying the poor percentage while keeping all other variables at the mean values

Table 8: Average MSE across 10-folds
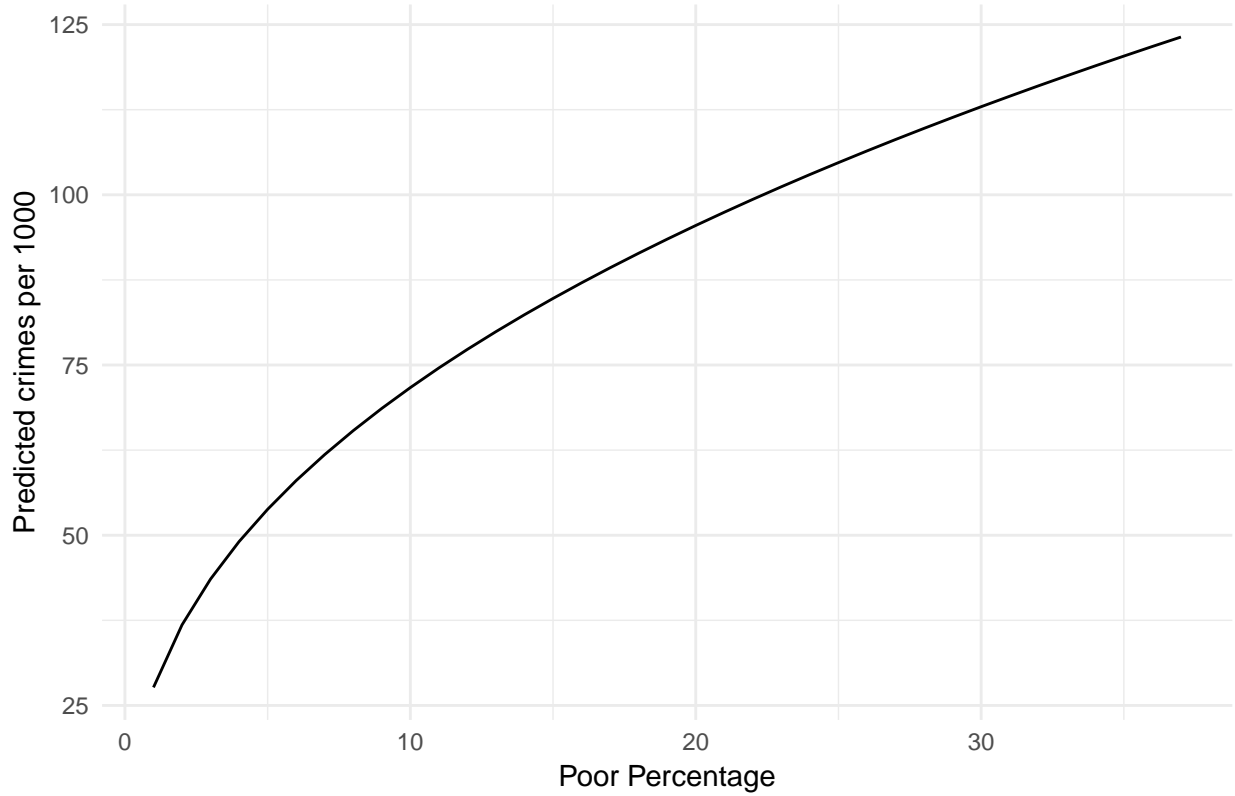
Additive	NB	RegionInter	AllInter	Full	Simple
395.69	415.05	430.86	439.82	500.43	605.53

Test Validation

At this point of the analysis everything had been conducted on the training 80% of the dataset. Cross validation was used as part of the model building process which means that the results will be biased. We kept the test set held out until the end when all models were finalized so that the estimated mean squared error is a better indication of the true mean squared error. The results on the test sets are shown in Table 9 for all of the data, and Table 10 for prediction not trained including Kings county data. See the appendix for reasoning behind dropping Kings county.

Table 9: Test MSE will all data from test included

Additive	NB	RegionInter	AllInter	Full	Simple
235.61	236.57	248.13	269.4	309.62	471.37

Table 10: Test MSE with Kings County excluded

NB	Additive	AllInter	RegionInter	Full	Simple
248.82	252.86	256.69	256.73	310.86	469.59

We can see from that including Kings County (NY) leads to slightly better results in the test set. Based on these results, we can see that both the negative binomial and the additive model without interactions performs the best and almost identically. Since linear regression is faster and simpler to interpret the results our final model is the Log Transformed model with no interactions, in the models section.

Using this as our final model, we can fit this model to the test selection data. This can be seen as a more conservative approach to our parameters since it is on the test set rather than the training.

