REGRESSION DIAGNOSTICS

Richard Lee Rogers

The Problem: The Error Term

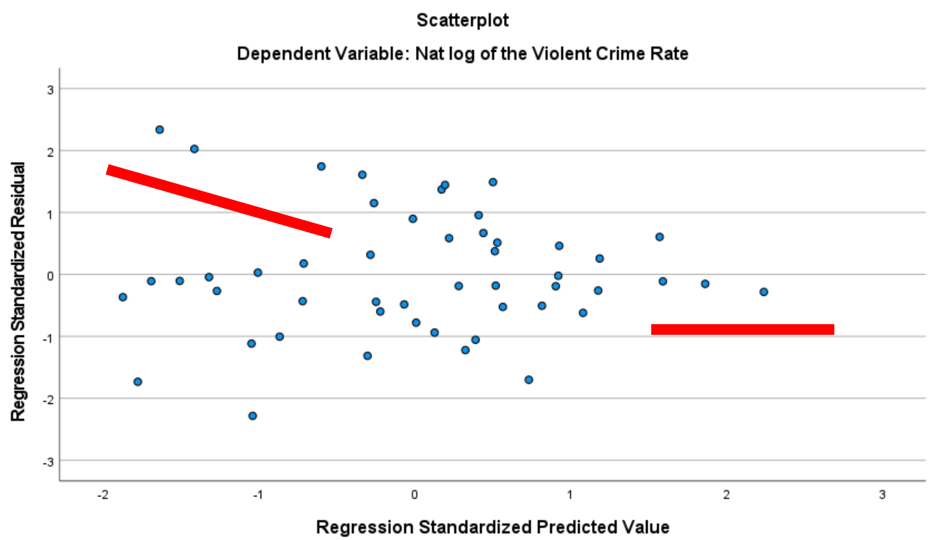
$$y = b_o + b_1 x + e$$

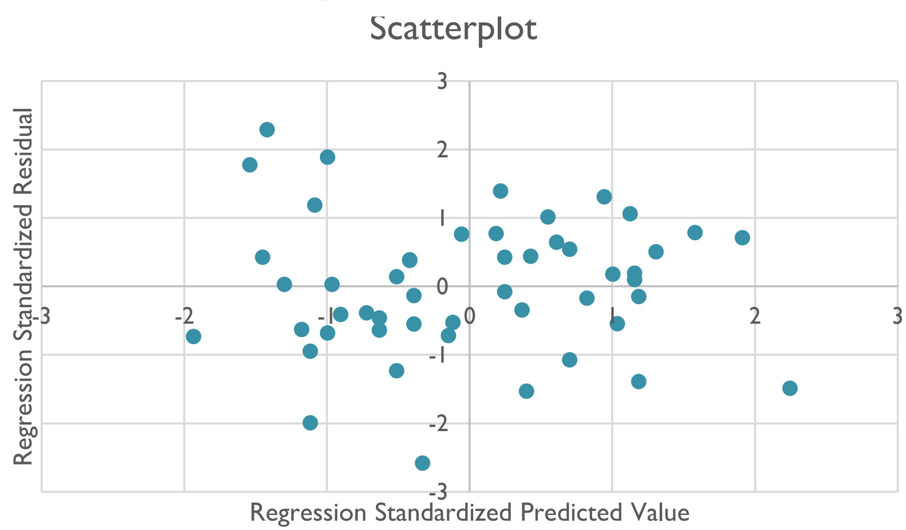
The Problem: The Error Term

$$y = b_o + b_1 x + e$$

homoskedastic

Comparing Two Residual Plots





Our Topics

- Residual plots
- Homoskedasticity/heteroskedasticity tests
- Influence and leverage statistics

LNVIOLENT on POVERTY_RATE

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.390[a]	.152	.135	.36068

a. Predictors: (Constant), Poverty Rate

b. Dependent Variable: Nat log of the Violent Crime Rate

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	1.123	1	1.123	8.636	.005[b]
	Residual	6.244	48	.130		
	Total	7.368	49			

a. Dependent Variable: Nat log of the Violent Crime Rate

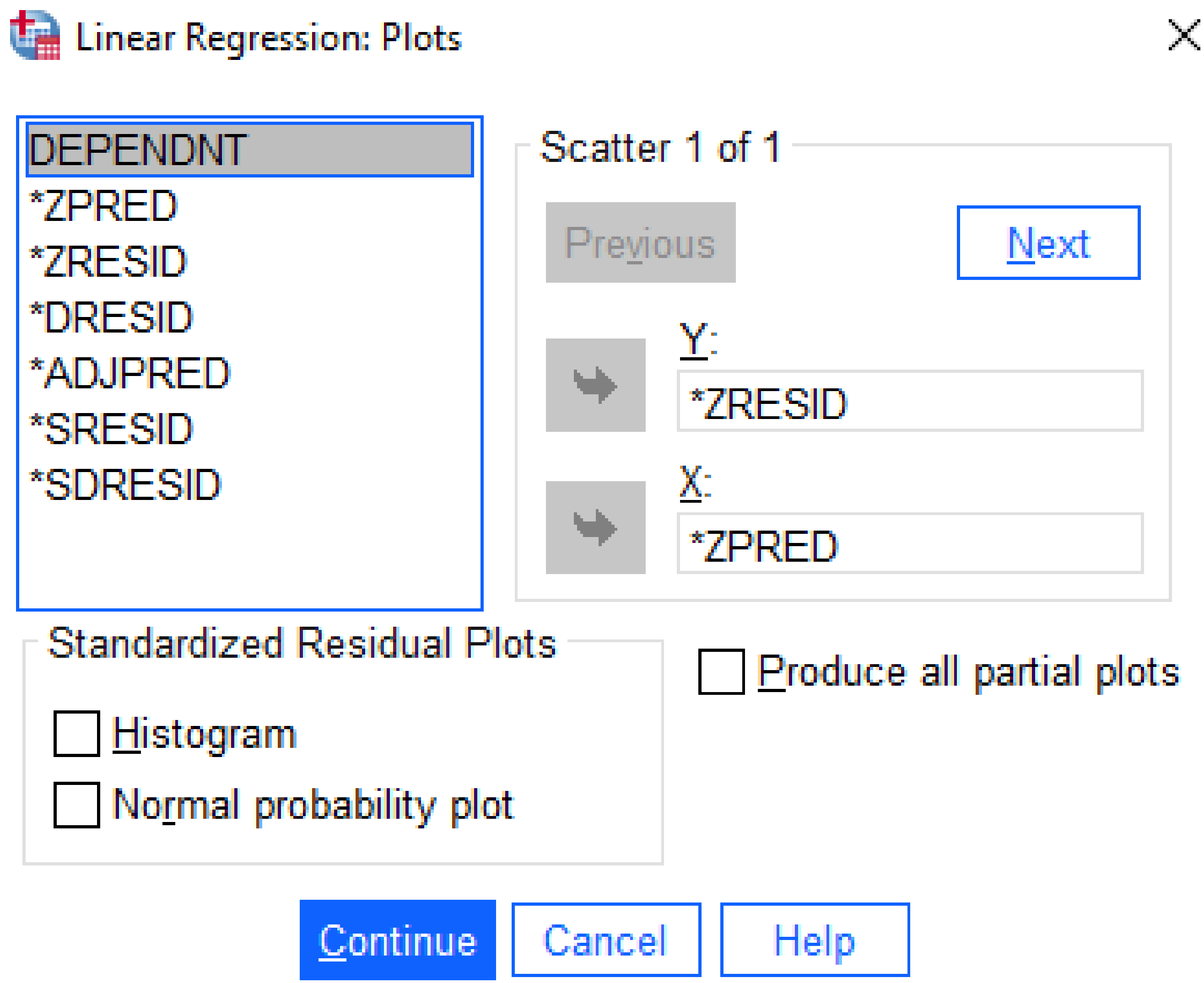
b. Predictors: (Constant), Poverty Rate

Coefficients[a]

