



The Hunt for Vehicle Theft Predictors

12.09.2017

Tayler O'Connell
Jon Hsu
Erin Tsai
Kenneth Owen



Table of Contents

Introduction.....2

Objectives.....2

Methods and Procedures.....3-4

Preliminary Analysis.....5-7

Analysis.....7-18

Conclusion.....18-19

Bibliography.....20

Introduction

Have you ever been the victim of a car theft? If you answered no, you are among the minority who has not yet been affected by the nationwide problem of motor vehicle theft. According to the FBI¹, there are hundreds of thousands of automobiles stolen every year in the United States. Because of the sheer number of thefts, the amount of property loss each year totals in the billions, with the loss per vehicle averaging around six thousand dollars. With auto theft being such a prolific and costly issue throughout the United States, our group has decided to use statistical analysis to help shed some light on the issue and possibly even provide useful insight into how to prevent future thefts from occurring.

Objective

The objective of this research paper is to determine if there are any statistically significant predictors for auto theft. The study is based on data from the U.S. Census Bureau which provides census information for every county in the U.S. for the years 1981-2008². Using regression analysis, the group plans to identify whether or not factors like age, unemployment, income, and education have an

¹ [FBI Motor Vehicle Theft](#)

² [United States Census Bureau](#)

effect of the amount of auto thefts per year in the United States. Once a model is determined, key predictors will be used to inform a time series regression.

Methods and Procedures

In order for us to meet our objectives, we used multiple linear regression to determine which factors are predictors for the number of motor vehicle thefts per capita at a 95% confidence level. This enabled us to create a model to predict the number of thefts per region. We then applied regression diagnostics to check the model assumptions of linearity, normality, and constant variance. We also performed a analysis of the trends in motor vehicle theft from 20001 to 2016, and compared those to trends in the prediction factors over the same time period.

The following list shows the variables that were used for this study. Some variables were converted into per capita numbers as a precaution against heteroscedasticity.

Initial Regression

Output variable:

- Number of motor vehicle thefts known to police in 2008 (per 10000 residents)

Input variables:

- Resident population total (July 1 - estimate) 2008
- Civilian labor force unemployment rate during 2008

- Educational attainment - persons 25 years and over - high school graduate (includes equivalency) 2005-2009 (per capita)
- Per capita personal income in 2007
- Median age of resident population: (April 1 - complete count) 2010
- People of all ages in poverty - percent 2008
- Local government finances - direct general expenditures for police protection FY 2002 (per capita)
- Geographical region

Time Series Regression³

Output Variable:

- Number of vehicles stolen in the United States (2000-2016)⁴ (per 10,000 residents)

Input Variables:

- United States resident population total (July 1 - estimate) 2000-2016
- United States unemployment ages 16 and over 2000-2016⁵ (percentage)
- Period/Year (e.g. 2000="0", 2001="1", etc.)

