AN UNDERSTANDING OF CRIME IN THE US

CODE

STAT 324

AUTHOR
Sydney Potkey and Alejandro Gomez

PUBLISHED
March 14, 2025

1 ABSTRACT

This report investigates factors influencing the amount of crime present in 400 US cities. Through visualizations, statistical tests, and variable analysis, this report provides insight as to what variables could potentially predict the prevalence of crime. Furthermore, predictive methods are used to test the plausibility of predicting crime.

2 INTRODUCTION

The United States ranks high among developed countries for both violent and property crimes. It is estimated that rates of violent crime in the US have at time been up to 9 times that of European countries [1].However elevated crime is not an issue for all US counties. The discrepancies in rate of crime for different areas of the country vary greatly. This study aims to gain insight as to what factors lead to a risk of increased crime in US counties.

Specifically this report investigates potential effects of high school graduation rate, poverty, and population on the amount of crime per 100,000 residents in 400 US counties. Understanding what factors lead to increased rates of crime is a first step in lowering the US crime rate as a whole.

The information that this study aims to collect could be a useful tool for policy makers, criminal justice professionals, and many others interested in the fight against crime in the US.

3 AKNOWLEDGEMENTS

1. Kalish, C. B. (n.d.). International crime rates. Bureau of Justice Statistics. https://bjs.ojp.gov/library/publications/international-crime-rates-0

4 MATERIALS AND METHODS

The original source of this data set is unknown. Our observational units are 400 singular US counties. The response variable being assessed is serious crimes per 100,000 residents. The data set contains 12 potential explanatory variables, 4 of which are investigated in this report. The quantitative variables population(in 100,00s), high school graduation rate, and poverty measured as the percent of population living below the poverty line, along with the categorical variable region (North East, North Central, South, West) are examined in relation to crime. It is unknown whether random sampling was implemented, but we do assume the study was observational.

5 SPLIT THE DATA

Our data on Crime in the US contains 440 counties/observations. For the purposes of this analysis, we will be randomly splitting the observations into two different groups:

1. testdata: Random sample of 20% (88 Observations)
2. traindata: The remaining 80% (352) of observations

The purpose of splitting the data is to use 20% of observations at the end of the analysis to validate our proposed model. The remaining 80% will be used in our main data analysis processes.The seed was set to 922, to ensure consistent results.

6 DATA VISUALIZATION

Having successfully split the data for analysis, we can now move on to the data visualization phase. Here we will be analyzing the associations between each variable through a Scatter plot and Correlation Matrix.

]