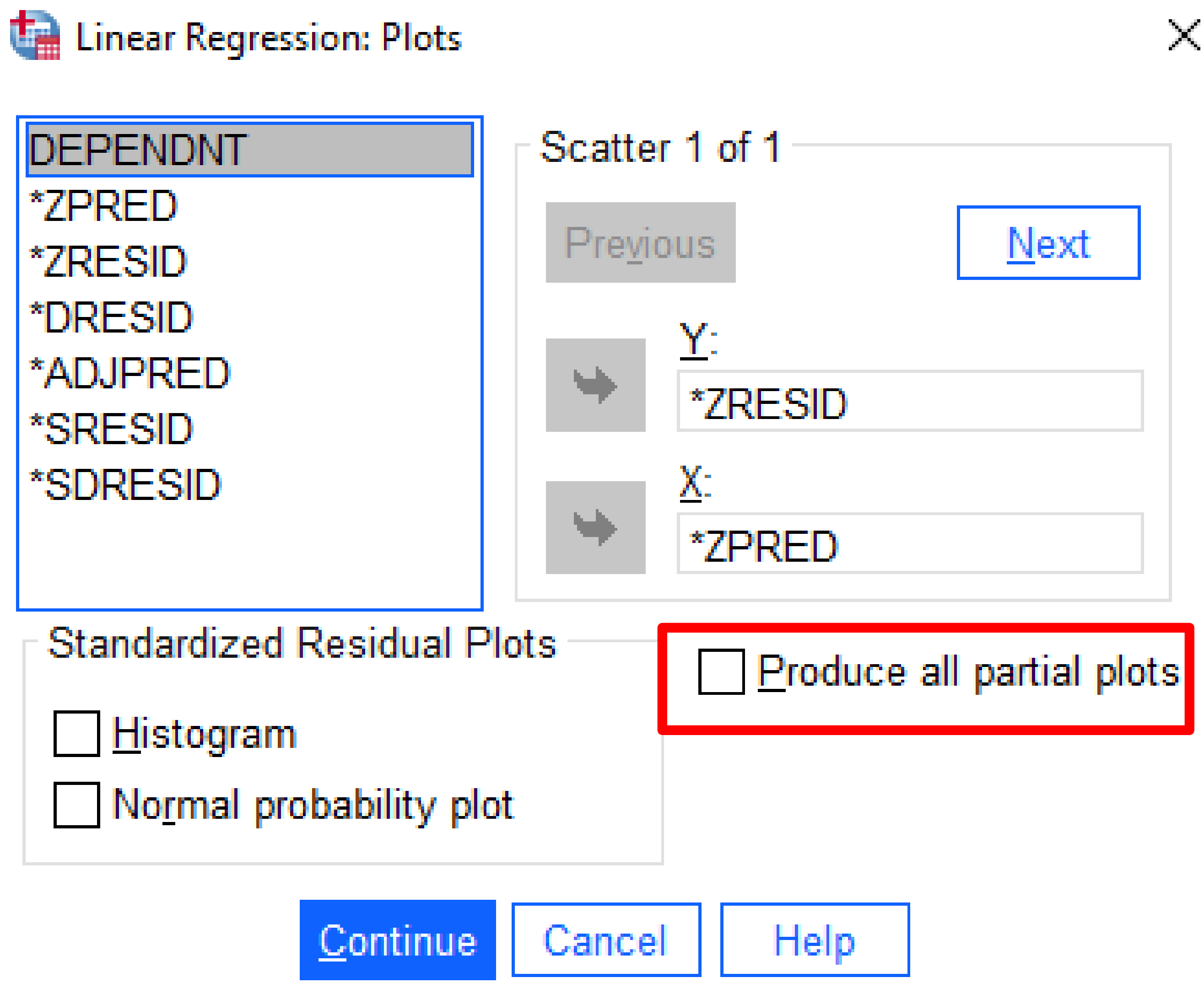
Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Collinearity Statistics	
		B	Std. Error	Beta			Tolerance	VIF
1	(Constant)	5.100	.243		21.029	.000		
	Poverty Rate	.046	.016	.390	2.939	.005	1.000	1.000

