PREDICTORS OF VIOLENT CRIME RATE IN THE UNITED STATES

DANIEL J. CHOI (TF: SHAUN DOUGHERTY)

Abstract. The following report seeks to provide explanations of the variation in violent crime rates throughout the United States by analyzing the statistical significance of relationships between the violent crime rate and seven potential predictor variables. A multiple linear regression showed that two of the seven potential predictor variables, teen birth rates and the percentage of the population living in metropolitan areas, proved to be strongly significant ($p < 0.001$). As each of these variables increase, the violent crime rate rises. This finding suggests that states with greater teenage promiscuity and larger cities exhibit higher violent crime rates. Despite demonstrating a strong statistical correlation, these results cannot be used to support a causal relationship. Still, though, the results of this report provide important insights into violent crime rates and potential ways to address them.

1. Introduction

Stories and comic books often play off the fear of violent crime, invariably inserting a super-hero sized punch at just the right time to save any would-be victims. In the real world, however, violent crime is a very real concern for all. This fear-inducing category of offenses includes murders and non-negligent manslaughter, forcible rapes, robberies, and aggravated assaults. These crimes so negatively affect society that it behooves any statistician to perform an analysis in an attempt to shed light on possible sources of this problem. This report examines seven various social and economic factors that could help explain the variance in violent crime rates throughout America's fifty states. Variables that would probably exhibit indirect relationships with violent crime such as divorce rates and level of education are analyzed together with variables that may more directly affect violent crime rates like unemployment and poverty rates. Using a multiple linear regression, this report seeks to uncover strong predictors of violent crime rates in the United States with the hope that its findings may help policymakers decide on ways to best alleviate this problem.

2. Materials and Methods

The response variable "viocrime" is defined as violent crime offenses rate per 100,000 individuals in 2008 by the Federal Bureau of Investigation, U.S. Department of Justice. The candidate predictor variables analyzed in this report include: the percentage of the population who has completed a bachelor's degree in 2004 ("bdegree"), the percentage of children who attend religious services at least weekly as reported by parents in 2003 ("religion"), the percentage of the population living in poverty in 2008 ("poverty")[1], the teen birth rate per 1,000 people in 2006 ("teenbirth")[2], the percentage of the population unemployed in 2009 ("unemployment")[3], the percentage of the population living in metropolitan areas ("metro"), and the divorce rate per 1,000 individuals in 1994 ("divorce"). Each variable includes data for the 50 states in America. No transformations were used, as the data appear to be normally distributed (based on histograms and quantile normal plots).

[1]The federal poverty threshold for a family of four in the 48 contiguous states was $21,834 in 2008.
[2]Population includes females aged 15-19.
[3]Unemployment within the civilian non-institutional population aged 16 years and older.

The most significant predictor variables were selected by a step-down multiple linear regression, where $\alpha < .05$. The step-down multiple linear regression procedure calls for the the response variable, *viocrime*, to be regressed on all the predictor variables, and then for the least significant variable (the variable with the highest p-value) to be removed. This step continues until all remaining predictor variables are significant at the $\alpha < .05$ significance level ($p < 0.05$).

3. Results

Table 1. State Characteristics (N = 50)

Variable		Mean	Median	Stdev	Min	Max
viocrime	(per 100,000)	399.1	345.65	172.7	117.5	729.7
bdegree	(percent)	26.26	25.5	4.66	16.3	37.4
religion	(percent)	55.06	54.4	10.1	28.1	72.2
poverty	(percent)	16.5	15.46	3.52	9.69	27.13
teenbirth	(per 1,000)	41.26	40.1	12.72	18.7	68.4
unemployment	(percent)	7.58	7.4	1.81	4.2	12
metro	(percent)	73.22	74.95	18.29	29.79	100
divorce	(percent)	4.8	4.7	1.29	2.4	9

The summary statistics over 50 states for each of the variables is presented in Table 1 (above). The mean and median values of each variable are all fairly equal. This fact, in addition to examination of the appropriate scatter plots and histograms supports the conclusion that the variables are all sufficiently normal, and that none of the variables needed to be transformed. The maximum for the violent crime rate (South Carolina 729.7/100,000 individuals) does not on first glance seem to be very high, but considering the severity of what a violent crime entails and the population of some of these states, it becomes apparent that this violent crime rate is quite high. For example, South Carolina's rate multiplied by its population yields a total of 32,274 violent crimes per year in South Carolina. The maximum for the divorce variable is a bit of an outlier (Nevada at 9 percent), but given the prevalence of hastily planned marriages that occur in the state, the high divorce rate is understandable. The metro variable exhibits a high variance, with a minimum at 29.79% an a maximum at 100%, but this is not altogether surprising - certain states like New Jersey and Rhode Island are almost completely metropolitan areas while other states like Wyoming and Vermont have little in the way of metropolitan areas.