▶ Code

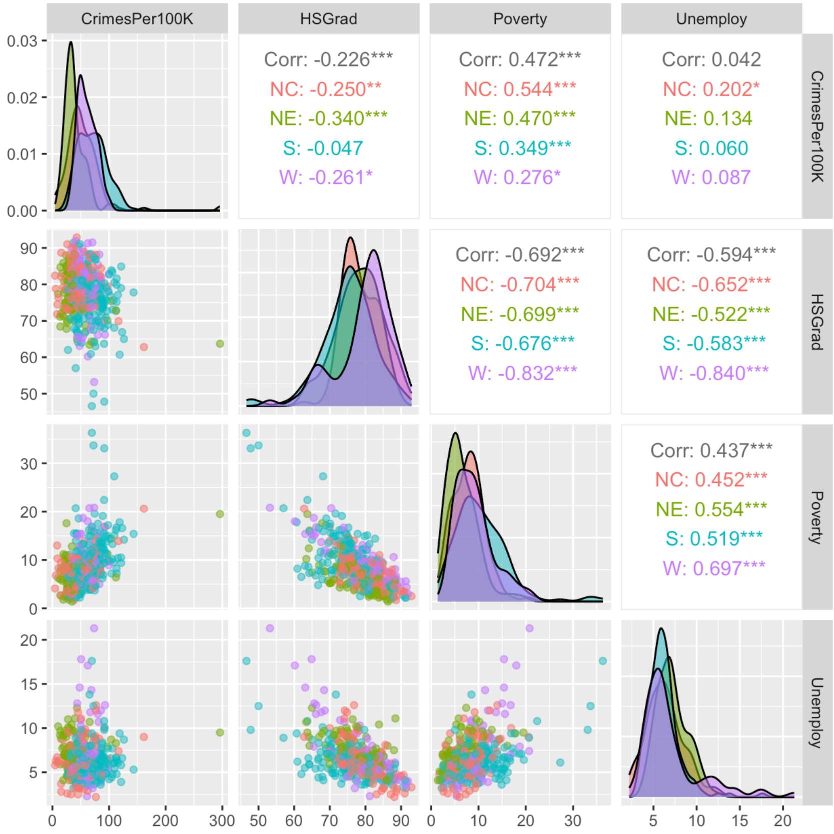


Figure 1: Matrix Scatterplot of Crime Rate Data

Figure 1 provides a brief look into the interaction between crimes per 100,000 residents and four explanatory variables through a scatter plot matrix. Poverty, measured by percent of population below the poverty line has the strongest linear association with crimes per 100,00 residents with a correlation coefficient of -0.692, suggesting a moderate negative linear association. High school graduation rate has the lowest linear association with crime rate reporting a correlation coefficient of -.226.

The results from Table 1 support the associations reported in the scatter plot matrix.

▶ Code

Table 1: Correlation Matrix of Elephant Data

	CRIMESPER100K	POP100K	HSGRAD	POVERTY
CrimesPer100K	1.000	0.280	-0.226	0.472
Pop100K	0.280	1.000	-0.017	0.038
HSGrad	-0.226	-0.017	1.000	-0.692
Poverty	0.472	0.038	-0.692	1.000

Table 1 displays the individual correlations of each explanatory variable with crime rate. The results support those found in the scatter plot matrix with poverty again having the strongest correlation to crimes per 100k residents and high school graduation rate having the lowest.