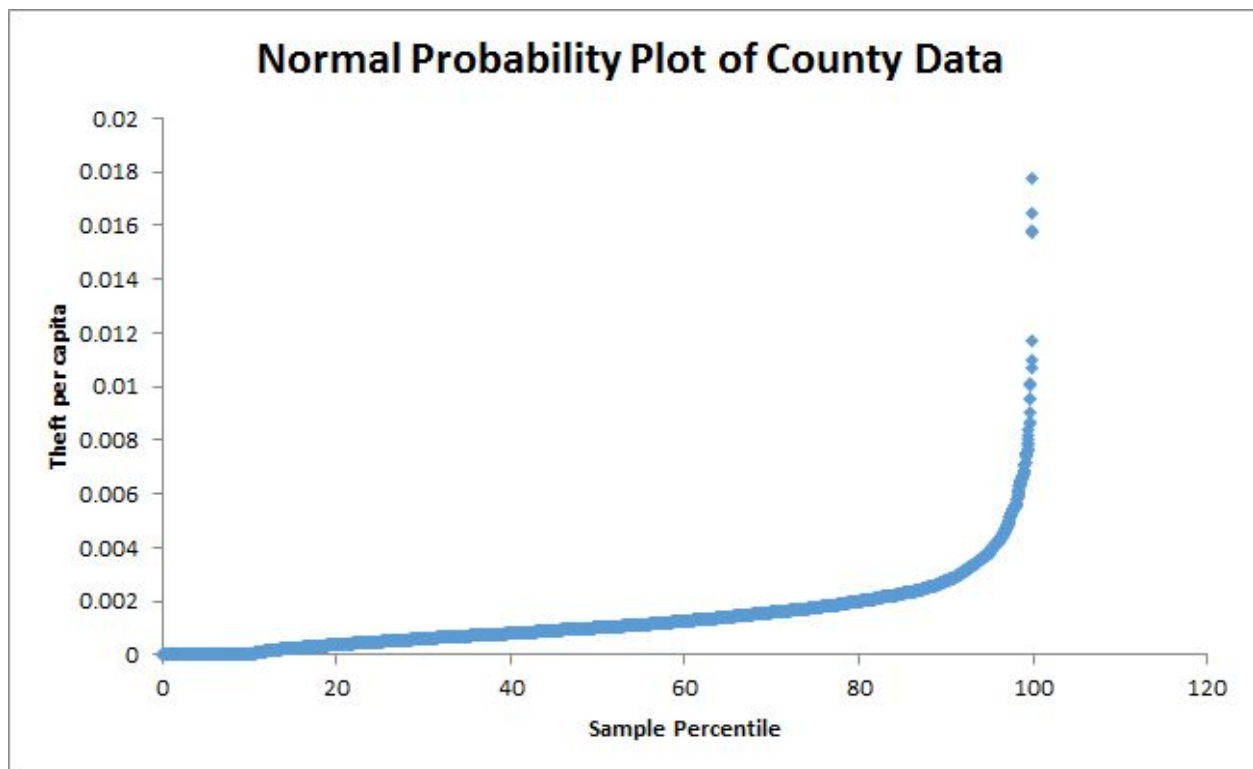
³ Recent time series required additional sources, due to decennial nature of U.S. Census

⁴ US Department of Justice Federal Bureau of Investigation, Uniform Crime Reports