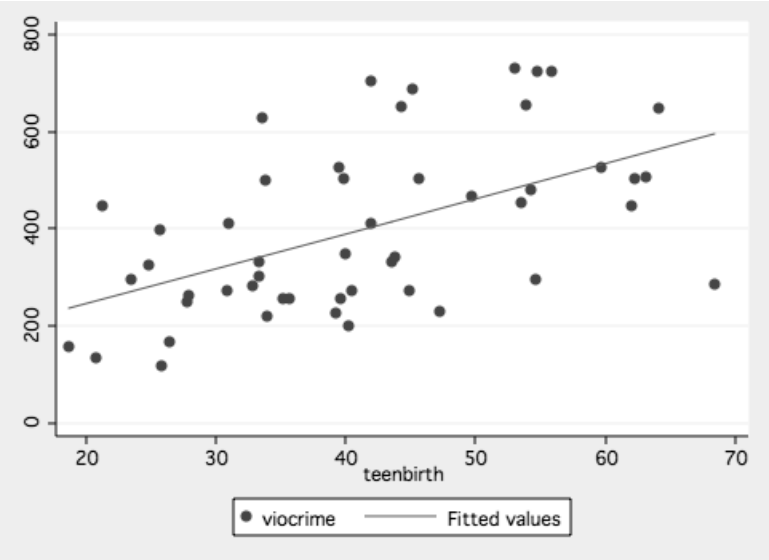
Table 2. Predictors of Violent Crime Rate

Variable	Coefficient
teenbirth	7.932***
	(1.352)
metro	4.787***
	(.9404)
Adjusted r^2	0.5188
Sample Size (N)	50

*P< 0.05 **P< 0.01 ***P< 0.001

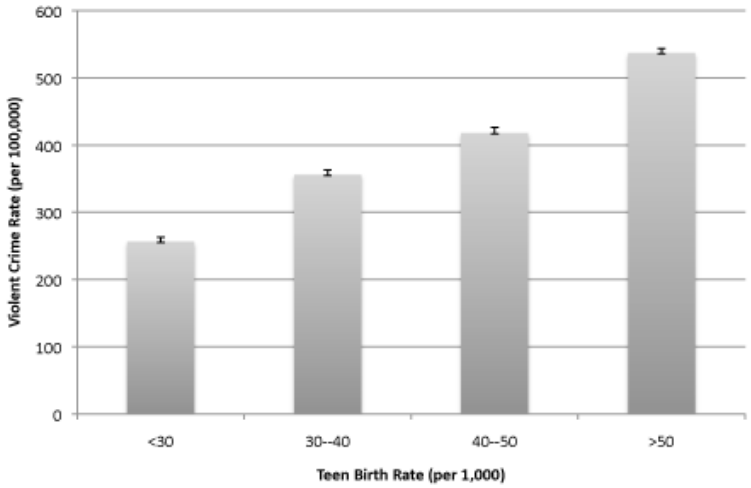
The final regression model is given by the formula: $hat(viocrime) = -278.7212 + 7.932(teenbirth) + 4.787(metro)$. The step-down linear regression procedure reveals that two of the original seven potential predictor variables are significant at the $p < 0.05$ level. Table 2 (above) lists these two variables, *teenbirth* and *metro*, along with their respective coefficients, significance, and standard errors. Both variables are strongly significance, with $p < 0.001$. The value of $r^2 = 0.538$ indicates that the regression model including these two variables can account for 53.8% of the variance in violent crime rate per 100,000 people.