▶ Code

```
Analysis of Variance Table

Model 1: CrimesPer100K ~ Region + Poverty
Model 2: CrimesPer100K ~ Region + Poverty + Region:Poverty
  Res.Df    RSS Df Sum of Sq      F    Pr(>F)
1    435 228936
2    432 216423  3     12513 8.3259 2.152e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

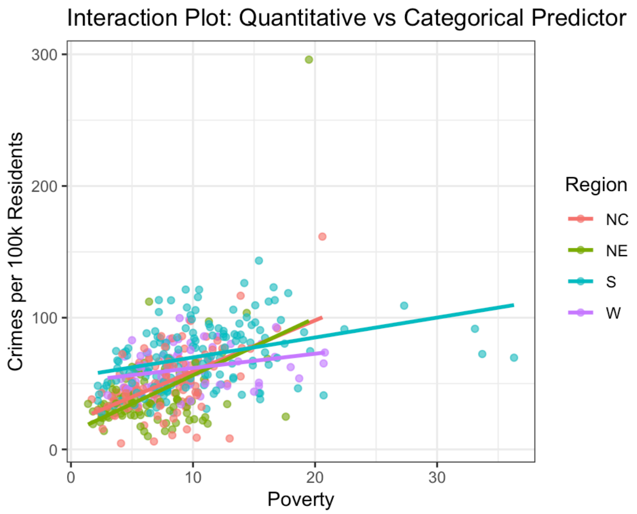


Figure 2: Scatter plot displaying poverty rate as the explanatory variable and serious crimes per 100,000 residents as the response. The colors represent the 4 US regions where data was collected. A linear regression has been applied for each region.

In Figure 2 you can see the linear regression lines for each region present in the data set. Each line displays the relationship between Poverty and Crime Rate for each region. The North Central and North East regions have the steepest regression lines giving evidence that high school graduation rate has stronger linear association with crime rate in these regions. There are a few apparent outlines with the most extreme coming from the North East region.

This scatter plot displays the interaction between poverty rate and region as predictors of crime. The colors represent each of the 4 regions from which data were collected. A linear regression line has been applied to each region to better visualize underlying relationships. Although not parallel, the lines all show the same general positive trend, as poverty increases crime increases regardless of region. However this positive trend is steeper for some regions, giving evidence of interaction. For example the slope of the North East's regression line is notably steeper than that of the Western region. This potentially implies that increases in poverty do not effect the crime rate in all regions equally.

In order to decide if we should account for interaction in our model a partial f test was ran comparing a reduced model with only region and poverty as predictors to a full model including the interaction term. The partial F test rendered a large F-statistic of 8.326, and a small p-value of 2.152e-05, giving strong evidence that the interaction term is significant in a model with region and poverty. For this region the interaction term was added to our model.

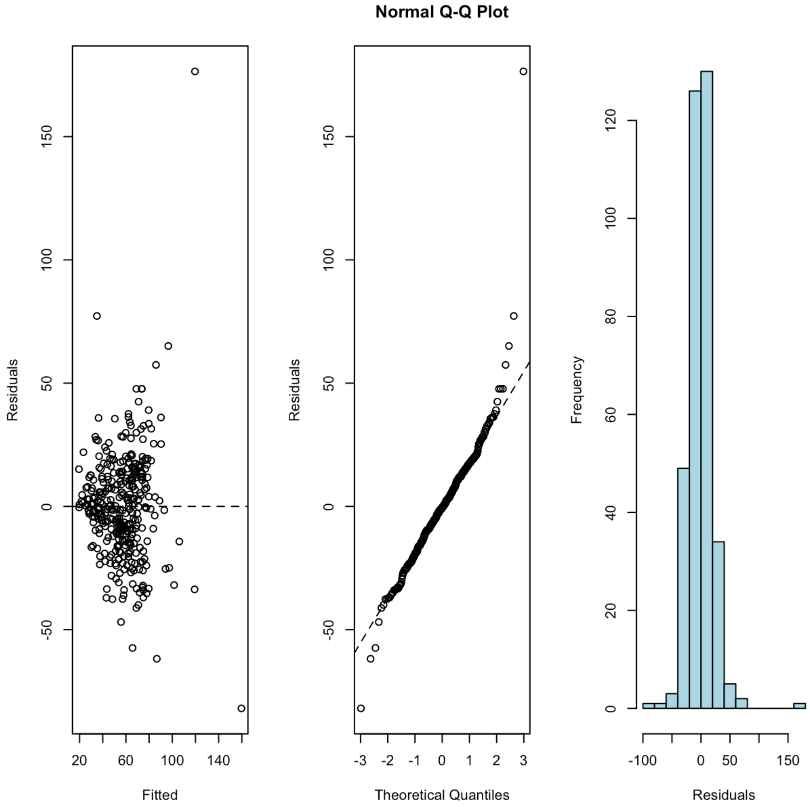
7 VARIABLE PRE-PROCESSING

This section discusses the essential transformation techniques required to fit a linear model under proper conditions.

A linear model is suitable for data that meets the FINE requirements. To understand if our data does this, we must take a look at 3 plots. ….

7.1 VERIFYING ORIGINAL DATA MODEL CONDITIONS

▶ Code



▶ Code

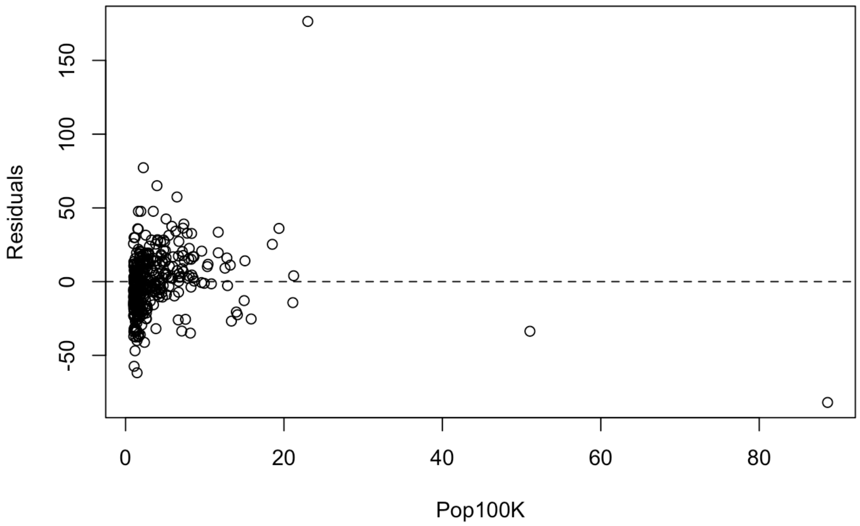
Form: Taking a look at the residuals vs. fitted plot, the residuals seem to be randomly scattered around our reference line, with the exception of one point. We can conclude that Form is suitable.