a. Dependent Variable: Nat log of the Violent Crime Rate

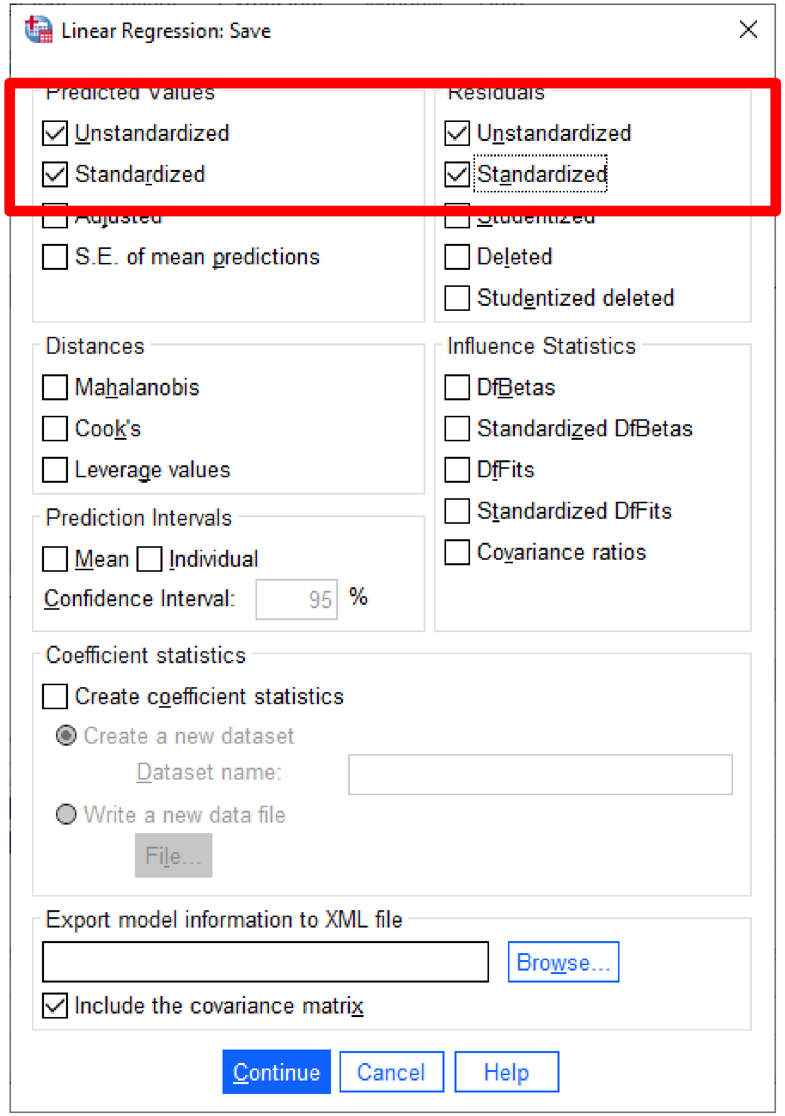
#1: DIRECT TO PLOTS

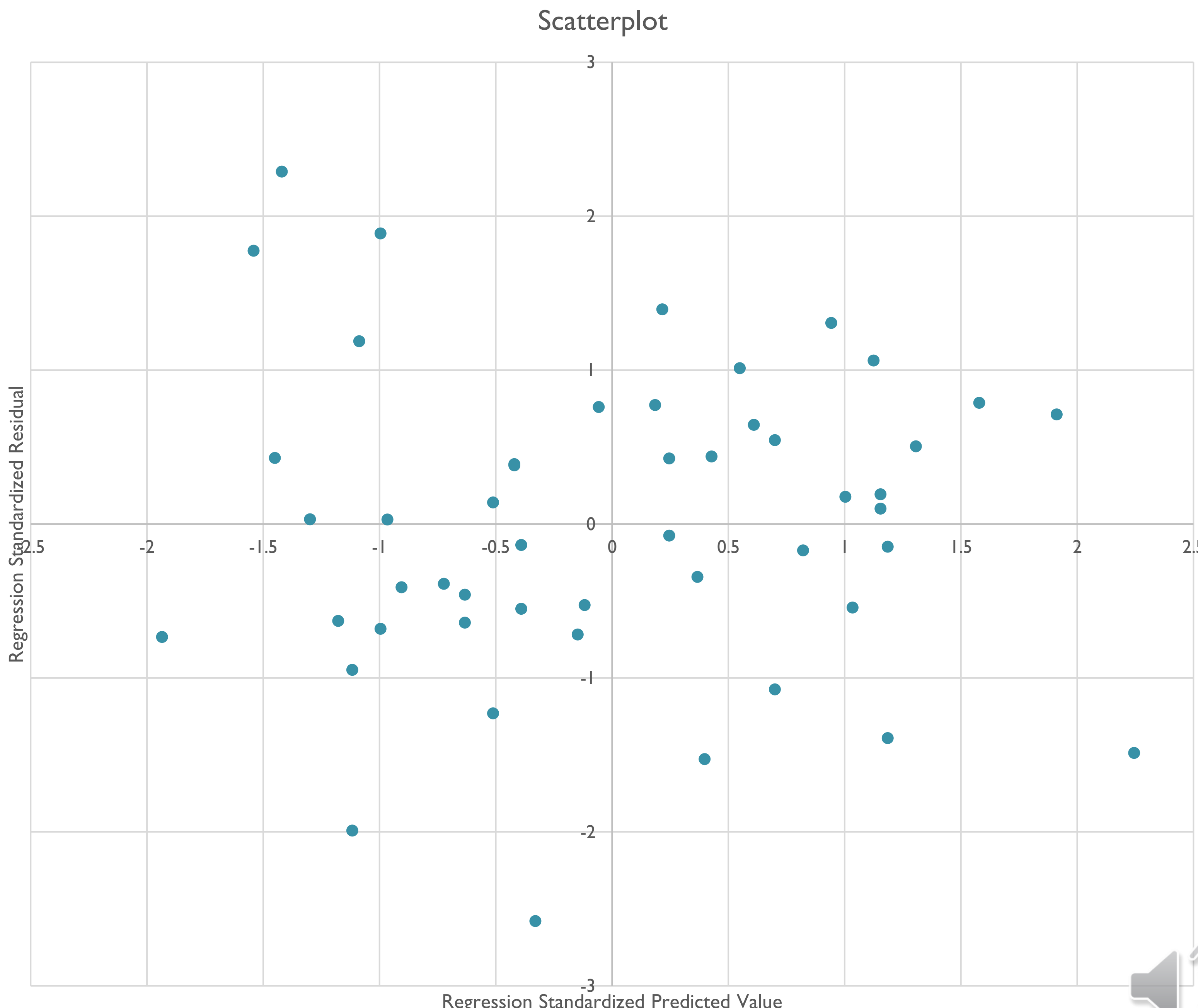


#1: DIRECT TO PLOTS



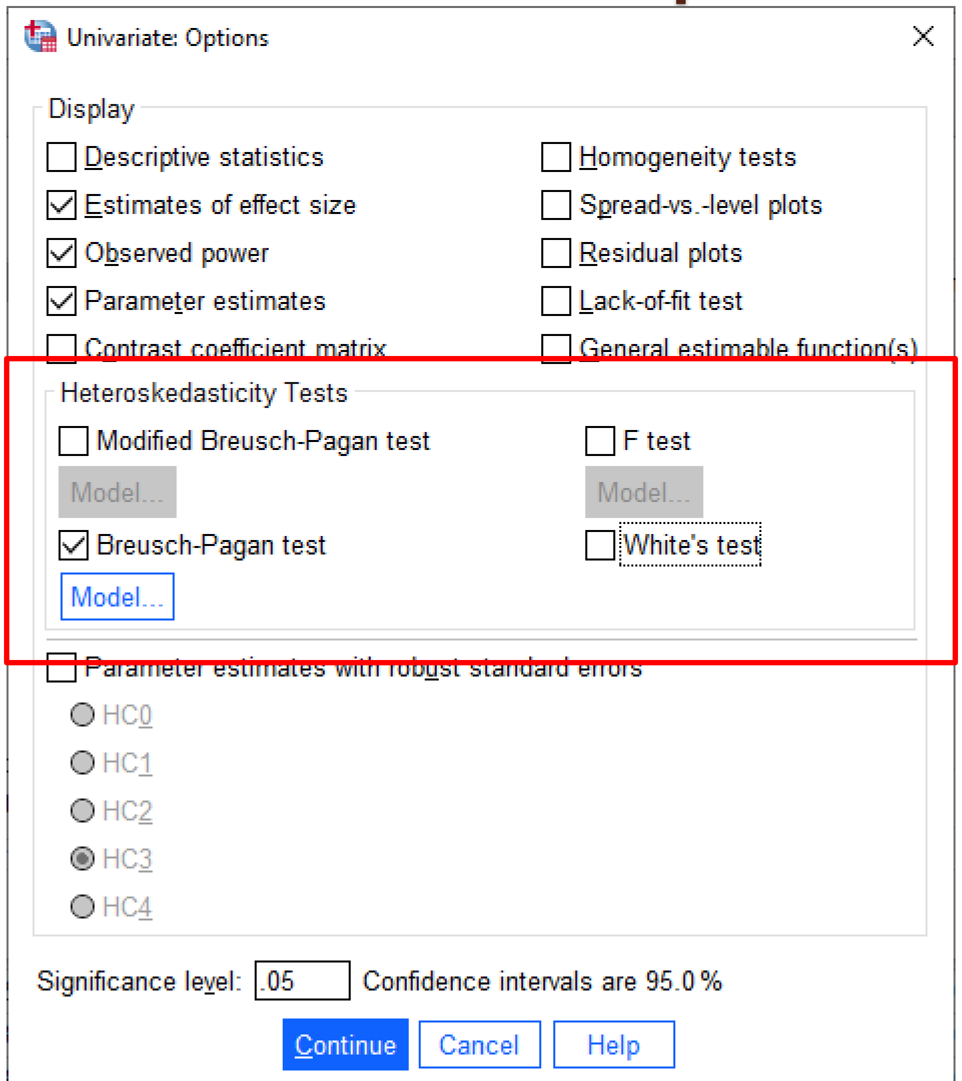
#2: SAVE VALUES TO DATASET





Scatterplot
3
2
1
0
-1
-2
-3
-2.5
-2
-1.5
-1
-0.5
0
0.5
1
1.5
2
2.5
Regression Standardized Residual
Regression Standardized Predicted Value