Table 11: Test set coefficients and p-values

	Estimate	Std. Error	$Pr(>|t|)$
(Intercept)	-385.074	183.899	0.040
log(percapitaarea)	-4.265	2.431	0.083
log(popul)	8.626	3.467	0.015
log(pop1834)	23.087	10.855	0.037
log(percapitabeds)	8.167	4.689	0.086
log(poors)	24.635	6.459	0.000
log(percapitaincome)	23.818	16.421	0.151
regionnortheast	-20.565	7.189	0.005
regionmidwest	-6.943	7.233	0.340
regionsouth	3.083	6.712	0.647

We have $R^2_{adj} = 0.614$ and $R^2 = 0.654$. Which means that 65.42 percent of the error is explained by our model. The p-values in Table 11are the most conservative, since we have not seen any of the data when testing this hypothesis. In this way, it is like a properly run experiment. We do not want to fish for statistically significance by creating hypotheses based on our data set. Instead if we formulate our hypothesis beforehand we should use the data set to validate our hypothesis.

Final Coefficients

Combining the training and test data sets and fitting our final model on this gives us our final coefficients. We want to use all the data if we are going to predict for more counties in the future. We can interpret these estimates exactly the same as for the training and test data. The p-values are artificially low, due to a large sample size and the fact that 80% of the data is from the training set which we used to generated the hypothesis.

Table 12: Final coefficients and p-values

	Estimate	Std. Error	Pr(>|t|)
(Intercept)	-387.885	92.606	0.000
log(percapitaarea)	-5.009	1.276	0.000
log(popul)	6.108	1.697	0.000
log(pop1834)	18.891	6.415	0.003
log(percapitabeds)	4.327	2.163	0.046
log(poors)	25.606	3.345	0.000
log(percapitaincome)	25.910	8.400	0.002
regionnortheast	-18.795	3.311	0.000
regionmidwest	-9.823	3.303	0.003
regionsouth	4.531	3.009	0.133

We find it easier to understand some of these estimates with plots. In Figure 3 we show how the predicted crime rates change when holding all variables constant (keeping the other variables at the mean) while varying only the population. We range the population from the minimum 100,043 to the maximum 8,863,164.

Figure 3: Varying the population while keeping all other variables at the mean values
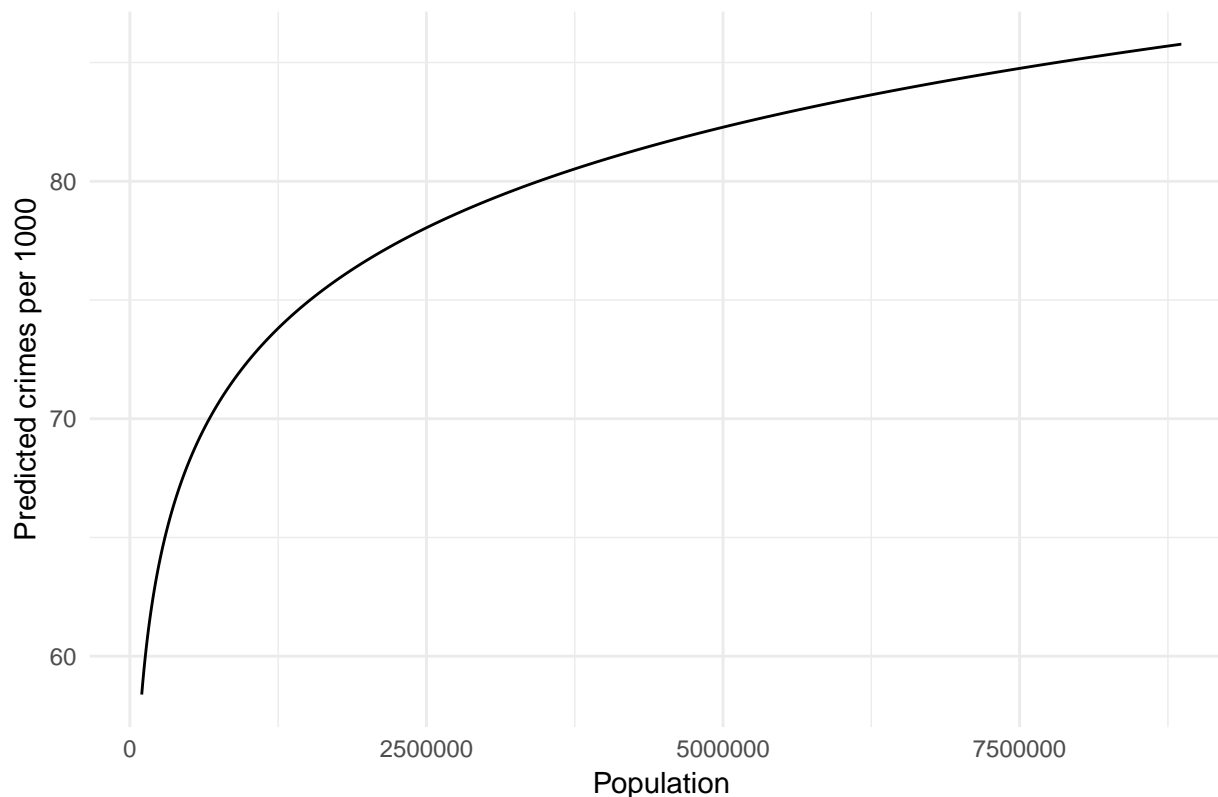
Conclusion