⁵ Bureau of Labor Statistics

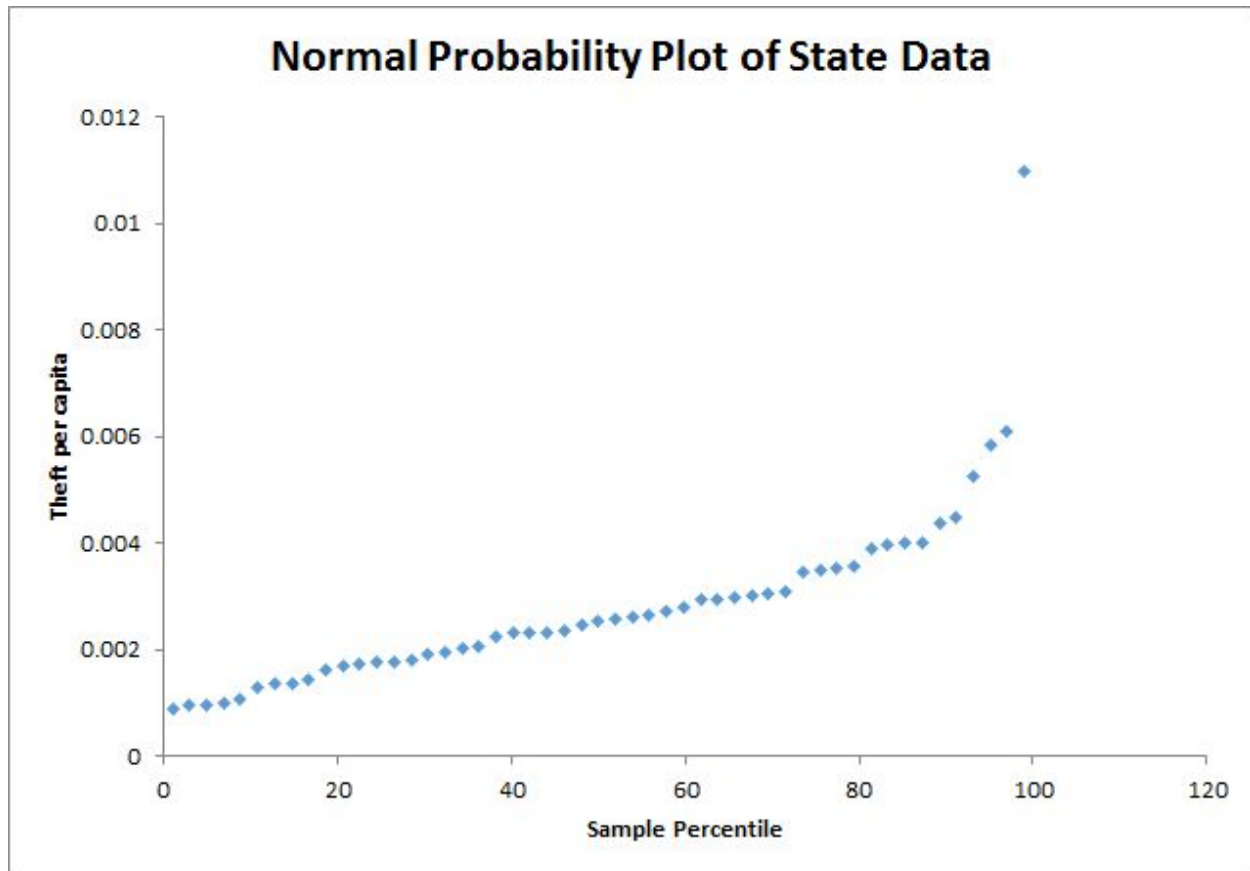
Preliminary Analysis

The data from the Census Bureau covered 3141 counties in the United States. However, a preliminary analysis showed that applying linear regression techniques on the county level data would produce an invalid model. The normal probability plot indicates that the errors do not follow a normal distribution. This violates a fundamental assumption of the ordinary least squares method. Therefore, an alternative data set was required.



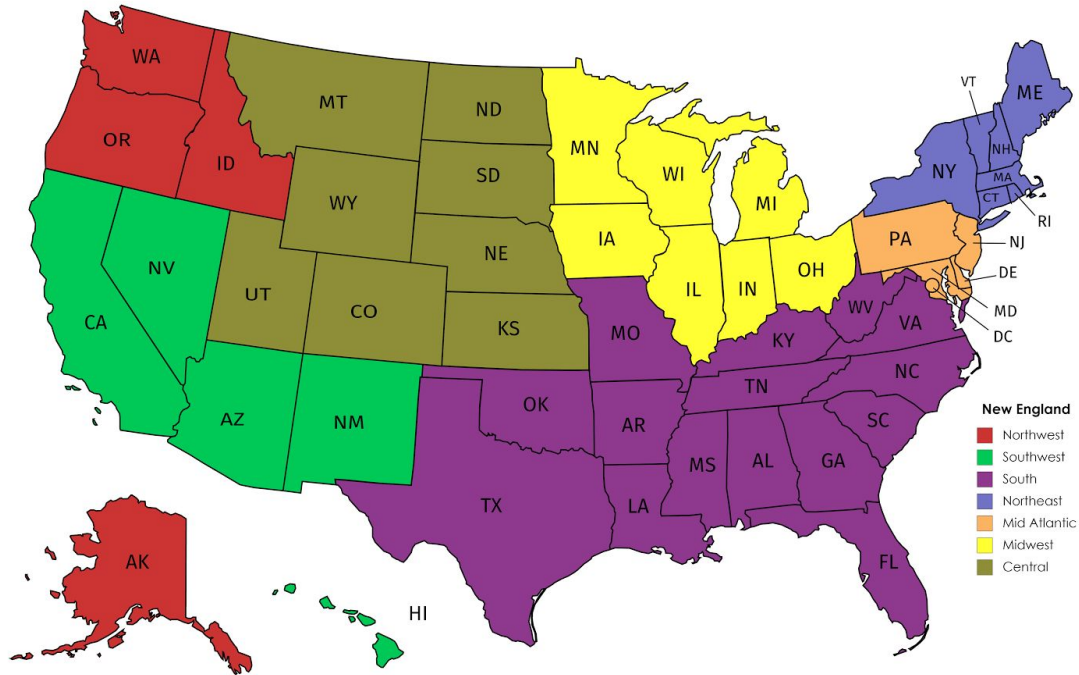
The Census Bureau also provided data at the state level. A preliminary analysis of the state level data shows a much cleaner normal probability plot. There

is one outlier corresponding to the District of Columbia. The rest of the observations follow a fairly linear pattern, so the normality assumption is satisfied. For the subsequent analyses, the District of Columbia observation was excluded.