Equal Variance: Using the Residual Histogram, we can see that there the residuals are approximately normally distributed, but the Residual vs. Fitted plot shows us that fanning is present, so Equal Variance has been violated

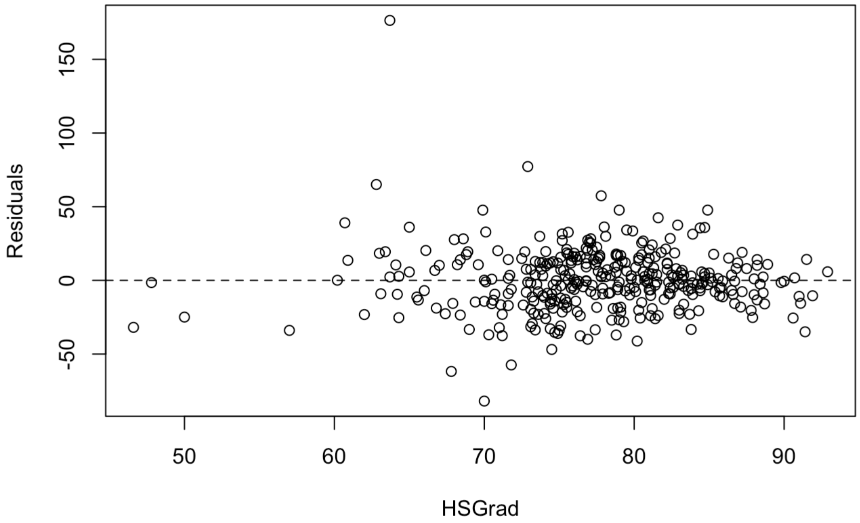
Normality: The QQPlot follows does not follow linear path relatively well, flaring out on each end, violating normality.

Because equal variance and normality have been violated, we may want to make a transformation on the response variable. However, we need to look at the residual plots of each individual explanatory variable to decide whether to increase or decrease power.

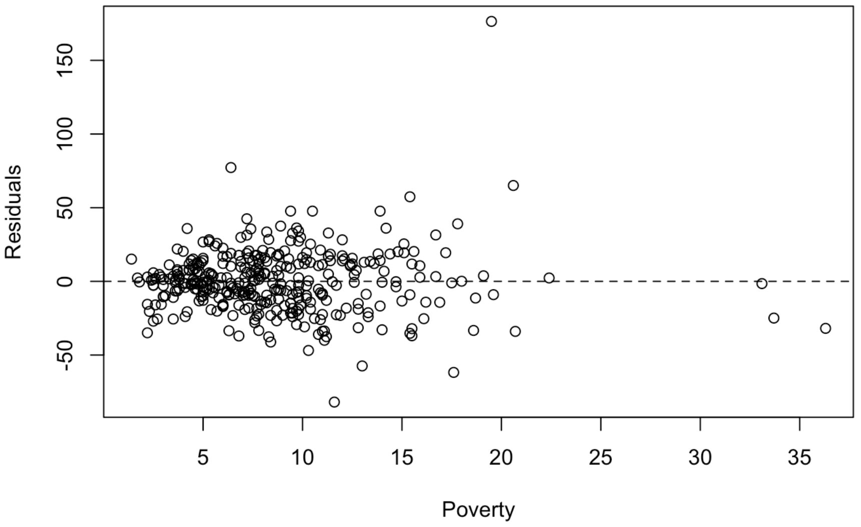
▶ Code


Residuals
150
100
50
0
-50
0
20
40
60
80
Pop100K

▶ Code


Residuals
150
100
50
0
-50
50
60
70
80
90
HSGrad

▶ Code



The 3 visuals above represent each individual variable's relationship with Crime Rate. While the 2 scatter plots containing Poverty and High School Graduation rate seem to maintain relatively linear form, the population visual contains a "C" shaped curve, indicating increasing the power of population in the linear model.

– Data Abnormalities –

While all 3 variables have some outliers, Pop100K seems to have very large outliers and an unusual residual plot reguardless of transformation. Standard deviation is also smaller than the rest, when ignoring the large outliers.

Using what we now know about the model conditions, we will increase y power using the square root, and increasing the population power by squaring it. The model:

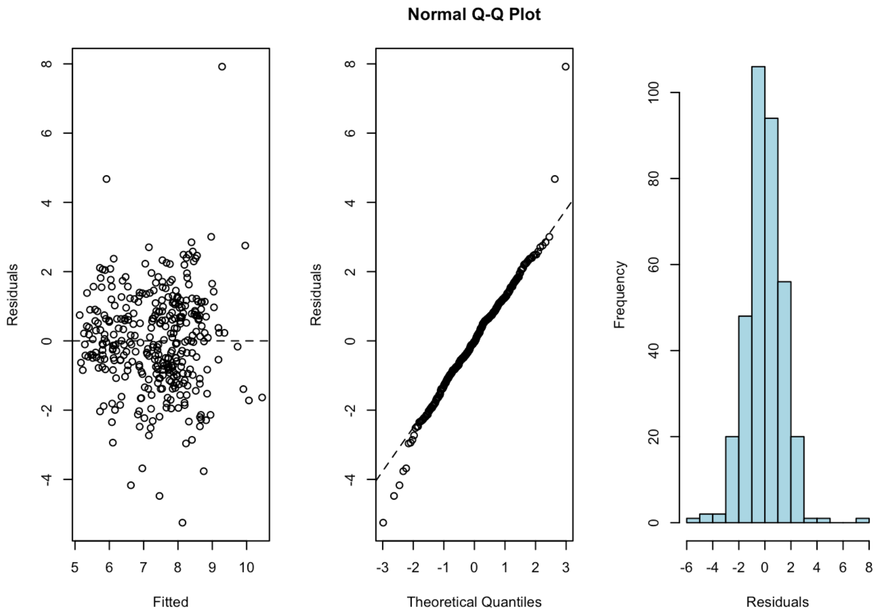
This model is formatted as code input for R Studio, and is not meant as an equation that reflects the actual Model Equation

$$\mathrm{CrimesPer100K}^{0.5} = \mathrm{Pop100K}^{2} + \mathrm{HSGrad} + \mathrm{Poverty} + \mathrm{Region} + \mathrm{Poverty:Region})$$

– Residual Analysis –

With the data now pre-processed and transformed, we can proceed to examine the residual diagnostics.

▶ Code



▶ Code

Form: The residual vs. fitted plot shows no patterns, assuring correct form.

Equal Variance: The residual vs. fitted plot shows no fanning or bow tie patterns with the majorty of points residing between 4 and -4, and the Residual Histogram follows an approximately normal distribution, assuring Equal Variance.

Normality: The QQ Plot follows a linear path very well, with the exception of one point, assuring normality.

7.2 LACK OF FIT TEST

We are also able to perform a lack of fit test, considering our Poverty variable has duplicates and equal variance and normality are satisfied. We will perform this test to confirm our transformed model has linear form:

▶ Code