In this analysis we investigated the crime rates for 440 counties in the United States. The data was split into 80% training data and 20% test data. The test data was untouched until all models were finalized.

We compared multiple linear regression models and a negative binomial model and compared the results. The models were compared using 10-fold cross validation as part of the model building process as well as for validation. The most predictive variables in the data set were the population, the percentage of the population between 18 and 34, the per captia beds, the percentage of poor people, the per capita income, and the region in the United States. We showed that the difference between West and Northeast crime rate had a statistically significant difference and well as the difference between West and Midwest.

Appendix

Data Summary

| Variable	Description
id	identification number, 1–440.
county	county name.
state	state abbreviation.
area	land area (square miles).
popul	estimated 1990 population.
pop1834	percent of 1990 CDI population aged 18–34.
pop65plus	percent of 1990 CDI population aged 65 years old or older.
phys	number of professionally active nonfederal physicians during 1990.
beds	total number of hospital beds, cribs and bassinets during 1990.
crimes	total number of serious crimes in 1990 (including murder, rape, robbery, aggravated assault, burglary, larceny-theft, motor vehicle theft).
higrads	percent of adults (25 yrs old or older) who completed at least 12 years of school.
bachelors	percent of adults (25 yrs old or older) with bachelor's degree.
poors	Percent of 1990 CDI population with income below poverty level.
unemployed	percent of 1990 CDI labor force which is unemployed.
percapitaincome	per capita income of 1990 CDI population (dollars).
totalincome	total personal income of 1990 CDI population (in millions of dollars).
region	Geographic region classification used by the U.S. Bureau of the Census, where 1=Northeast, 2 = Midwest, 3=South, 4=West.1
percapitabeds	number of beds (see descriptions above) per capita
percapitaarea	area per capita
percaptiaphys	physicians per capita

Model Diagnostics

Studentized Residuals

Studentized Residuals
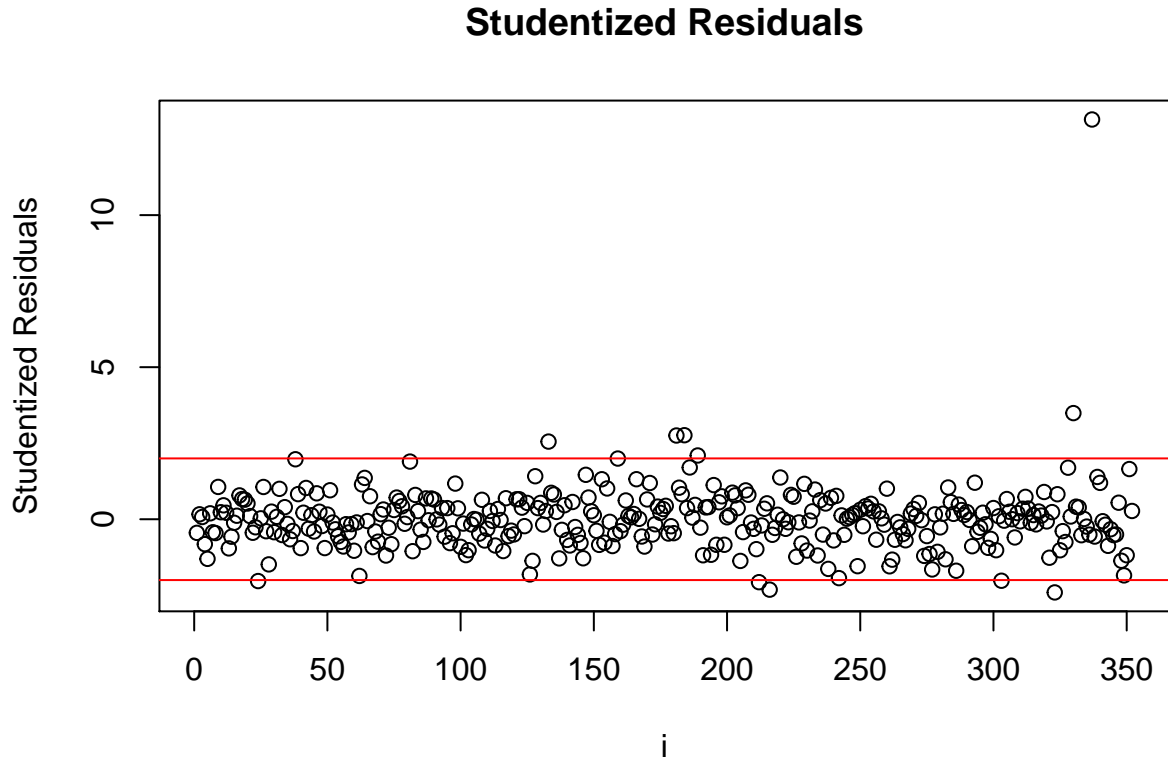
Table 14: Studentized Residuals with red lines at ± 1.96