FIGURE 1. Association between Violent Crimes Rate and Teen Birth Rate



Though both *teenbirth* and *metro* have P-values < 0.001, the *teenbirth* variable has a higher t-statistic, so it seems to be the most influential factor. Figures 1 and 2 provide a more detailed analysis of its relationship to the violent crimes rate. Figure 1 (above) displays the association between violent crimes rate and the teen birth rate, fit with a least squares regression line. The points fit somewhat closely to the line, certainly in a linear manner, and no stark outliers are present. The positive slope of the least squares regression line corroborates the positive coefficient for *teenbirth* in Table 2 listed above. In Figure 2 (below), the relationship between violent crimes rate and teen birth rate is plotted for four different groupings of teen birth rate. First, the data are separated into four groups based on teen birth rate (birth rates per 1,000 less than 30, between 30 and 40, between 40 and 50, and greater than 50). Then, the mean violent crime rate in each grouping is plotted, along with standard error bars. As one can see, as the teen birth rate increases, the mean violent crime rate also increases, which further corroborates the positive coefficient for *teenbirth*. Also, the standard error bars are very small, which means that the amount of error is small and therefore the observed differences in means between groups are significant.

FIGURE 2. Effect of Teen Birth Rate on Violent Crimes Rate



4. Conclusions

The final regression model reveals important findings about the relationships between the response and predictor variables. Two predictor variables, the teen birth rate and the percentage of population living in metropolitan areas, were determined to exhibit a statistically significant relationship with the violent crime rate. Teen birth rate was positively correlated with violent crimes rate ($p < 0.001$). An increase in the teen birth rate of 1 per 1,000 individuals was associated with an increase in violent crimes rate of 7.932 per 100,000 individuals holding the metropolitan proportion constant. The proportion of the population living in metropolitan areas was also positively correlated with violent crimes rate ($p < 0.001$). An increase in the metropolitan proportion of 1% was associated with an increase in violent crimes rate of 4.787 per 100,000 individuals holding the teen birth rate constant.

It is important to remark upon the various limitations and weaknesses of this regression of the violent crimes rate. First, though a statistically significant relationship was observed between the predictor and response variables, this does not imply or support a causal relationship, although causation may exist in one or more of these cases. To support causation, a rigorous study would have performed, i.e., a carefully designed randomized experiment. The existence of multiple studies corroborating these relationships and causations would help support the argument for causality. Another weakness is the possibility of confounding variables. Perhaps people in a community with many teen births would have different morals than people in a community with fewer teen births, and it is actually these morals that can accurately predict the rate of violent crimes. Unfortunately, morals are difficult to quantify, which is why this study has instead focused on possible indicators of morals (religious activity, teen birth rate, divorce rate). The data also may give rise to ecological fallacies, in that individuals who give birth as teens or live in metropolitan areas may be believed to be more likely to commit a violent crime. This, however, would be an incorrect assumption, as the nature of specific individuals cannot be based solely on aggregate statistics.

The conclusions detailed in this report have widespread policy implications, especially as states continue try to combat crime in an increasingly violent world. Prior to implementing any policies aimed at addressing the significant predictor variables of this study, however, more studies must be undertaken to support these claims. Future studies should attempt to establish causation and to further analyze the effects of living in metropolitan areas to try to pinpoint which effects are most closely associated with violent crimes. Once causation and the correct predictor variables are determined, states can begin to enact policies to fight the rise in violent crimes.

References

[1] State Master <http://www.statemaster.com>
[2] Kaiser State Health Facts <http://www.statehealthfacts.org/index.jsp>
[3] Centers for Disease Control and Prevention <http://www.divorcereform.org/94staterates.html>
[4] Analysis Software: Stata/SE 10.0

Appendix A. Original Regression

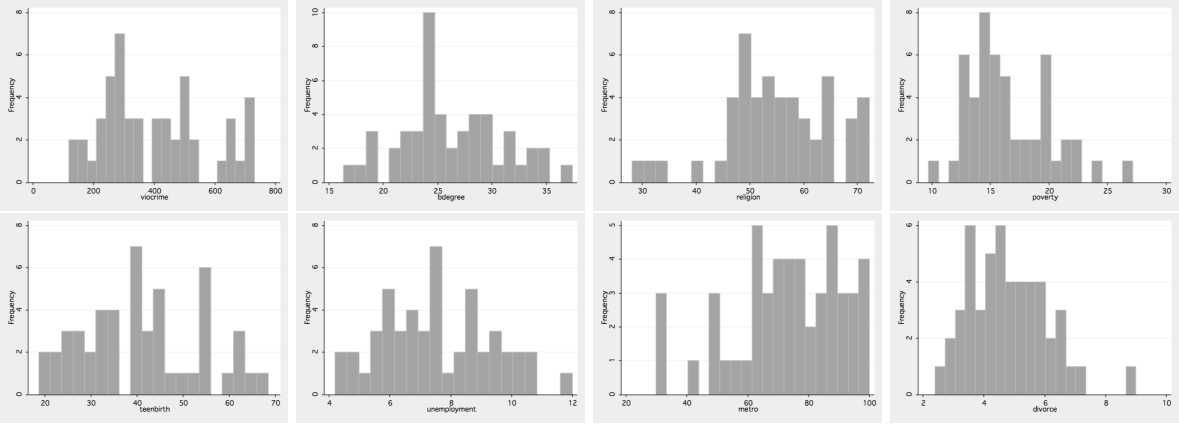
Source	SS	df	MS
Model	816258.717	7	116608.388
Residual	645575.225	42	15370.8387
Total	1461833.94	49	29833.3458