Analyze>General Linear Model>Univariate>Options



Analyze>General Linear Model>Univariate>Options

Breusch-Pagan Test for Heteroskedasticity[a,b,c]

Chi-Square	df	Sig.
1.804	1	.179

a. Dependent variable: Nat log of the Violent Crime Rate

b. Tests the null hypothesis that the variance of the errors does not depend on the values of the independent variables.

c. Predicted values from design: Intercept + PovertyRate

Breusch-Pagan Test for Heteroskedasticity[a,b,c]

Chi-Square	df	Sig.
6.427	1	.011

a. Dependent variable: Nat log of the Violent Crime Rate

b. Tests the null hypothesis that the variance of the errors does not depend on the values of the independent variables.

c. Predicted values from design: Intercept + LnPop

LNVIOLENT on LNPOPTOT

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.375[a]	.141	.123	.36318

a. Predictors: (Constant), lnpoptot

b. Dependent Variable: lnviolent

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	1.037	1	1.037	7.858	.007[b]
	Residual	6.331	48	.132		
	Total	7.368	49			

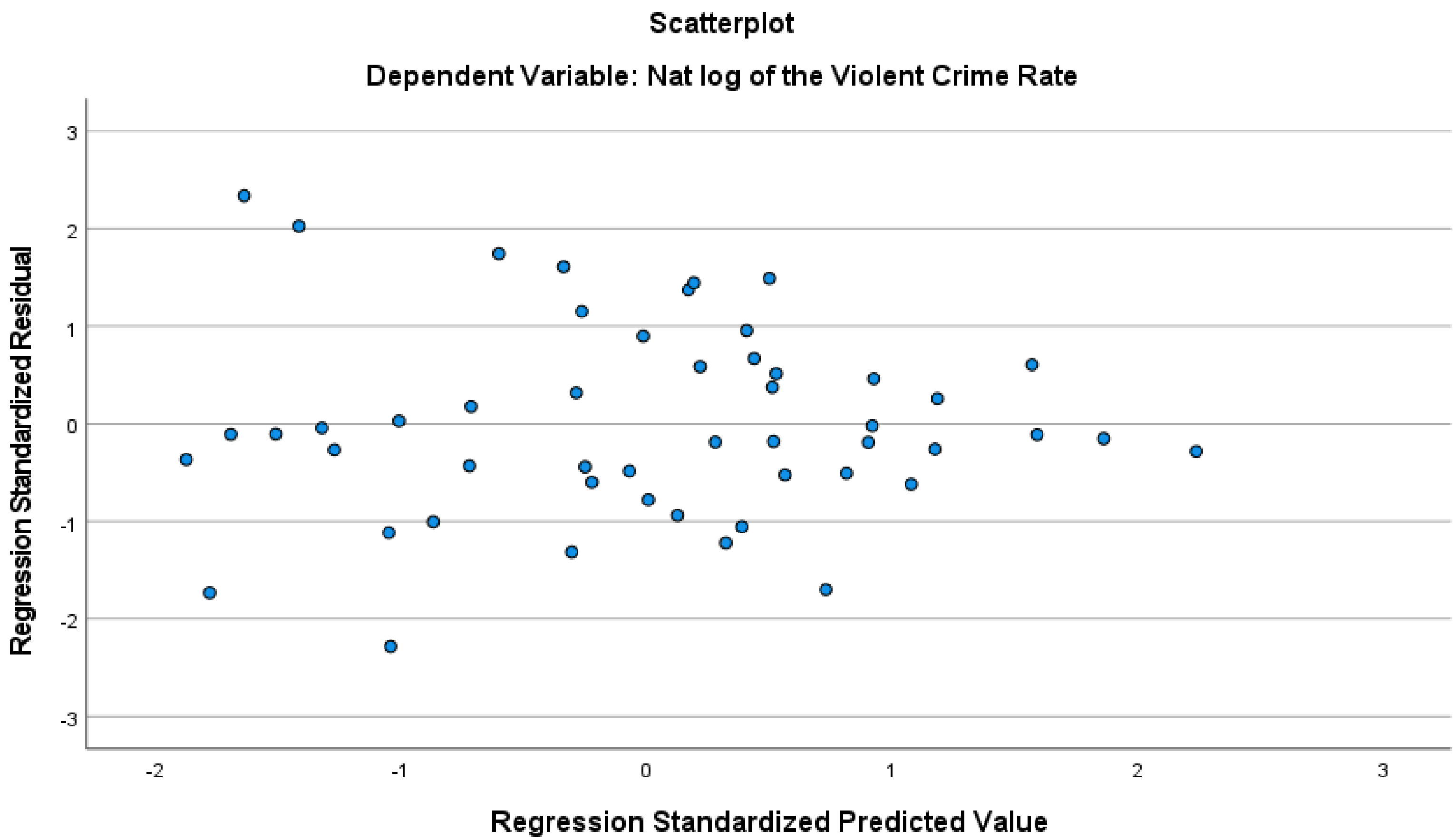
a. Dependent Variable: lnviolent

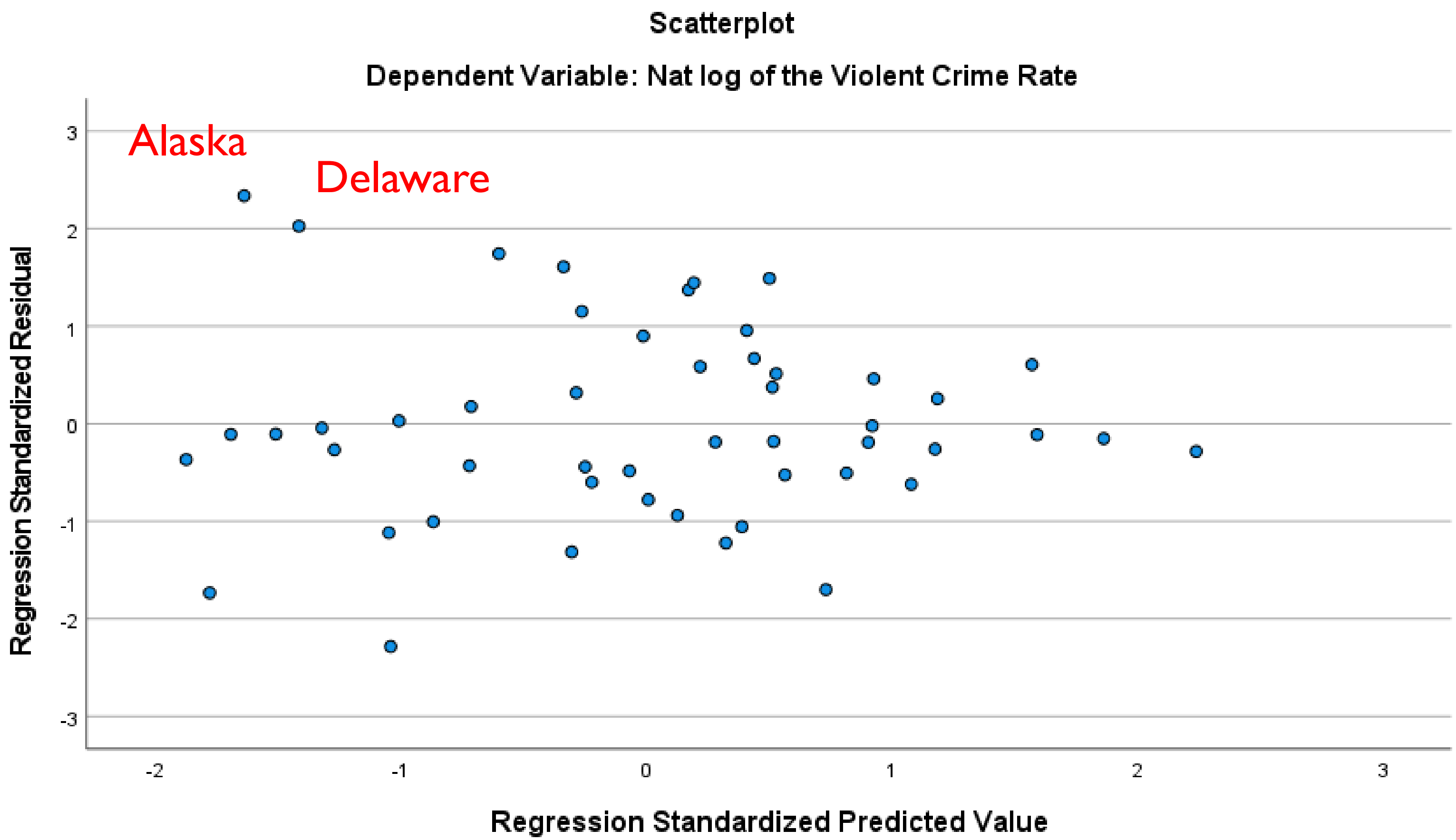
b. Predictors: (Constant), lnpoptot

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	3.636	.773		4.705	.000
	lnpoptot	.143	.051	.375	2.803	.007

a. Dependent Variable: lnviolent

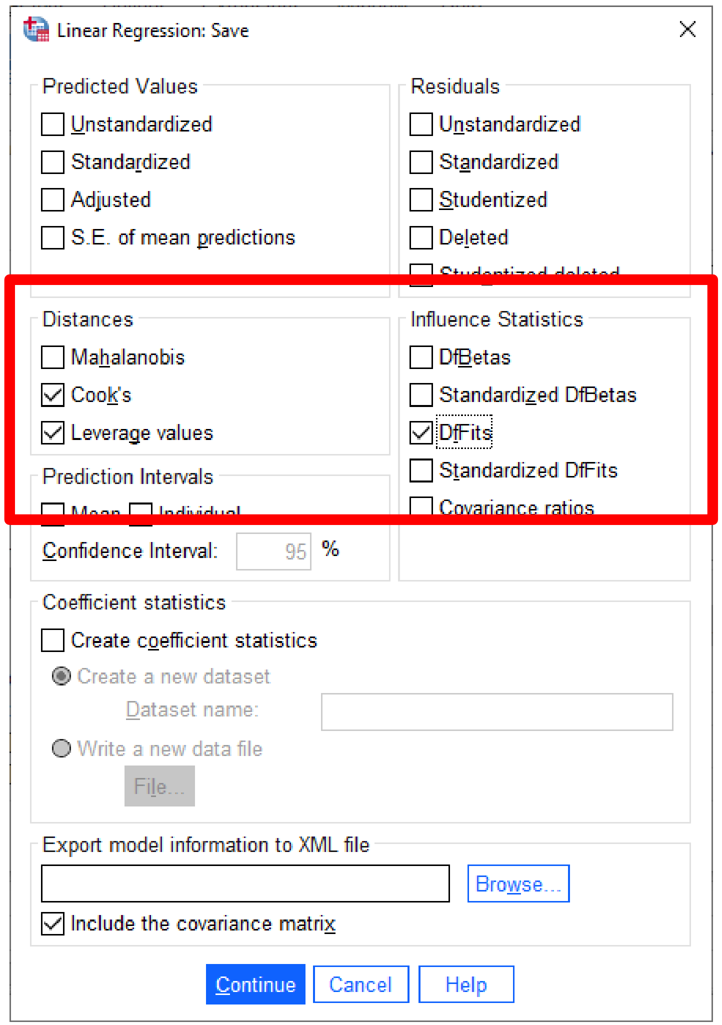

Scatterplot
Dependent Variable: Nat log of the Violent Crime Rate
Regression Standardized Residual
3
2
1
0
-1
-2
-3
-2
-1
0
1
2
3
Regression Standardized Predicted Value


Scatterplot
Dependent Variable: Nat log of the Violent Crime Rate
Alaska
Delaware
Regression Standardized Residual
3
2
1
0
-1
-2
-3
-2
-1
0
1
2
3
Regression Standardized Predicted Value