county	state	Student Residuals
Kings	NY	13.144
Atlantic	NJ	3.490
Wyandotte	KS	2.763
Ector	TX	2.753
Leon	FL	2.553
Delaware	IN	2.406
Jefferson	KY	2.315
Pulaski	AR	2.097
Worcester	MA	2.069
Monroe	IN	2.033
Columbiana	OH	2.022
Shawnee	KS	1.995
Clay	MO	1.972

We can see from the Studentized residuals that there are a few notable points. Kings county in NY is extremely far away from other points. We compared models with and without Kings county but we found that keeping Kings county in the data set had a better test mean squared error.
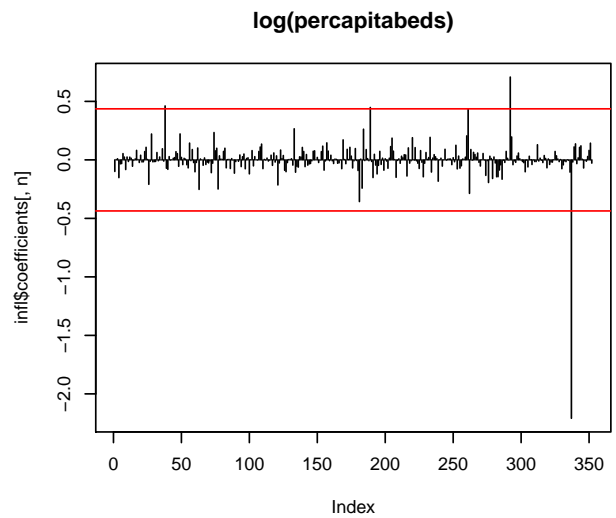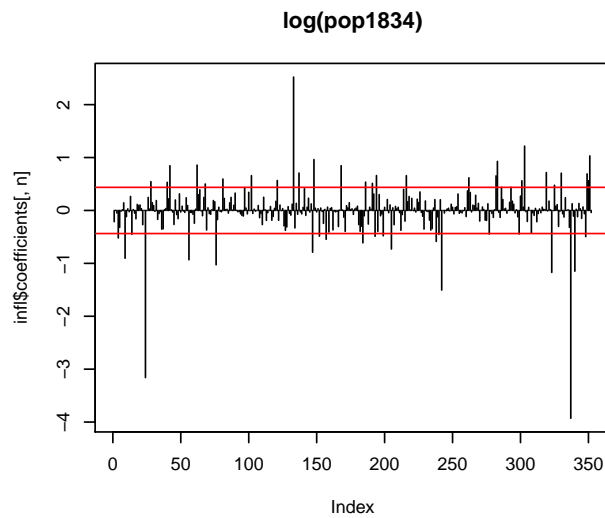
DFBETA on continuous coefficients

log(percapitaarea)

log(popul)

log(pop1834)

log(percapitabeds)

log(poors)