One of the input variables included in this study was geographic region. The Census Bureau data did not define any geographic regions, so this input variable was created manually. The method used to define geographic regions was to group states that were adjacent to each other and had similar auto theft rates. The resulting regions are shown on the map below. For the regression analysis, a set of

6 binary variables were used to describe the 7 geographic regions. The South was chosen as the default region and no binary variable was used for that region.



Created with mapchart.net ©

Analysis

Initial Regression

After determining which data set to analyze and what input variables to include, a linear regression model was created. The preliminary regression (all independent variables used) resulted in an adjusted R^2 of 0.715 and overall F_{calc} of

10.66. This would suggest a well-fitted, significant model. However, the initial model showed that many of the input variables were not significant at a 95% confidence level based on their p-value.

<u>Input variable</u>	<u>P-value</u>
Population	0.03302
Unemployment rate	0.030364
High school graduates	0.088585
Income	0.332707
Median age	0.872866
Poverty level	0.255857
Police spending	0.066062
Northwest region	0.09999
Southwest region	0.013246
Central region	0.192233
Midwest region	0.268193
Mid Atlantic region	0.490715
Northeast region	0.019696

From the initial linear regression, the model was simplified iteratively by removing the input with the largest p-value above 0.05 and recalculating the regression. As some inputs were removed, other inputs saw their p-values change. Eventually, only 4 significant inputs remained. The final model is described below.

Regression Statistics

Multiple R	0.813168
R Squared	0.661243
Adjusted R Squared	0.631131
Standard Error	7.460156
Observations	50

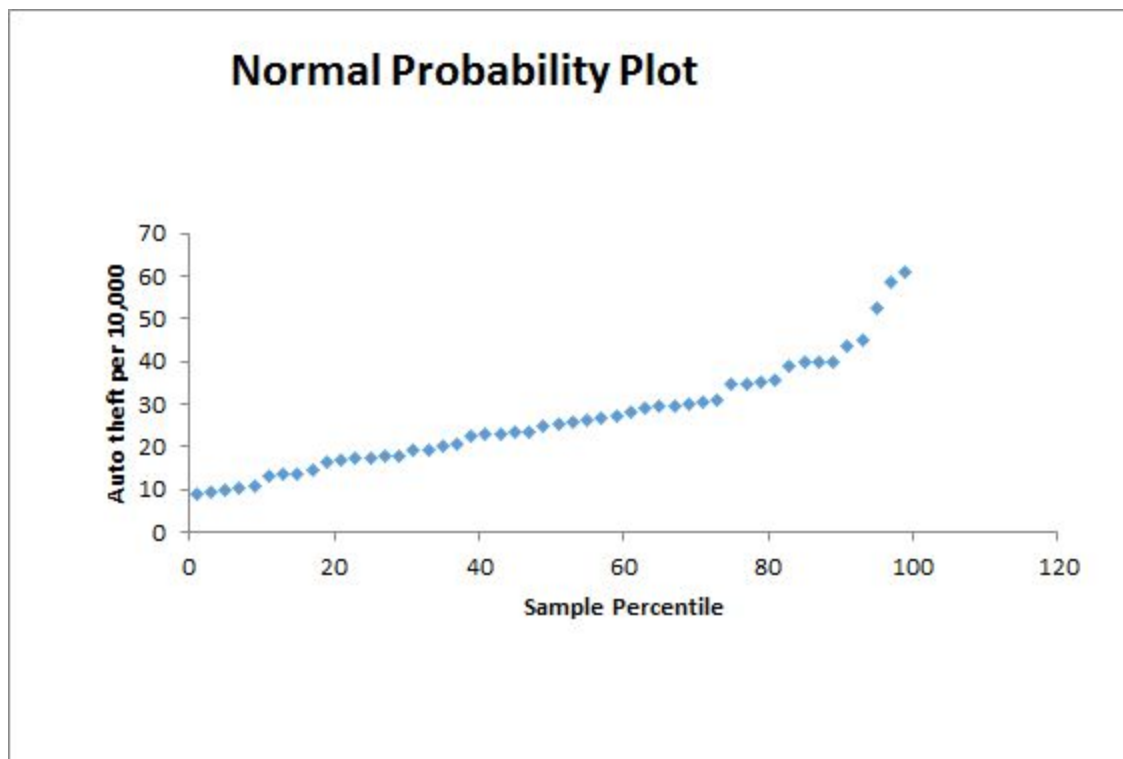
<i>ANOVA</i>	<i>df</i>	<i>SS</i>	<i>MS</i>	<i>F</i>	<i>Significance F</i>
Regression	4	4888.55285	1222.138	21.95961	4.20042E-10
Residual	45	2504.426436	55.65392		
Total	49	7392.979286			