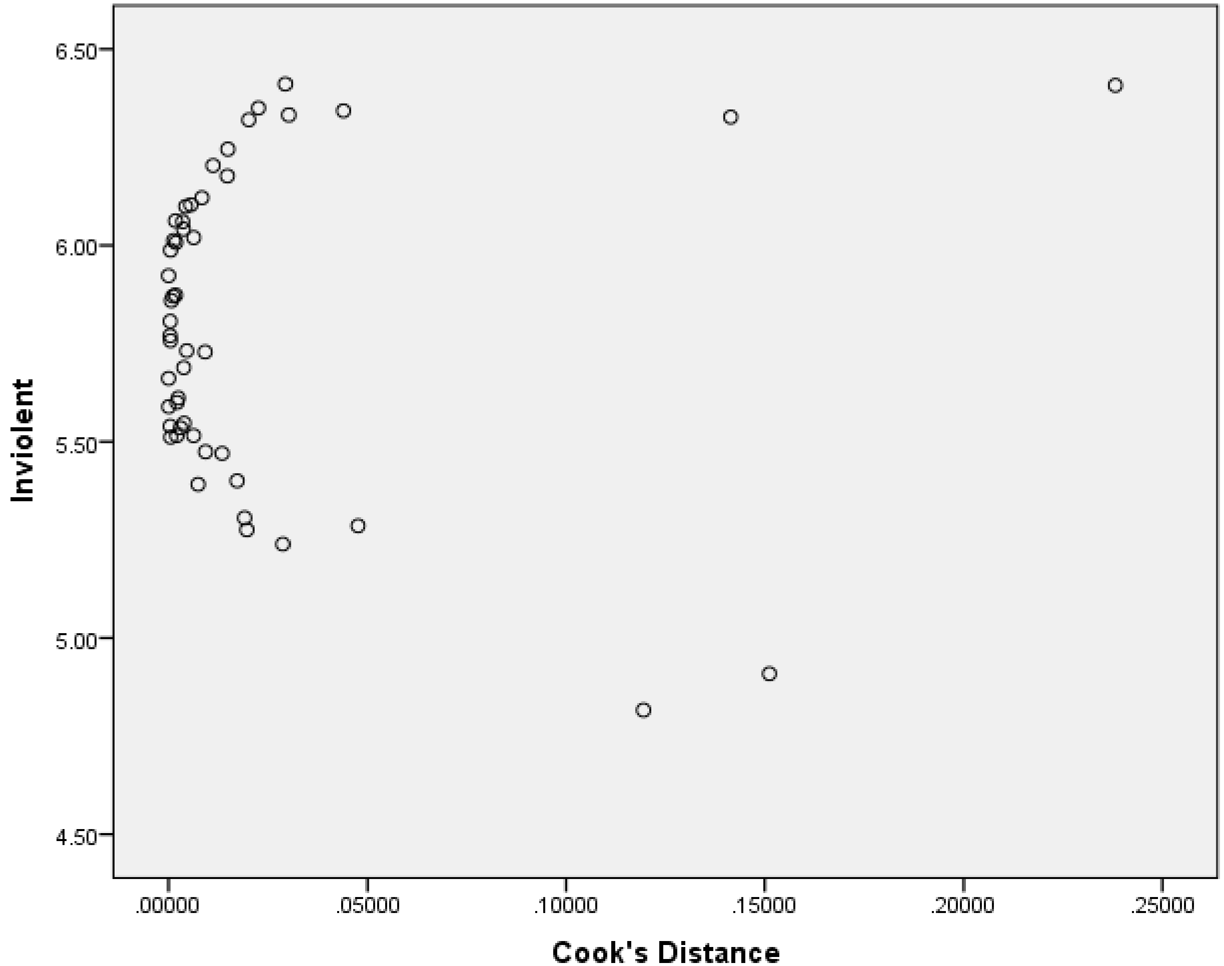
Influence and Leverage Statistics

- Cook's D
- DFITS (DiFits or Dffits)
- Centered Leverage Statistics

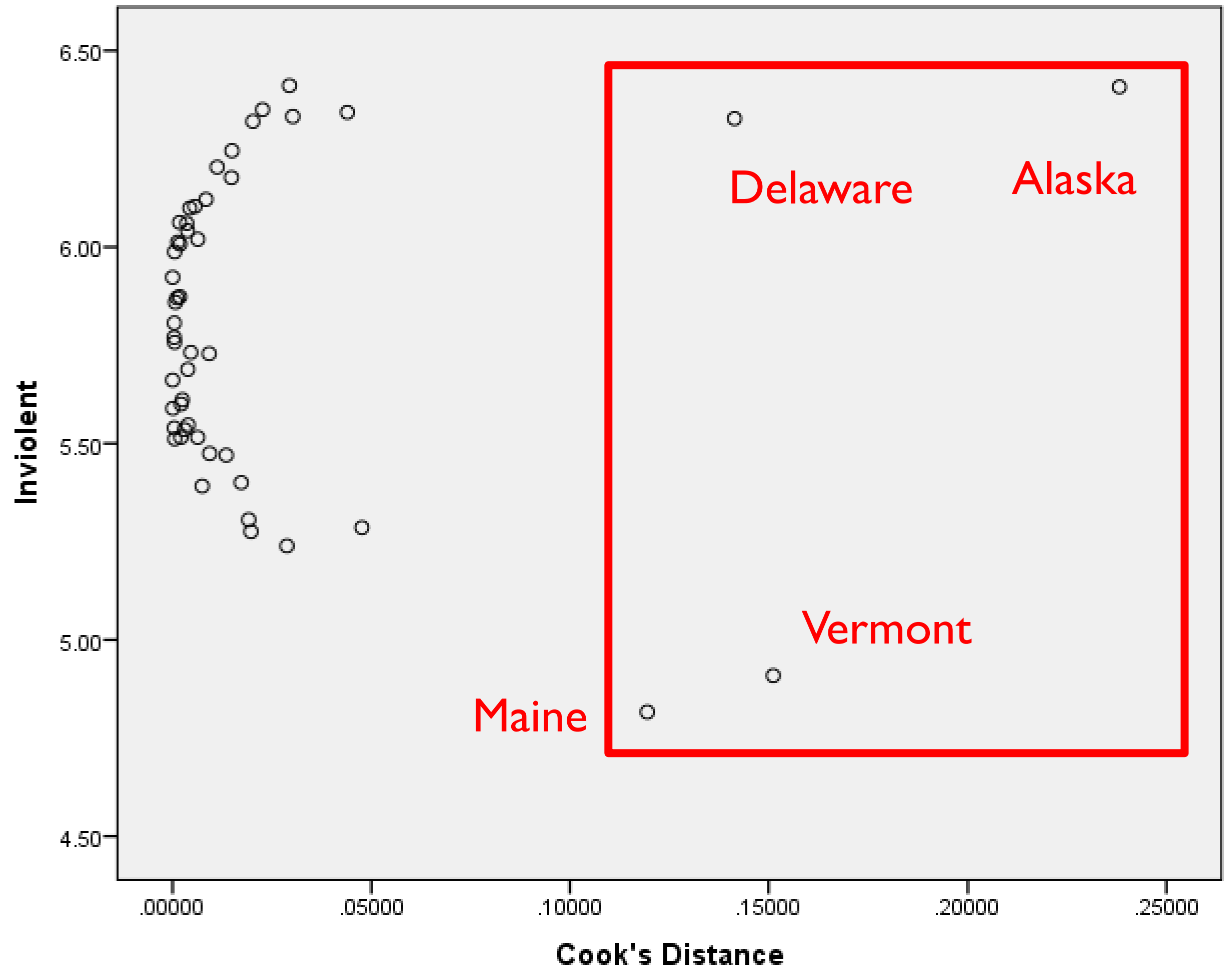
Analyze>Regression>Linear>Save