log(percapitaincome)
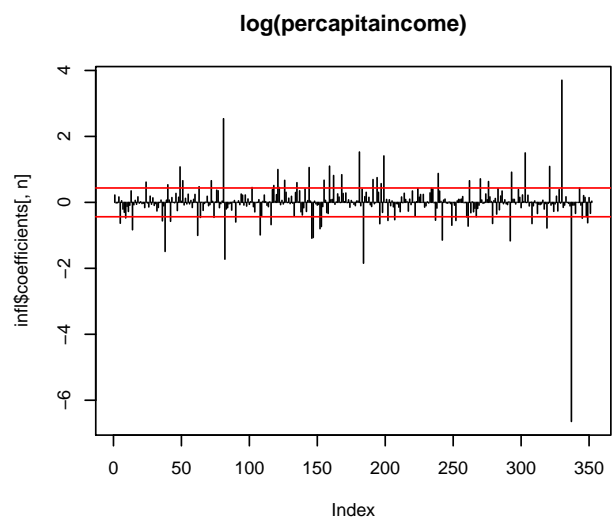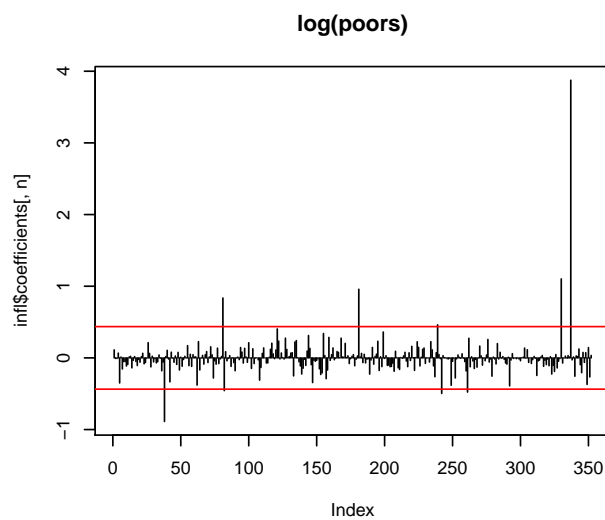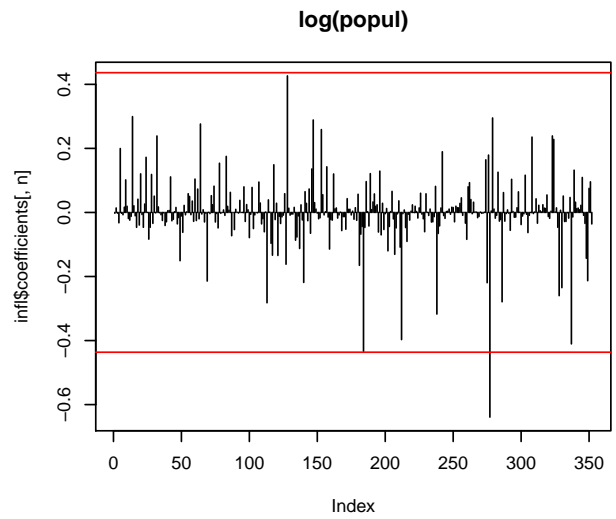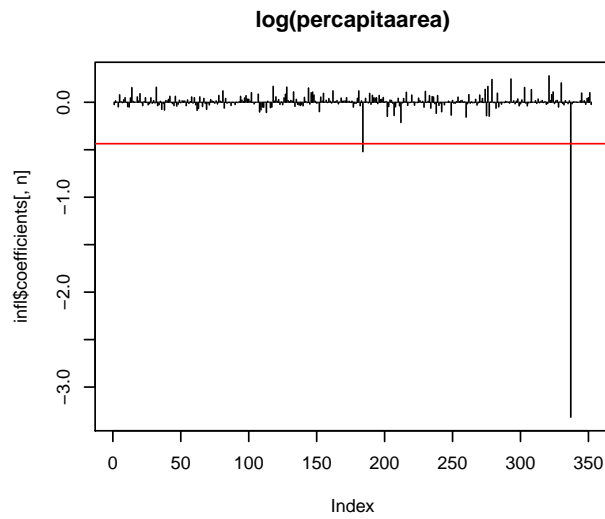
Table 15: High > 2/sqrt(n) DFBETA values

county	coefficient	DFBETA
Kings (NY)	log(percapitaincome)	-6.644
Kings (NY)	log(pop1834)	-3.929
Kings (NY)	log(poors)	3.874
Atlantic (NJ)	log(percapitaincome)	3.704
Kings (NY)	log(percapitaarea)	-3.317
Monroe (IN)	log(pop1834)	-3.160
Fulton (GA)	log(percapitaincome)	2.537
Leon (FL)	log(pop1834)	2.521
Kings (NY)	log(percapitabeds)	-2.209
Wyandotte (KS)	log(percapitaincome)	-1.845
Westchester (NY)	log(percapitaincome)	-1.723
Ector (TX)	log(percapitaincome)	1.527
Pitt (NC)	log(pop1834)	-1.505
Columbiana (OH)	log(percapitaincome)	1.499
Clay (MO)	log(percapitaincome)	-1.488
Utah (UT)	log(percapitaincome)	1.408
Columbiana (OH)	log(pop1834)	1.217
Delaware (IN)	log(pop1834)	-1.171
Fairfax_County (VA)	log(percapitaincome)	-1.167
Manatee (FL)	log(pop1834)	-1.148
Pitt (NC)	log(percapitaincome)	-1.144
Atlantic (NJ)	log(poors)	1.102
Shawnee (KS)	log(percapitaincome)	1.098
Philadelphia (PA)	log(percapitaincome)	1.089
San_Mateo (CA)	log(percapitaincome)	-1.087