<i>Predictors</i>	<i>Coefficients</i>	<i>Standard Error</i>	<i>t Stat</i>	<i>P-value</i>	<i>Lower 95%</i>	<i>Upper 95%</i>
Intercept	25.75101	8.950786911	2.876955	0.006119	7.723204126	43.77882
Unemployment Rate	3.586839	0.867861395	4.132963	0.000154	1.838876167	5.334801
High school Grad %	-0.9618	0.382578899	-2.51398	0.015583	-1.73235089	-0.19124
Southwest	20.81367	3.739694463	5.565607	1.37E-06	13.28153872	28.3458
Northeast	-8.82897	3.075060286	-2.87115	0.006214	-15.0224604	-2.63548

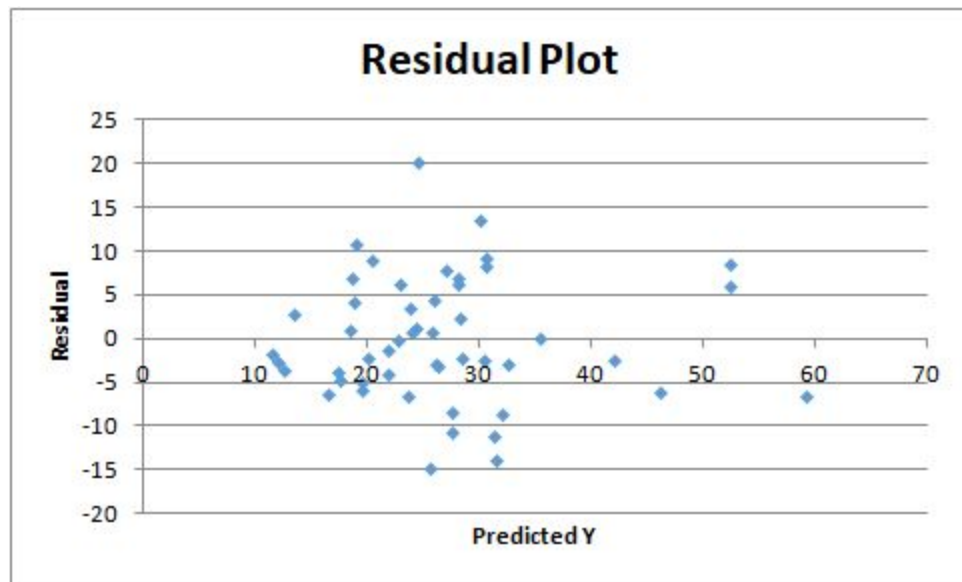
The new model has an adjusted R^2 of 0.63, suggesting 63% of the variability in auto thefts per capita is explained by unemployment rate, high school graduation rate, and geographic region. The overall F statistic increased significantly.

Diagnostics

A normal probability plot was used to check for normality of errors. The plot shows a fairly linear trend, which does not provide enough evidence for us to reject the hypothesis that the errors are normally distributed.



The plot of residuals shows no evidence of heteroscedasticity. There is no apparent "funnel-in" or "fan-out" pattern.



To evaluate the possibility of multicollinearity, we checked the correlation matrix for all independent variables. Based on the correlation matrix, most values are very small. There are a few values that around 0.5 or -0.5. However, we do not believe that they should cause any immediate concern. Under Klein's Rule, we should worry about the stability of the regression coefficient estimates only when a pairwise predictor correlation exceeds the multiple correlation coefficient R (the square root of R^2). Based on our preliminary analysis, we do not have any values that exceed approximately 0.8, so there is no multicollinearity problem.