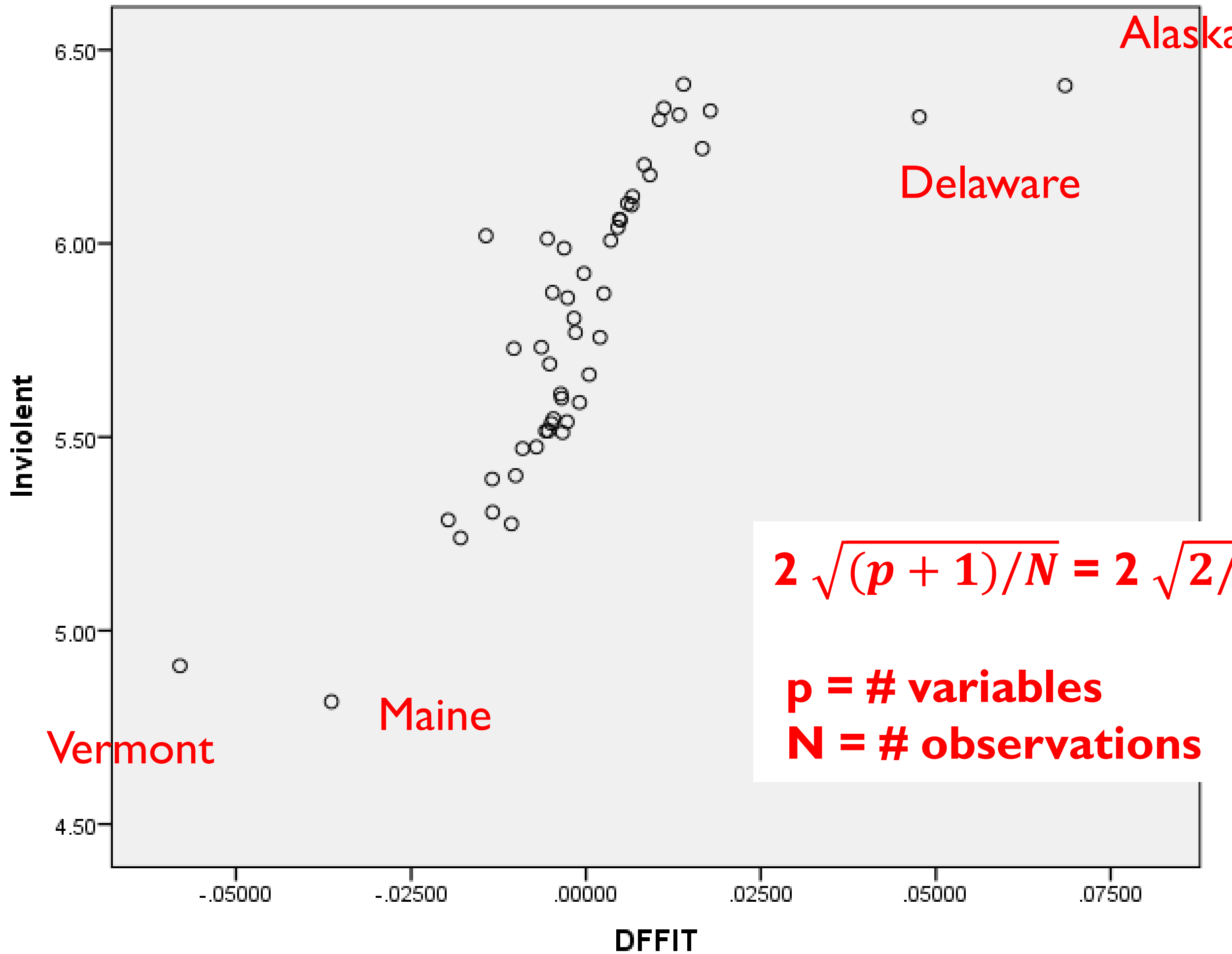
LNVIOLENT on Cook's D



LNVIOLENT on Cook's D



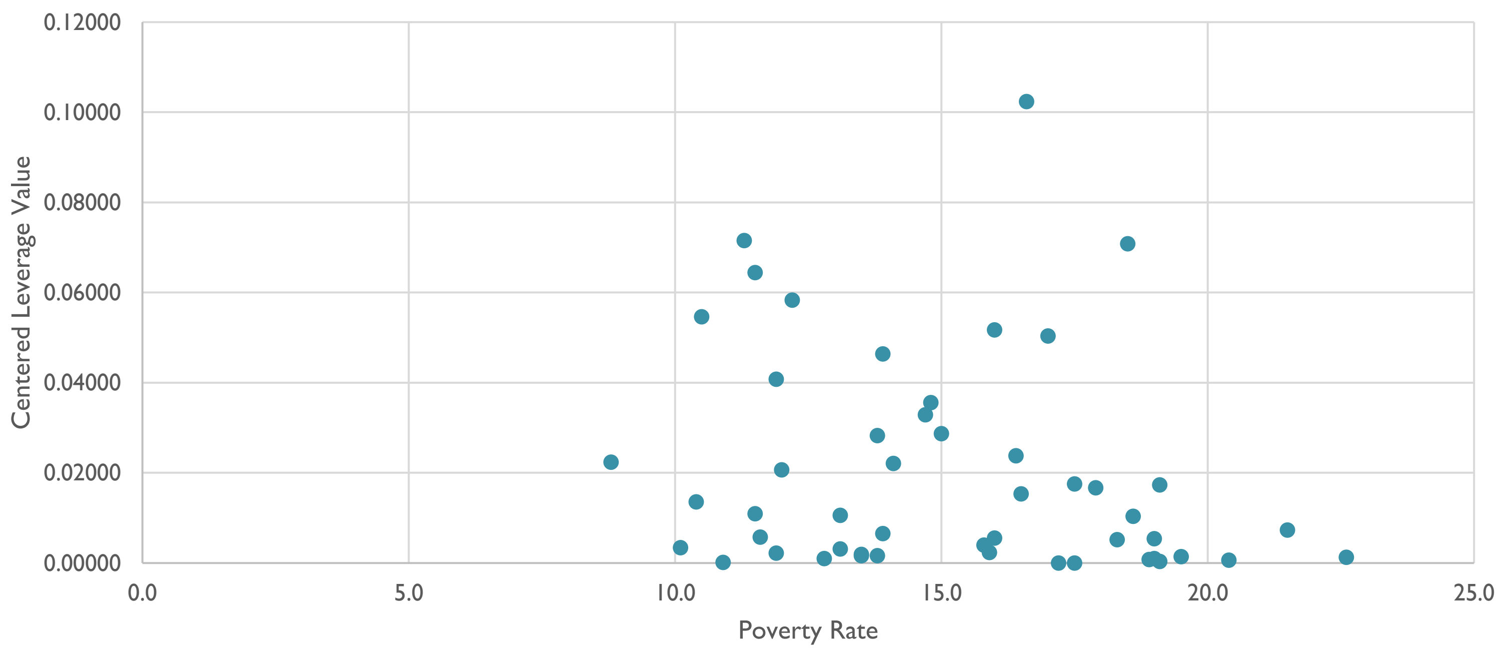
LNVIOLENT on DFFITS



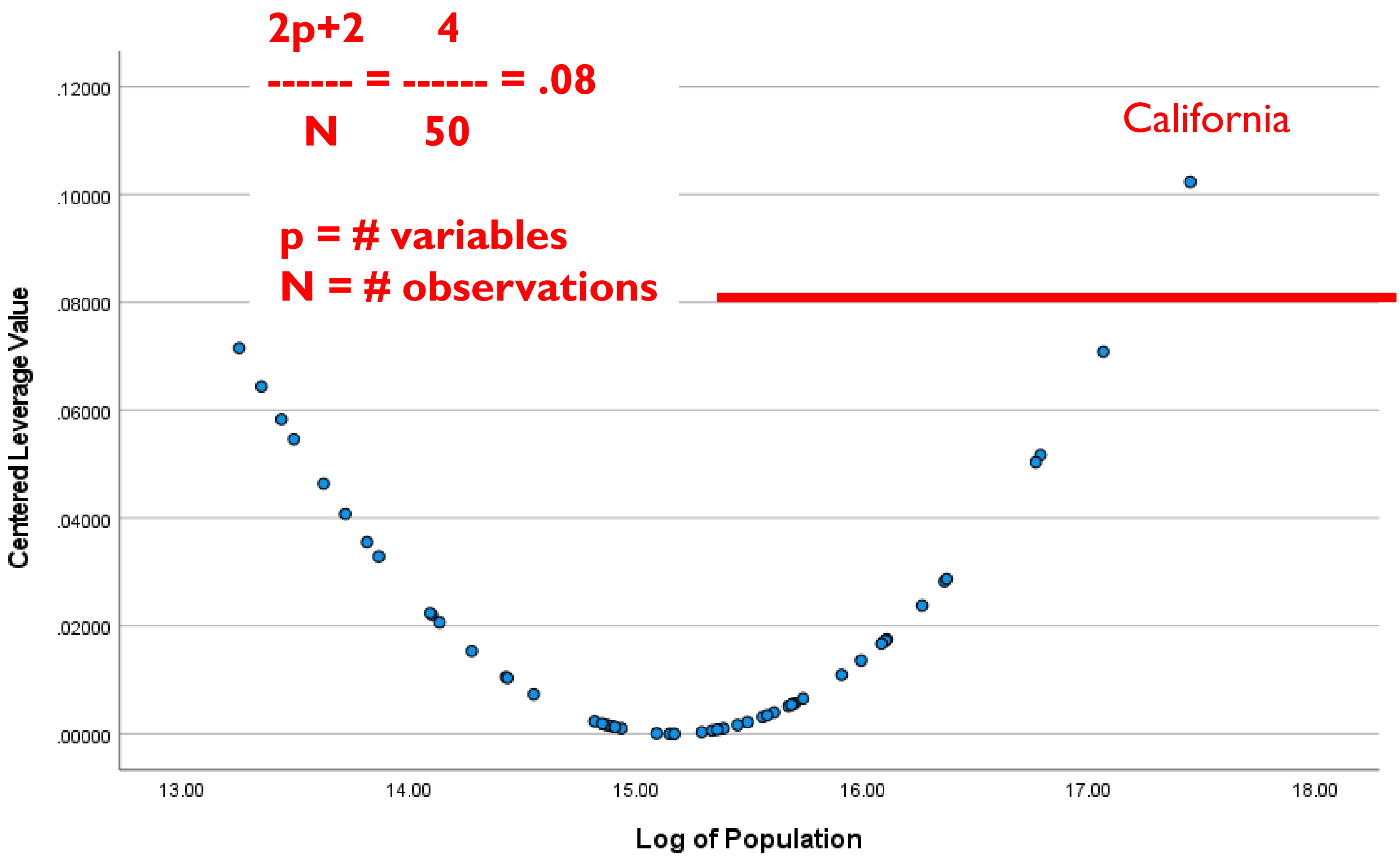
$2\sqrt{(p+1)/N} = 2\sqrt{2/50} = .4$

p = # variables