```
Analysis of Variance Table

Model 1: CrimesPer100K^0.5 ~ I(Pop100K^2) + HSGrad + Poverty + Region + 
    Region:Poverty
Model 2: CrimesPer100K^0.5 ~ as.factor(Poverty)
  Res.Df    RSS  Df Sum of Sq      F Pr(>F)
1    342 701.71                            
2    210 489.04 132    212.67 0.6918  0.989
```

Because the p-value is very high, we have extremely strong evidence that the least squares line for our transformed linear model is not statistically significantly worse than the separate means model. In other words, correct form can be assumed for our model.

7.3 CONCLUSION

After analyzing the Residual vs Fitted Plot, QQ Plot, Residual Histogram, and Lack of Fit Test, we can confirm that all conditions are reasonably met and proceed with statistical inference.

8 FIT A LINEAR MODEL

Having conducted thorough residual diagnostics, we are now equipped to proceed with statistical inference and prediction under our transformed linear model.

▶ Code

?(caption)

```
Call:
lm(formula = CrimesPer100K^0.5 ~ I(Pop100K^2) + HSGrad + Poverty + 
    Region + Region:Poverty, data = traindata)

Residuals:
    Min      1Q  Median      3Q     Max 
-5.2504 -0.8338 -0.0337  0.8644  7.9198 