	<i>Pop.</i>	<i>Unemp.</i>	<i>High School</i>	<i>Income</i>	<i>Median age</i>	<i>Poverty level</i>	<i>Police sp./capita</i>	<i>Region1</i>	<i>Region2</i>	<i>Region3</i>	<i>Region4</i>	<i>Region5</i>	<i>Region6</i>
Population	1.000												
Unemployment	0.001	1.000											
High school	-0.046	0.163	1.000										
Income	0.050	-0.303	-0.275	1.000									
Median age	-0.030	0.004	0.578	0.045	1.000								
Poverty level	-0.017	0.447	0.092	-0.527	-0.156	1.000							
Police spending/capita	0.034	-0.016	-0.150	0.319	-0.025	-0.104	1.000						
Region1	-0.005	0.067	-0.146	0.029	-0.026	-0.055	0.099	1.000					
Region2	0.017	0.114	-0.188	0.060	-0.025	0.003	0.231	-0.047	1.000				
Region3	-0.018	-0.403	-0.055	0.104	0.113	-0.169	0.025	-0.096	-0.089	1.000			
Region4	-0.010	0.145	0.198	0.044	0.062	-0.242	0.000	-0.111	-0.103	-0.213	1.000		
Region5	0.007	-0.025	0.083	0.169	0.020	-0.130	-0.001	-0.044	-0.041	-0.085	-0.098	1.000	
Region6	0.008	-0.028	-0.048	0.180	0.041	-0.118	0.009	-0.047	-0.044	-0.091	-0.104	-0.042	1.000

The final linear regression model is:

(Auto thefts per 10,000) = 25.75 + 3.587*(Unemployment rate) - 0.9618*(High school graduate %) + 20.814*(SW region) - 8.8*(NE region)

This model is logical. Auto thefts are positively correlated with unemployment rate and negatively correlated with high school graduate rates. It appears that being in the Southwest region increases the likelihood of auto theft, while the Northeast region decreases it. All other regions were determined to be insignificant predictors in the model.

Time Series Regression

For the time series regression spanning 2000-2017, regional variables were excluded, since we were interested in the regression at a national level. Given the small coefficient and difficulty finding reliable per annum data, high school

graduate percentage was also removed from predictor consideration. An initial plot of vehicle thefts/10,000 revealed a polynomial trend.



Given Occam's razor, we determined that a linear regression with an acceptably high adjusted r^2 would suffice in place of a quadratic model. An initial correlation matrix revealed no violation of Klein's Rule.

	<i>Year</i>	<i>%Unemployed</i>
<i>Year</i>	1	
<i>%Unemployed</i>	0.392013	1

Time Series Regression Results

The regression produced an overall significant result at a 95% confidence level, with F_{calc} equal to 95.3. The predictors Time and Unemployment rate were significant.

<i>Regression Statistics</i>	
Multiple R	0.965182736
R Square	0.931577713
Adjusted R Square	0.921803101
Standard Error	2.602506185
Observations	17

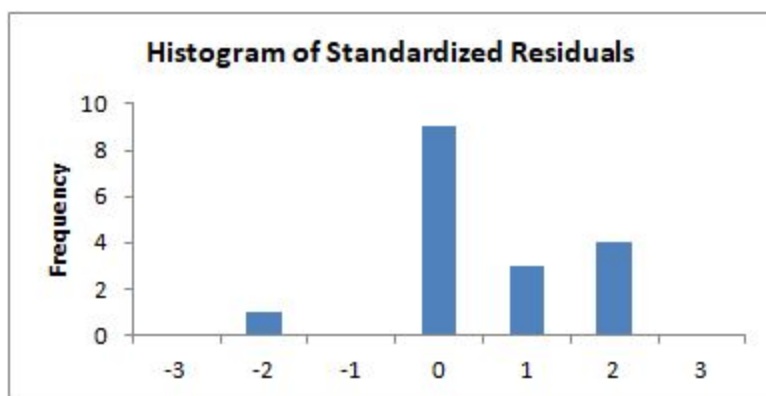
ANOVA	df	SS	MS	F	Significance F
Regression	2	1291.020333	645.5101664	95.30585	7.02074E-09
Residual	14	94.8225382	6.773038443		
Total	16	1385.842871			

Predictors	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%
Intercept	53.58088768	2.369503748	22.61270434	2.02E-12	48.4988076	58.66296776
Year	-1.50749033	0.140053125	-10.7637036	3.73E-08	-1.807874405	-1.20710625
%Unemployed	-1.48828923	0.398521926	-3.734522819	0.00222	-2.343033745	-0.633544707