Table 15 shows the sorted extreme $DFBETA$ values for each of the coefficients. Take for example Monroe (IN), which has a huge influence on the $pop1834$ parameter. If we look at the data, we see that Monroe has 45.8% of its population between 18 and 34. We see that Kings (NY), influences many of the coefficients and has some of the most extreme values. It is by far the most influential and extreme point in the data.

Final Model Diagnostic Plots
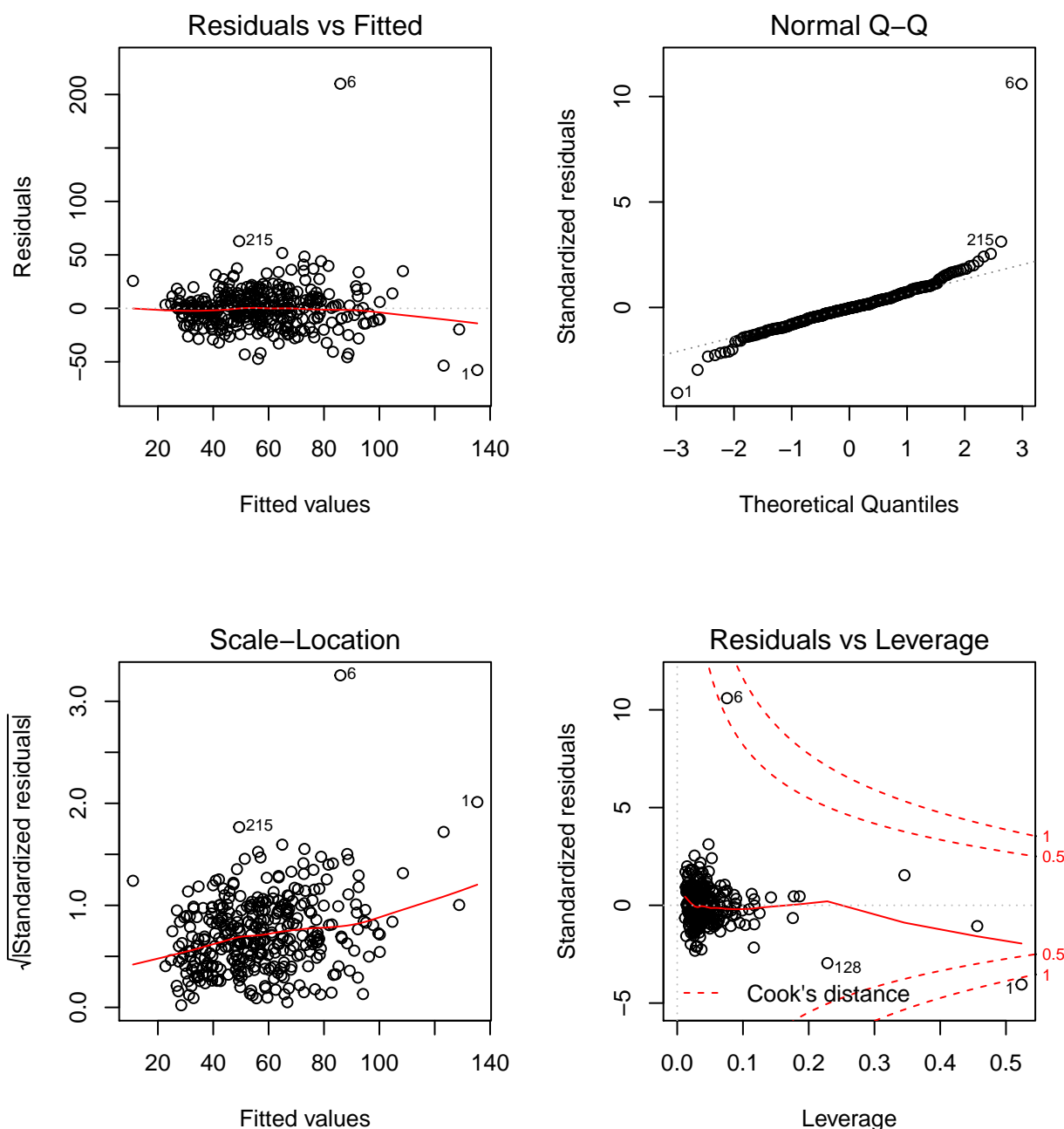
We can see from the diagnostic plots that the residuals and fitted values are approximately symmetric and the residuals don't appear to be increasing or decreasing as the fitted