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)    
(Intercept)       1.2188429  1.5290888   0.797  0.42594    
I(Pop100K^2)      0.0003697  0.0001751   2.111  0.03548 *  
HSGrad            0.0434859  0.0166794   2.607  0.00953 ** 
Poverty           0.2914403  0.0480636   6.064 3.52e-09 ***
RegionNE         -0.2196814  0.5156956  -0.426  0.67038    
RegionS           2.5335539  0.4664860   5.431 1.06e-07 ***
RegionW           1.7573332  0.6381386   2.754  0.00620 ** 
Poverty:RegionNE -0.0186428  0.0634734  -0.294  0.76916    
Poverty:RegionS  -0.1734865  0.0492089  -3.526  0.00048 ***
Poverty:RegionW  -0.1597916  0.0689748  -2.317  0.02111 *  
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.432 on 342 degrees of freedom
Multiple R-squared:  0.3597,	Adjusted R-squared:  0.3428 
F-statistic: 21.35 on 9 and 342 DF,  p-value: < 2.2e-16
```

- All of the following information is in context of ?@tbl-mod1-summary.

8.1 MODEL EQUATION

To understand the context of our analysis, the Model Equation is provided below:

$$\sqrt{\widehat{Crime}} = 3.067 + 0.000387(Pop100K^2) + 0.034(HSGrad) + 0.153(Poverty) - 0.480(RegionNE) \\ + 1.051(RegionS) + 0.558(RegionW) - 0.0186428(Poverty)(RegionNE) \\ - 0.1734865(Poverty)(RegionS) - 0.1597916(Poverty)(RegionW)$$

8.2 MODEL UNDERSTOOD

Because the model is transformed, prediction from this model does not directly predict Crime Rate. Each explanatory variable ($Pop100K^2$, HSGrad, Poverty, and Region) contributes to the prediction of $\sqrt{Crime\,Rate}$. Because of this we will need to square any results in order to correctly interpret crime per 100,000 residents. So how well does our model do its job?

8.3 MODEL EVALUATION

This particular model does a moderate job in predicting $\sqrt{Crime\,Rate}$, considering only 35.97% of variation is accounted for by this model, which is fairly low. However, the model utility test p-value does imply that at least one of the variables are statistically significant predictors of $\sqrt{Crime\,Rate}$. As well, each individual variable is a statistically significant predictor of $\sqrt{Crime\,Rate}$, after adjusting for all other variables, as seen by the low p-values. While reducing the amount of Unknown Error could be improved, the predictors in our model do seem to be efficient.

Interpretation of R^2: 35.97% of variation in $\sqrt{Crime\,Rate}$ is accounted for by $Pop100K^2$, HSGrad, Poverty, and Region. This value is fairly low, as stronger models would be expected to have an R^2 value above 50%.

Interpretation of S: The typical deviation of an individual $\sqrt{Crime\,Rate}$ from the mean is 1.432.

- The units for $\sqrt{Crime\,Rate}$ are $\sqrt{\text{crimes per } 100,000 \text{ people}}$

8.4 COEFFICIENT INTERPRETATIONS

After further approval of our model, individual slope coefficients in our model may be interpreted, considering they hold valuable information about their relationship with $\sqrt{Crime\,Rate}$. However, we must keep in mind that each β_i hold a relationship with respect to our transformed response variable, not necessarily the original Crime Rate variable.

Intercept: The expected $\sqrt{Crime\,Rate}$ when all explanatory variables are 0 and the county is in North Central America, is 3.0668.

$Pop100K^2$: Each increase of one in $Pop100K^2$ is associated with an increase of 0.000387 in predicted $\sqrt{Crime\,Rate}$, after adjusting for the other variables.

RegionW: The expected difference in $\sqrt{Crime\,Rate}$, if a county is in the Western United States, is 0.558 greater than a county in North Central America, with any given Pop100K, HSGrad, and Poverty Rate.

- The units for $\sqrt{Crime\,Rate}$ are $\sqrt{\text{crimes per } 100,000 \text{ people}}$
- These are just a few interpretations of the many β_i in our linear model.

8.5 VERIFYING THE ABSENCE OF MULTICOLLINEARITY

One last way we ca verify our model is sufficient for inference is by verifying the low influence of multicollinearity.

▶ Code

VARIABLE	GVIF
I(Pop100K^2)	1.031141
HSGrad	2.363429
Poverty	8.666790

VARIABLE	GVIF
Region	5.757984
Poverty:Region	7.392001

A couple of the GVIFs are above 5, indicating that collinearity will slightly affect our inferences, however none are above the severse threshold of 10. Thus, we should be slightly cautious when making inferences, but our model's credibility is not severely violated.

We can now implement our model for statistical inference.

9 STATISTICAL INFERENCE

With the model fitted and evaluated, we can now shift our focus to inference, where we will compare ours to a less complex model to ensure significance. We will also interpret confidence intervals to apply our discoveries.

Recall ?@tbl-mod-summary

– Model Utility Test –

Under the context of our linear model, the hypotheses being tested are as follows:

$$\begin{aligned} H_o : \beta_{Pop100K^2} = \beta_{HSGRad} = \beta_{Poverty} = \beta_{RegionNE} = \beta_{RegionS} = \beta_{RegionW} \\ = \beta_{(Poverty)(RegionNE)} = \beta_{(Poverty)(RegionS)} = \beta_{(Poverty)(RegionW)} = 0 \end{aligned}$$

$$H_A : \text{Atleast } 1\, \beta_i \neq 0$$

Our model utility test reveals an overall F statistics of 21.35 on 9 and 342 degrees of freedom, and a p-value of 2.2e-16. Both values give evidence that our model is a statistically significant predictor of crime rate per 100k residents.

– Partial F-Test –

Another way we can gage how effective our model is in predicting Crime Rate, is by comparing it to a much more basic model. Here, we will be comparing our model with one only containing our most correlated predictor, Poverty.

▶ Code