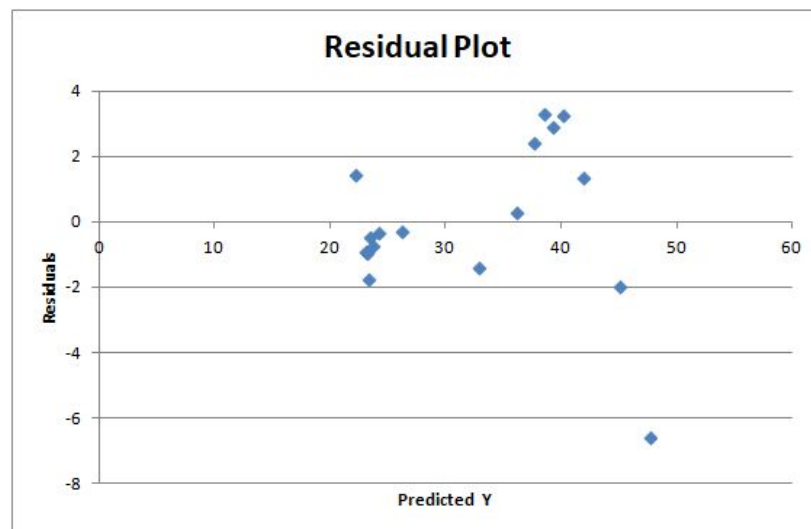
RESIDUAL OUTPUT			
Observation	Predicted Vehicles Stolen per 10,000	Residuals	Standard Residuals
2000	47.67734042	-6.567585488	-2.697801164
2001	45.01642594	-1.927317782	-0.791694325
2002	41.95863433	1.357176861	0.55749458
2003	40.14108375	3.300570945	1.355792651
2004	39.30332358	2.937556781	1.206675439
2005	38.47796581	3.308875541	1.359203973
2006	37.67741287	2.458797599	1.010013046
2007	36.15752013	0.332710762	0.136669326
2008	32.88888755	-1.37974635	-0.566765566
2009	26.19718975	-0.280739433	-0.115320793
2010	24.20600542	-0.298802578	-0.122740685
2011	23.70311032	-0.713304766	-0.293007899
2012	23.47306825	-0.441548076	-0.181376994
2013	23.01978279	-0.873130974	-0.35866054
2014	23.29823953	-1.738861424	-0.714281128
2015	23.14261192	-0.921654764	-0.378592909
2016	22.24283969	1.447003146	0.594392989

Diagnostics

A histogram of standardized residuals reveals what may be a skewed-left distribution. This may suggest errors are non-normal. Given the values of the standardized residuals, we can see that there are no outliers and one unusual observation (year 2000). Given the lack of outliers and our small sample size, we consider this a mild violation.



The plot of residuals does not show obvious signs of heteroscedasticity. The single observation with a large residual corresponds to the unusual observation (year 2000). That said, increasing the sample size could reveal a “fan-out” pattern.



The Durbin-Watson test statistic was calculated to check for autocorrelation.

Observation	Predicted Vehides	Residuals		
	Stolen per 10,000	(et)	(et-et-1)^2	et^2
1	47.67734042	-6.56759	NA	43.1332
2	45.01642594	-1.92732	21.5320844	3.71455
3	41.95863433	1.357177	10.7879051	1.84193
4	40.14108375	3.300571	3.77678057	10.8938
5	39.30332358	2.937557	0.13177928	8.62924
6	38.47796581	3.308876	0.13787762	10.9487
7	37.67741287	2.458798	0.72263251	6.04569
8	36.15752013	0.332711	4.52024524	0.1107
9	32.88888755	-1.37975	2.93250936	1.9037
10	26.19718975	-0.28074	1.2078162	0.07881
11	24.20600542	-0.2988	0.00032628	0.08928
12	23.70311032	-0.7133	0.17181206	0.5088
13	23.47306825	-0.44155	0.0738517	0.19496
14	23.01978279	-0.87313	0.1862638	0.76236
15	23.29823953	-1.73886	0.74948921	3.02364
16	23.14261192	-0.92165	0.66782672	0.84945
17	22.24283969	1.447003	5.6105403	2.09382
Sum		0	53.2097403	94.8225
DW		0.561151		
n=17		k=2		
critical DW range=1.015-1.536				