Number of obs = 50
F(7, 42) = 7.59
Prob > F = 0.0000
R-squared = 0.5584
Adj R-squared = 0.4848
Root MSE = 123.98

viocrime	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
bdegree	1.095181	6.727715	0.16	0.871	-12.4819	14.67226
religion	-.4227962	2.692931	-0.16	0.876	-5.85735	5.011758
poverty	-11.49933	10.11164	-1.14	0.262	-31.90545	8.906785
teenbirth	11.84503	3.971902	2.98	0.005	3.829406	19.86065
unemployment	13.46418	12.9173	1.04	0.303	-12.60399	39.53235
metro	4.240531	1.307086	3.24	0.002	1.602725	6.878337
divorce	-20.32316	27.86911	-0.73	0.470	-76.5653	35.91898
_cons	-220.3084	337.2178	-0.65	0.517	-900.8416	460.2247

This first regression which includes all seven original, untransformed potential predictor variables depicts their influence upon our response variable, *viocrime*. The r^2 value of .558 indicates that this regression model already predicts the values of *viocrime* with some strength. The F-statistic, 7.59 (F(7,42)) yields us a significant P-value < 0.0001. Finally, we can observe that both variables *teenbirth* and *metro* are significant at the $\alpha = 0.05$ level.

Appendix B. Evaluation of Possible Transformations



From the above histograms of the outcome variable and each predictor variable, we can conclude that no variable is skewed strongly enough to warrant a transformation. All of the data have fairly normal distributions.

Appendix C. Regression Results

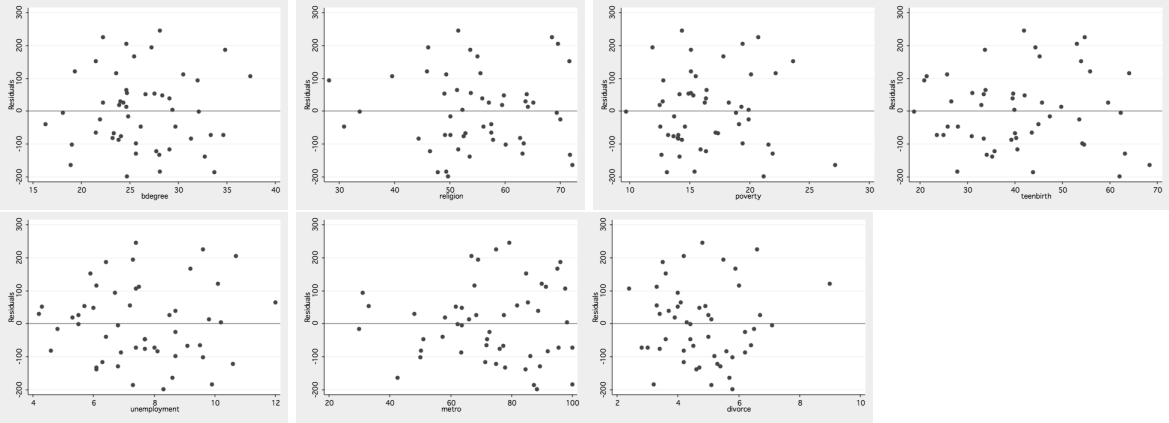
Source	SS	df	MS
Model	816258.717	7	116608.388
Residual	645575.225	42	15370.8387
Total	1461833.94	49	29833.3458