```
Analysis of Variance Table

Model 1: CrimesPer100K^0.5 ~ Poverty
Model 2: CrimesPer100K^0.5 ~ I(Pop100K^2) + HSGrad + Poverty + Region + 
    Region:Poverty
  Res.Df    RSS Df Sum of Sq      F    Pr(>F)    
1    350 885.75                                  
2    342 701.71  8    184.04 11.212 4.176e-14 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

$$\begin{aligned} H_o : \beta_{Pop100K^2} = \beta_{HSGRad} = \beta_{RegionNE} = \beta_{RegionS} = \beta_{RegionW} \\ = \beta_{(Poverty)(RegionNE)} = \beta_{(Poverty)(RegionS)} = \beta_{(Poverty)(RegionW)} = 0 \end{aligned}$$

$$H_A : \text{Atleast } 1\, \beta_i \neq 0$$

F-Stat: 13.488

P-Value: ≈ 0

Conclusion: Because the p-value is very small 0, we have very strong evidence that adding the predictors Pop100K^2, HSGrad, Region, and the interaction term to a model with just Poverty, statistically significantly improves the prediction of $\sqrt{\text{Crime Rate}}$.

To ensure that our interaction term is significant, we will also perform a partial F-test comparing a model with and without the interaction term.

▶ Code

```
Analysis of Variance Table

Model 1: CrimesPer100K^0.5 ~ I(Pop100K^2) + HSGrad + Poverty + Region
Model 2: CrimesPer100K^0.5 ~ I(Pop100K^2) + HSGrad + Poverty + Region + 
    Region:Poverty
  Res.Df    RSS Df Sum of Sq      F    Pr(>F)    
1    345 739.71                                  
```

```
2    342 701.71  3    37.998 6.1732 0.0004271 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

$$\beta_{(Poverty)(RegionNE)} = \beta_{(Poverty)(RegionS)} = \beta_{(Poverty)(RegionW)} = 0$$

$$H_A : \text{Atleast } 1\ \beta_i \neq 0$$

F-Stat = 6.1732

P-value = 0.0004271

Conclusion: Because the p-value is very small 0, we have very strong evidence that adding the interaction term between region and poverty, statistically significantly improves the prediction of $\sqrt{\text{Crime Rate}}$.

– Prediction and Confidence Intervals –

Moving onto further analysis, approval from our past examinations of our linear model allows us to confidently make predictions about the $\sqrt{\text{Crime Rate}}$ of counties with specific attributes. In this case, we will be performing 2 predictions:

▶ Code

```
[1] 57.28644
```

▶ Code

```
[1] 27.32771
```

▶ Code

```
       fit      lwr      upr
1 9.284325 6.219238 12.34941
```

▶ Code

```
       fit      lwr      upr
1 9.284325 8.077325 10.49133
```

The parameters chosen for confidence and prediction intervals come from the observed data for Kings New York. Kings NY is of interest because it's crime rate is more than 8 standard deviations above the mean crime rate, making it the city with the highest crime rate in this data set.

Mean crime rate: 57.2864 King NY cime rate: 295.987

Note: The statistics in the following interpretations have been squared to account for earlier transformations.

With 95% confidence, we predict that a city in the North East with a population of 230,000, high school graduation rate of 63.7%, and Poverty rate of 19.5% will experience crimes per 100k residents between 38.676 and 152.498. The average amount of crimes per 100k residents for a city with the same parameters is predicted with 95% confidence to be between 65.238 and 110.040.

These intervals are interesting as they do not capture the actual amount of crime recorded for Kings New York. This however is not entirely suprising as Kings NY was an extremely abnormal outlier (see fig coded scatter).

Because of this unique result we will conduct another confidence and prediction interval to ensure the predictive capabilities of our model. For the second test we have chosen to use parameters from Santa Barbra county which is not an outlier.

▶ Code

```
       fit      lwr      upr
1 7.434301 4.585736 10.28287
```

▶ Code

```
       fit      lwr      upr
1 7.434301 7.014291 7.854311
```

Prediction interval: (21.031, 105.74) Confidence Interval: (49.196, 61.685) With 95% confidence, we predict that a county with a population of 369,600, high school graduation rate of 80%, and poverty level of 7.4% will experience between 21.031 and 105.74 crimes per 100,000 residents. Furthermore we can narrow down the interval for the average amount of crime experienced per 100,000 residents to be between 49.169 and 61.685.

Both of these intervals contain the actual value reported for Santa Barbra county, verifying the predictive ability of our model.

10 MODEL VALIDATION

We will now turn to model validation to ensure that our model performs reasonably well outside of our training data.

▶ Code