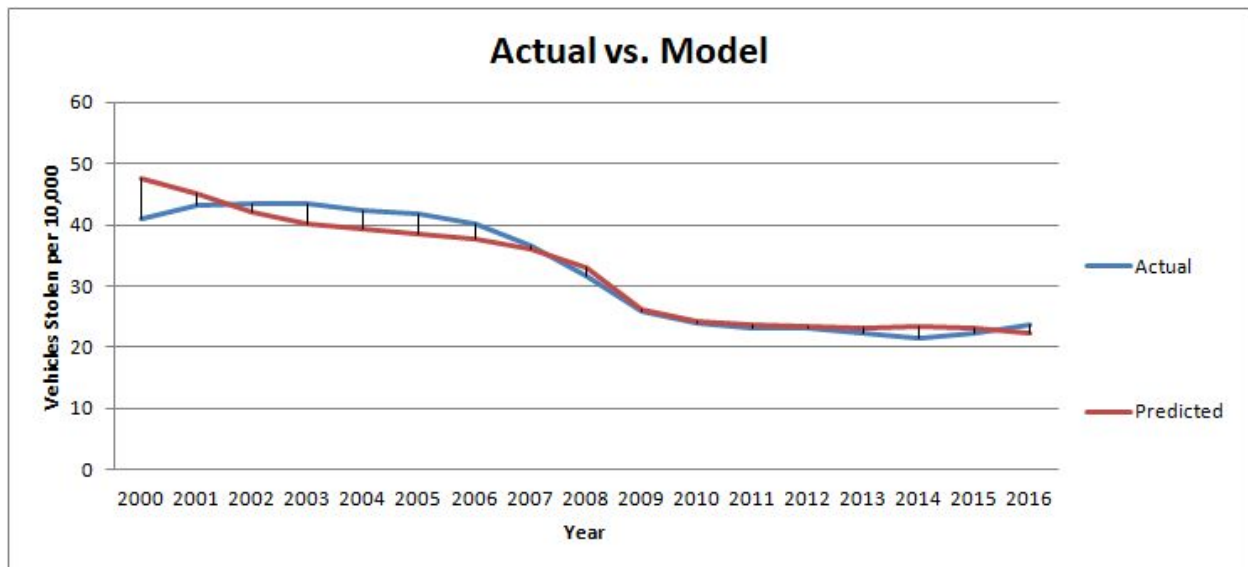
The Durbin-Watson test statistic (0.56) is outside of the acceptable range for $n=17$ and $k=2$. This means the time-series errors are positively autocorrelated. Despite this, the predictor coefficients are still unbiased and consistent. We can conclude that our model's fit is overstated, which is not surprising given the large adjusted r^2 (0.92).

Our final time series model is:

$$(\text{Auto thefts per 10,000}) = 53.581 - 1.488 * (\text{Unemployment rate}) - 1.507 * (\text{Year})$$


Interestingly, the coefficient for Unemployment rate has changed from positive in the initial state regression to negative in the national time series. This means that

an increase in unemployment correlates with a decrease in auto-thefts, nationwide. Additionally, the year coefficient tells us that the general trend for auto-thefts is downward.



Conclusion

From the initial set of input variables that might contribute to the rate of auto theft, it was determined that only 3 factors were statistically significant at a 95% confidence level: the unemployment rate, the percentage of the population with a high school diploma, and the geographic region of the country. It is unsurprising that unemployment has a positive correlation with auto theft while high school education has a negative correlation. The geographic differences were an interesting trend that might merit further study. There may be unidentified factors that cause the Southwestern states to have a higher rate of auto theft and the



Northeast to have a lower rate than the rest of the country. It is also interesting that this study suggests that improving employment and education could be a more productive way to reduce auto thefts than increasing police spending. However, it is important to note that we cannot determine causality from this analysis.

The time series analysis shows a downward trend in auto theft over time, which is a positive sign for the country. One unusual result is the negative correlation between the time series data and the unemployment rate. Further analysis is required to identify the reason behind this result.

Bibliography

1. Federal Bureau of Investigation

<https://ucr.fbi.gov/crime-in-the-u.s/2012/crime-in-the-u.s.-2012/property-crime/motor-vehicle-theft>

2. United States Census Bureau

<https://www.census.gov/support/USACdataDownloads.html>

3. Bureau of Labor Statistics

<https://data.bls.gov/pdq/SurveyOutputServlet>

4. United States Census Bureau U.S. and World Population Clock

<https://www.census.gov/popclock/>

5. U.S. Department of Justice, Federal Bureau of Investigation, Uniform Crime Reports

<https://ucr.fbi.gov/ucr-publications>