values increases. The normal QQ plot shows that we have larger tails than we would like. The estimates for the parameters do not dependent on it following a normal distribution but our p-values should be interpreted a little pessimistically. The two points that are labeled, 6 (Kings) and 1 (Los Angeles) are outliers. We can see that Los Angeles has a high leverage and also an extreme standardized residual which makes it influential. Kings county is closer to the mean for the dependent variables so it's leverage is lower. It is still marked by the Residuals vs Leverage plot as an issue as its residual is so high.
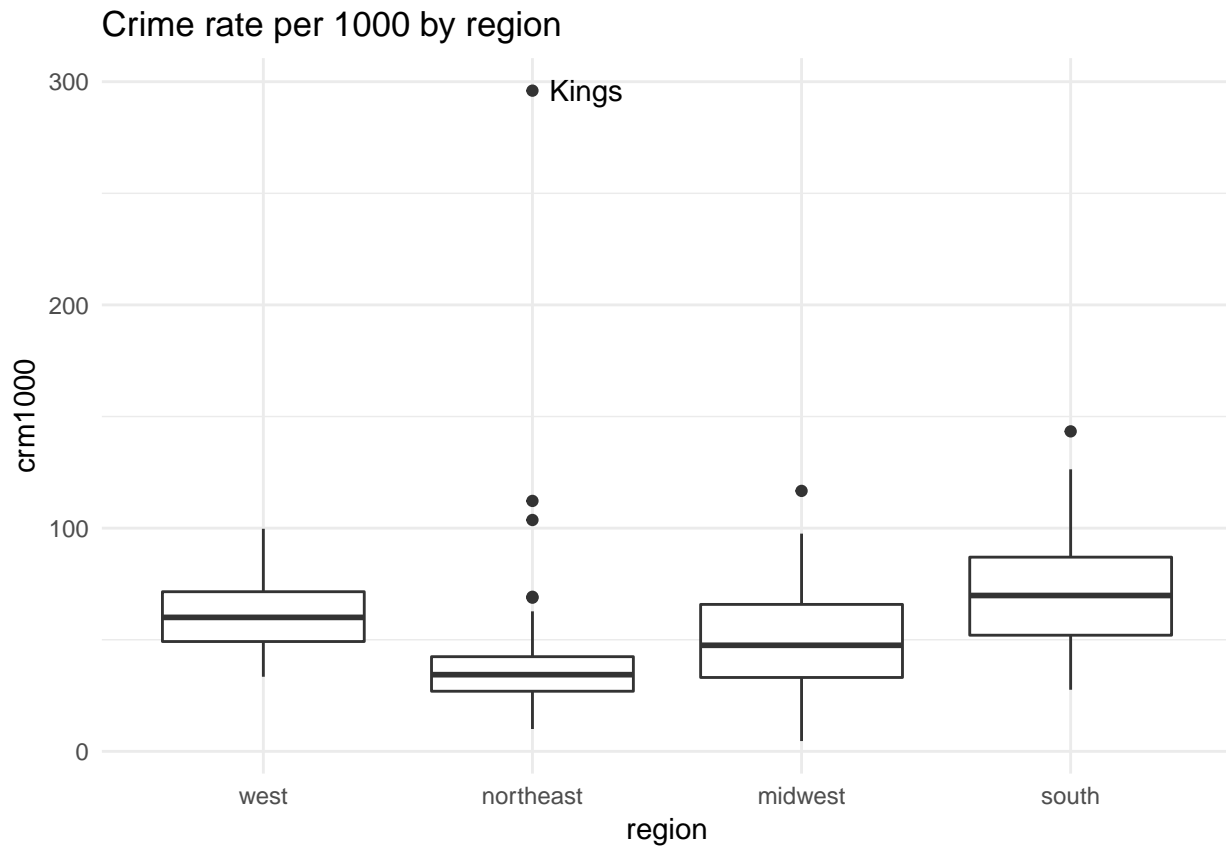
Correlations