```
Analysis of Variance Table

Response: CrimesPer100K^0.5
                Df Sum Sq Mean Sq F value    Pr(>F)    
I(Pop100K^2)     1  12.78  12.780  6.2286 0.0130406 *  
HSGrad           1  44.58  44.583 21.7291 4.507e-06 ***
Poverty          1 184.26 184.256 89.8029 < 2.2e-16 ***
Region           3 114.56  38.188 18.6121 3.312e-11 ***
Poverty:Region   3  38.00  12.666  6.1732 0.0004271 ***
Residuals      342 701.71   2.052                      
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We will now turn to model validation to ensure that our model performs reasonably well outside of our training data.

Using the test data that was subset earlier, MSPE or predicted mean square error was calculated to be about 1.3655. The MSE from our model is 2.052. These values are reasonably close and allow us to successfully verify that our model's predictive capabilities are acceptable.

11 CONCLUSION

The regression analysis of crime in US counties leads us to two major conclusions. Firstly a regression model with multiple predictors can significantly increase our ability to predict potential high and low crime counties. Through a model utility test we found that our model which considers, high school graduation rate, poverty level, region, population, and the interaction between region and poverty, significantly increases our ability to predict crime rates. Furthermore we performed a partial F-test comparing our model to one containing only poverty as a predictor for crime. This test allowed us to conclude that our more complex model does a significantly better job as a prediction tool than a model that attempts to estimate crime based on poverty alone.

What does this mean for real world crime? The findings in this report give insight as to what risk factors could be contributing to counties currently experiencing increased crime rates. More importantly our model allows the investigation of specific combinations of risk factors that may seem insignificant if viewed one at a time. However our model is by no means an exact predictor of crime.

One issue that appeared consistently throughout this report was a lack of a strong coefficient of determination. Our model failed to reach levels above 38% meaning that there is still a substantial amount of unexplained variation within the final model.

Another issue pertained to data collection. We were unable to find vital information about how, where, and who originally collected the data. In a future study knowing this information could enable the researcher to communicate better conclusions and aid in variable selection; the researcher could select variables for analysis that the original data collector found particularly important.

In a future study it would likely be beneficial to test more terms for interaction. We were able to decrease unexplained variation by including an interaction term between poverty and region. There is a possibility that other variables could have significant interaction with each other as well, leading to the possibility of a superior model.

12 APPENDIX

12.1 DATASET

The dataset has been submitted separately on Canvas alongside the project submission.

– Data set Variables –

1. CrimesPer100K: the number of serious crimes (per 100,000 people)
2. Area: Land area (in 100s of square miles)
3. Pop100K: Total population (in 100,000s of people)
4. Pop18-34: Percent of the population that is age 18 to 34 (in percent)
5. Pop65+: Percent of the population that is age 65 or over (in percent)
6. Physicians: Total number of practicing physicians
7. BedsPer100K: Number of hospital beds (per 100,000 people)
8. HSGrad: Percent of the population that has graduated from high school (in percent)

9. Bachelors: Percent of the population that has graduated from a 4-year college (in percent)

10: Poverty: Percent of the population that below the poverty line (in percent)

11. Unemploy: Percent of the adult population that is unemployed (in percent)

12. IncPerCapita: Income per person (in thousands of dollars)

13. Region: Location of the county within the U.S. (North East, North Central, South, West)

– Our variables –

1. testdata: Random sample of 20% (88 Observations) in the dataset

2. traindata: The remaining 80% (352) of observations in the dataset

Source: Canvas

12.2 LIST OF EXTRAS

- Title Page Image Address:

https://www.google.com/url?sa=i&url=https%3A%2F%2Fwww.redbubble.com%2Fi%2Fsticker%2FBad-Girl-Dog-Arrested-And-Incarcerated-by-smartnet77%2F44053999.EJUG5&psig=AOvVaw3fcPlAeMiLj-PE5NXwdjSJ&ust=1739772561519000&source=images&cd=vfe&opi=89978449&ved=0CBQQjRxqFwoTCIDu8_3Dx4sDFQAAAAAdAAAAABAE

- Interaction plot
- Test of significance for interaction term
- Acknowledgement
- Use of ggplot
- Improved presentation: theme, table of contents, in line code
- Color coded matrix scatterplot