Number of obs	= 50
F(7, 42)	= 7.59
Prob > F	= 0.0000
R-squared	= 0.5584
Adj R-squared	= 0.4848
Root MSE	= 123.98

viocrime	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
bdegree	1.095181	6.727715	0.16	0.871	-12.4819	14.67226
religion	-.4227962	2.692931	-0.16	0.876	-5.85735	5.011758
poverty	-11.49933	10.11164	-1.14	0.262	-31.90545	8.906785
teenbirth	11.84503	3.971902	2.98	0.005	3.829406	19.86065
unemployment	13.46418	12.9173	1.04	0.303	-12.60399	39.53235
metro	4.240531	1.307086	3.24	0.002	1.602725	6.878337
divorce	-20.32316	27.86911	-0.73	0.470	-76.5653	35.91898
_cons	-220.3084	337.2178	-0.65	0.517	-900.8416	460.2247

Because no variables needed to be transformed, these results mirror those found in Appendix A.

Appendix D. Evaluation of Possible Outliers



No outlier is really clear enough to be excluded, given the above graphs of residuals and each predictor variable. In the divorce scatterplot, Nevada is fairly far to the right (9% divorce rate), but it is not extreme enough to warrant exclusion. Also, Nevada is not close to being an outlier for any of the other variables, so dropping that individual could be bad for our model. The same is true of Mississippi in the poverty graph (Mississippi's poverty rate is 27).

Appendix E. Modeling Process Results

The step-down regression results for each step are included below:

Step 1: All variables

Source	SS	df	MS		
Model	816258.717	7	116608.388	Number of obs =	50
Residual	645575.225	42	15370.8387	F(7, 42) =	7.59
Total	1461833.94	49	29833.3458	Prob > F =	0.0000
				R-squared =	0.5584
				Adj R-squared =	0.4848
				Root MSE =	123.98

viocrime	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
bdegree	1.095181	6.727715	0.16	0.871	-12.4819	14.67226
religion	-.4227962	2.692931	-0.16	0.876	-5.85735	5.011758
poverty	-11.49933	10.11164	-1.14	0.262	-31.90545	8.906785
teenbirth	11.84503	3.971902	2.98	0.005	3.829406	19.86065
unemployment	13.46418	12.9173	1.04	0.303	-12.60399	39.53235
metro	4.240531	1.307086	3.24	0.002	1.602725	6.878337
divorce	-20.32316	27.86911	-0.73	0.470	-76.5653	35.91898
_cons	-220.3084	337.2178	-0.65	0.517	-900.8416	460.2247

Step 2: Remove religion

Source	SS	df	MS		
Model	815879.83	6	135979.972	Number of obs =	50
Residual	645954.112	43	15022.1887	F(6, 43) =	9.05
Total	1461833.94	49	29833.3458	Prob > F =	0.0000
				R-squared =	0.5581
				Adj R-squared =	0.4965
				Root MSE =	122.57

viocrime	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
bdegree	1.570234	5.940358	0.26	0.793	-10.40964	13.55011
poverty	-11.34091	9.946407	-1.14	0.261	-31.39975	8.717932
teenbirth	11.57946	3.552726	3.26	0.002	4.414703	18.74421
unemployment	13.88964	12.48579	1.11	0.272	-11.29037	39.06964
metro	4.191279	1.254409	3.34	0.002	1.661522	6.721036
divorce	-18.50442	25.05836	-0.74	0.464	-69.03941	32.03057
_cons	-256.0736	245.8057	-1.04	0.303	-751.7881	239.6408

Step 3: Remove bdegree

Source	SS	df	MS		
Model	814830.199	5	162966.04	Number of obs =	50
Residual	647003.744	44	14704.6305	F(5, 44) =	11.08
Total	1461833.94	49	29833.3458	Prob > F =	0.0000
				R-squared =	0.5574
				Adj R-squared =	0.5071
				Root MSE =	121.26

viocrime	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
poverty	-11.82157	9.674867	-1.22	0.228	-31.31998	7.676843
teenbirth	11.41835	3.462864	3.30	0.002	4.439407	18.39729
unemployment	13.41779	12.22622	1.10	0.278	-11.22254	38.05813
metro	4.344414	1.100798	3.95	0.000	2.125902	6.562926
divorce	-19.7554	24.34591	-0.81	0.421	-68.82137	29.31057
_cons	-201.8884	134.2059	-1.50	0.140	-472.3627	68.58592