County Feature Distances
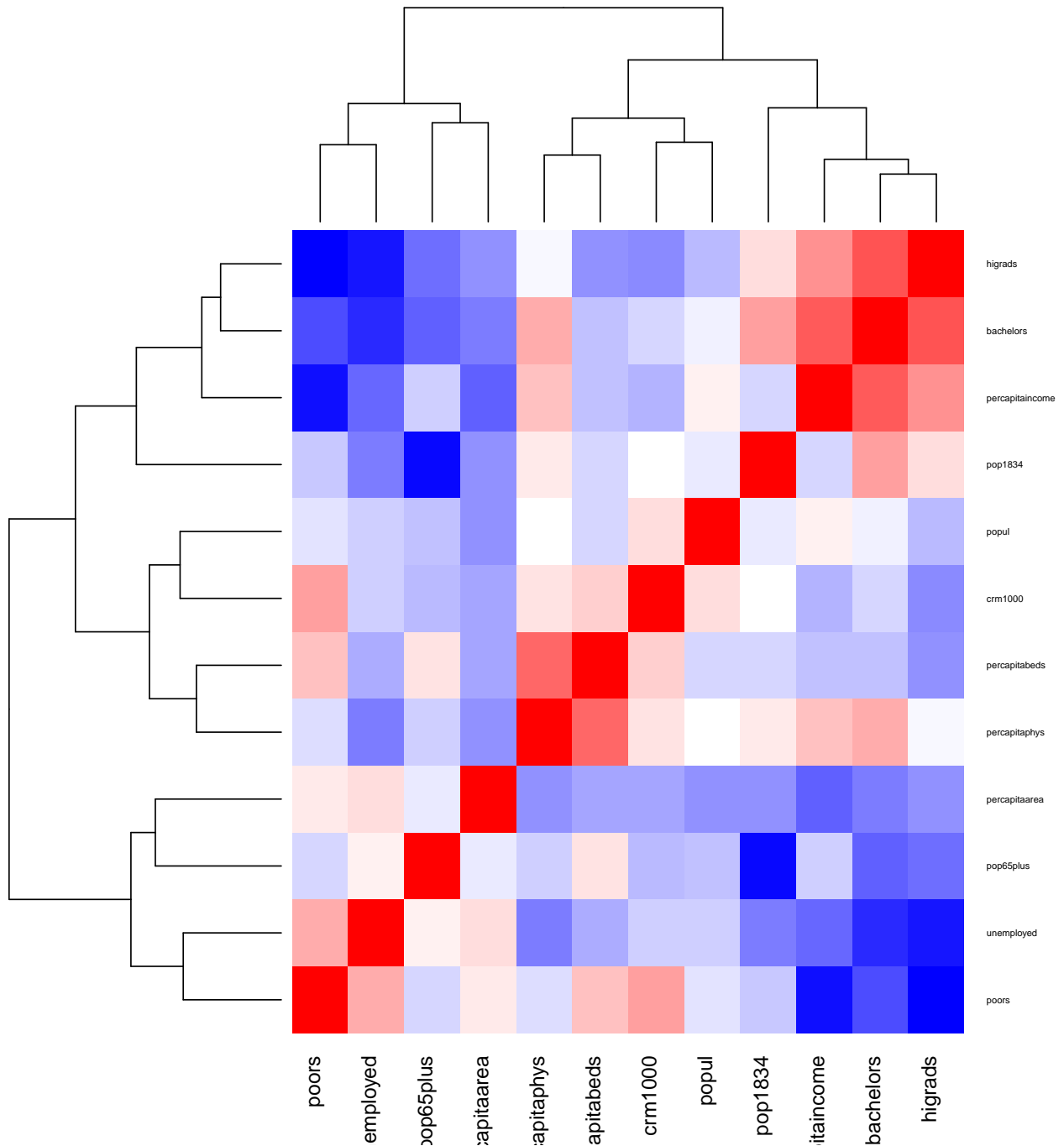
This clustered distance matrix was how we selected which variables to remove from the model. The blue variables are negatively correlated with each other and the red variables are positively correlated.

Outliers

Crime rate per 1000 by region

We can see that Kings county is high for its region and overall.

References

[1] Alan Agresti. *Categorical data analysis.* A Wiley-Interscience publication. Wiley, New York [u.a.], 1990.

[2] Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. *An Introduction to Statistical Learning: With Applications in R.* Springer Publishing Company, Incorporated, 2014.

[3] Philippe Rigollet. Statistics for applications, Fall 2016.

[4] Wikipedia contributors. Variance inflation factor — Wikipedia, the free encyclopedia. https://en.wikipedia.org/w/index.php?title=Variance_inflation_factor&oldid=878147754, 2019. [Online; accessed 13-January-2019].