Step 4: Remove divorce

Source	SS	df	MS		
Model	805148.008	4	201287.002	Number of obs	50
Residual	656685.934	45	14593.0208	F(4, 45)	13.79
				Prob > F	0.0000
Total	1461833.94	49	29833.3458	R-squared	0.5508
				Adj R-squared	0.5108
				Root MSE	120.8

viocrime	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
poverty	-7.330009	7.904816	-0.93	0.359	-23.25113	8.591108
teenbirth	9.19099	2.103086	4.37	0.000	4.955157	13.42682
unemployment	9.571345	11.22705	0.85	0.398	-13.0411	32.18379
metro	4.57749	1.058622	4.32	0.000	2.445317	6.709664
_cons	-266.9127	107.2466	-2.49	0.017	-482.9184	-50.90706

Step 5: Remove unemployment

Source	SS	df	MS		
Model	794541.811	3	264847.27	Number of obs	50
Residual	667292.132	46	14506.3507	F(3, 46)	18.26
				Prob > F	0.0000
Total	1461833.94	49	29833.3458	R-squared	0.5435
				Adj R-squared	0.5138
				Root MSE	120.44

viocrime	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
poverty	-5.397976	7.550472	-0.71	0.478	-20.59629	9.800336
teenbirth	9.068738	2.091952	4.34	0.000	4.857858	13.27962
metro	4.936754	.9682412	5.10	0.000	2.987786	6.885723
_cons	-247.522	104.4952	-2.37	0.022	-457.8599	-37.18403

Step 6: Remove poverty

Source	SS	df	MS		
Model	787127.481	2	393563.741	Number of obs	50
Residual	674706.462	47	14355.4566	F(2, 47)	27.42
				Prob > F	0.0000
Total	1461833.94	49	29833.3458	R-squared	0.5385
				Adj R-squared	0.5188
				Root MSE	119.81

viocrime	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
teenbirth	7.931995	1.352347	5.87	0.000	5.211424	10.65257
metro	4.787068	.9404028	5.09	0.000	2.895221	6.678914
_cons	-278.7212	94.45107	-2.95	0.005	-468.7322	-88.71016

Appendix F. The Final Model

Source	SS	df	MS
Model	787127.481	2	393563.741
Residual	674706.462	47	14355.4566
Total	1461833.94	49	29833.3458

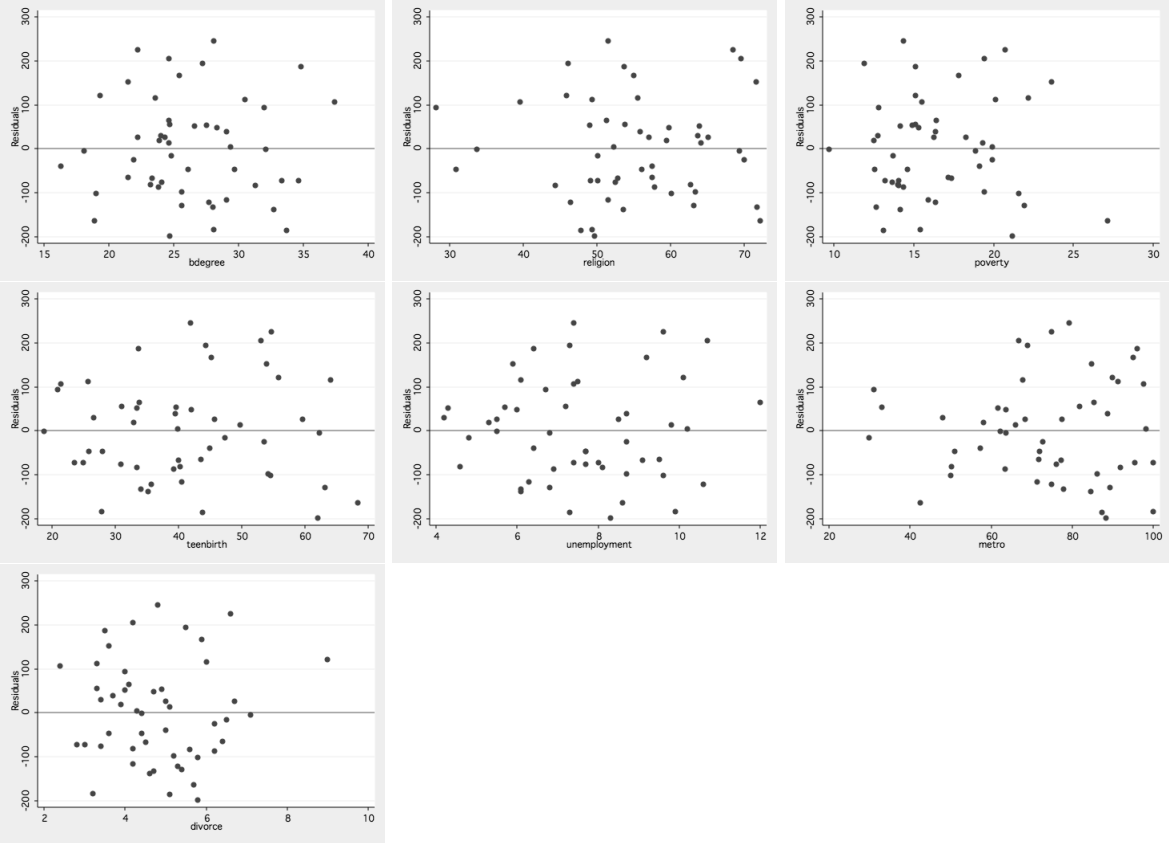
Number of obs	= 50
F(2, 47)	= 27.42
Prob > F	= 0.0000
R-squared	= 0.5385
Adj R-squared	= 0.5188
Root MSE	= 119.81

viocrime	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
teenbirth	7.931995	1.352347	5.87	0.000	5.211424	10.65257
metro	4.787068	.9404028	5.09	0.000	2.895221	6.678914
_cons	-278.7212	94.45107	-2.95	0.005	-468.7322	-88.71016

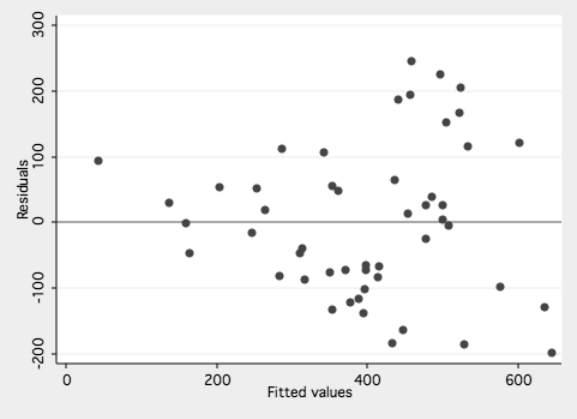
In this final multiple regression model, two variables remain that are both strongly statistically significant at the $\alpha = 0.05$ level - *teenbirth* and *metro*. The adjusted r^2 value (.5188) indicates that the predictor variables can explain 51.9% of the variation of the response variable, *viocrime*. Because it is over .5, this can be called a strong model. The F-statistic is 27.42 (F(2,47)), which gives us a P-value < 0.0001. This indicates that we have a good overall model. The formula of the linear regression is given by: $hat(viocrime) = -278.7212 + 7.932(teenbirth) + 4.787(metro)$

Appendix G. Assumptions Check

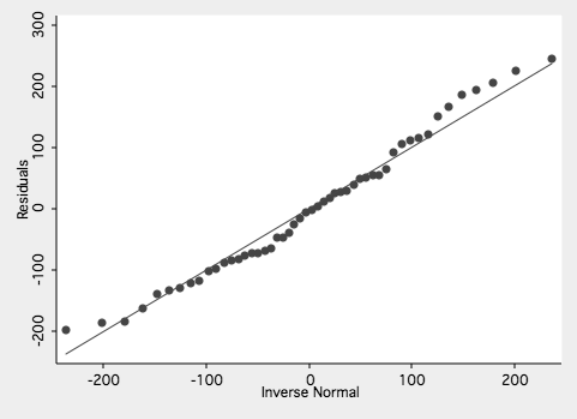
G.1. **Check for Linearity:** Observe random scatter of points in plots of residuals by predictor variables. We also see linearity (no curvature). Check.



G.2. **Check for Consistency of Variance:** Observe random scattering of points in graph of residuals by predicted values. Check.



G.3. **Check Normality of Residuals:** The points on the quantile normal graph follow closely a line. Check.



Appendix H. Standard Error Bar Graph Data

Group	N	Mean	Stdev	SE
<30	10	256.07	113.346372	7.08317993
30--40	14	355.942857	132.320611	7.01354128
40--50	13	417.8	173.329388	8.47984639
>50	13	536.961538	151.60362	6.54241338