N = # observations

LNVIOLENT on Centered Leverage Statistic



LNVIOLENT on Centered Leverage Statistic



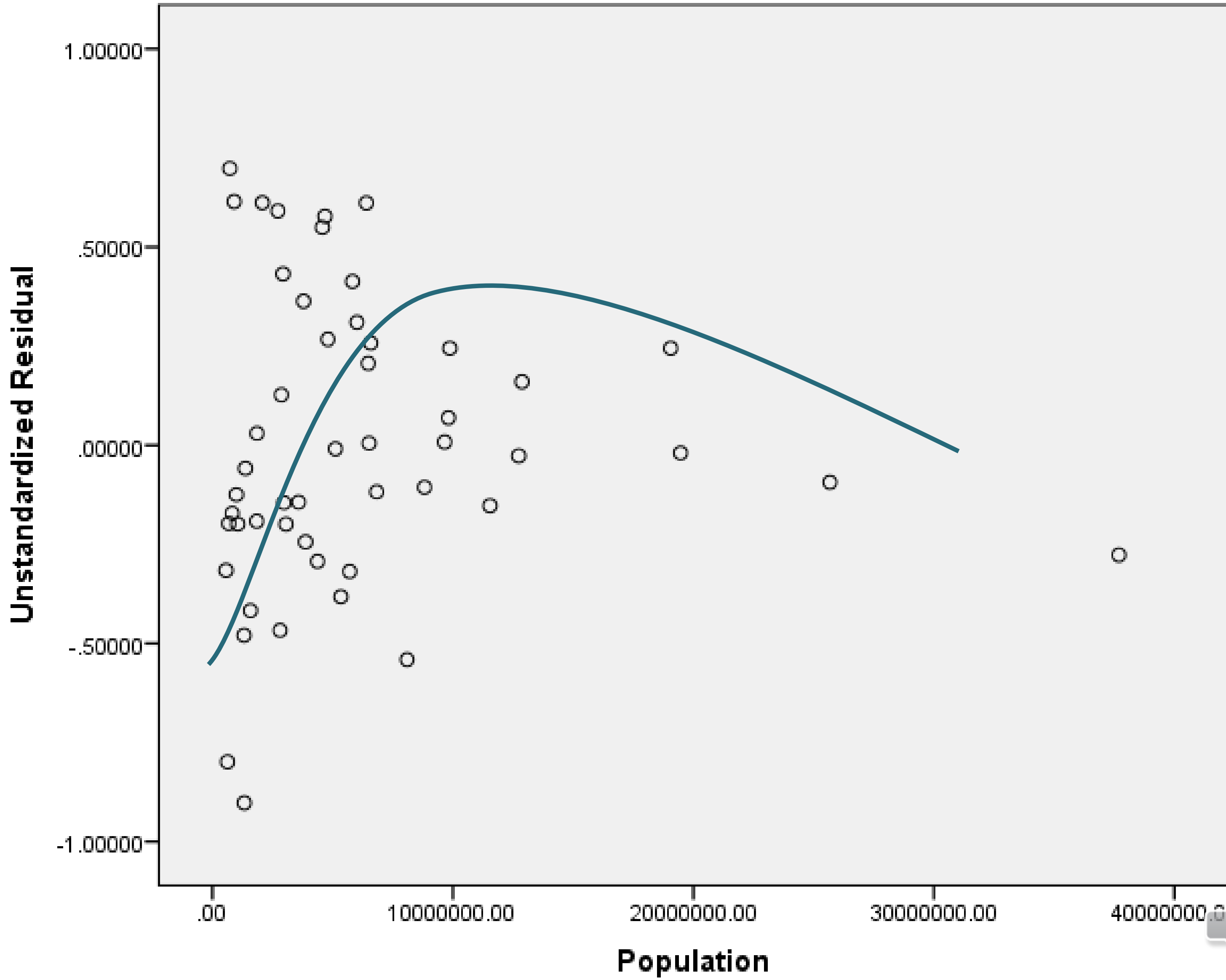
A Question

Which influence statistic is the best?

Possible Sources of Heteroskedasticity

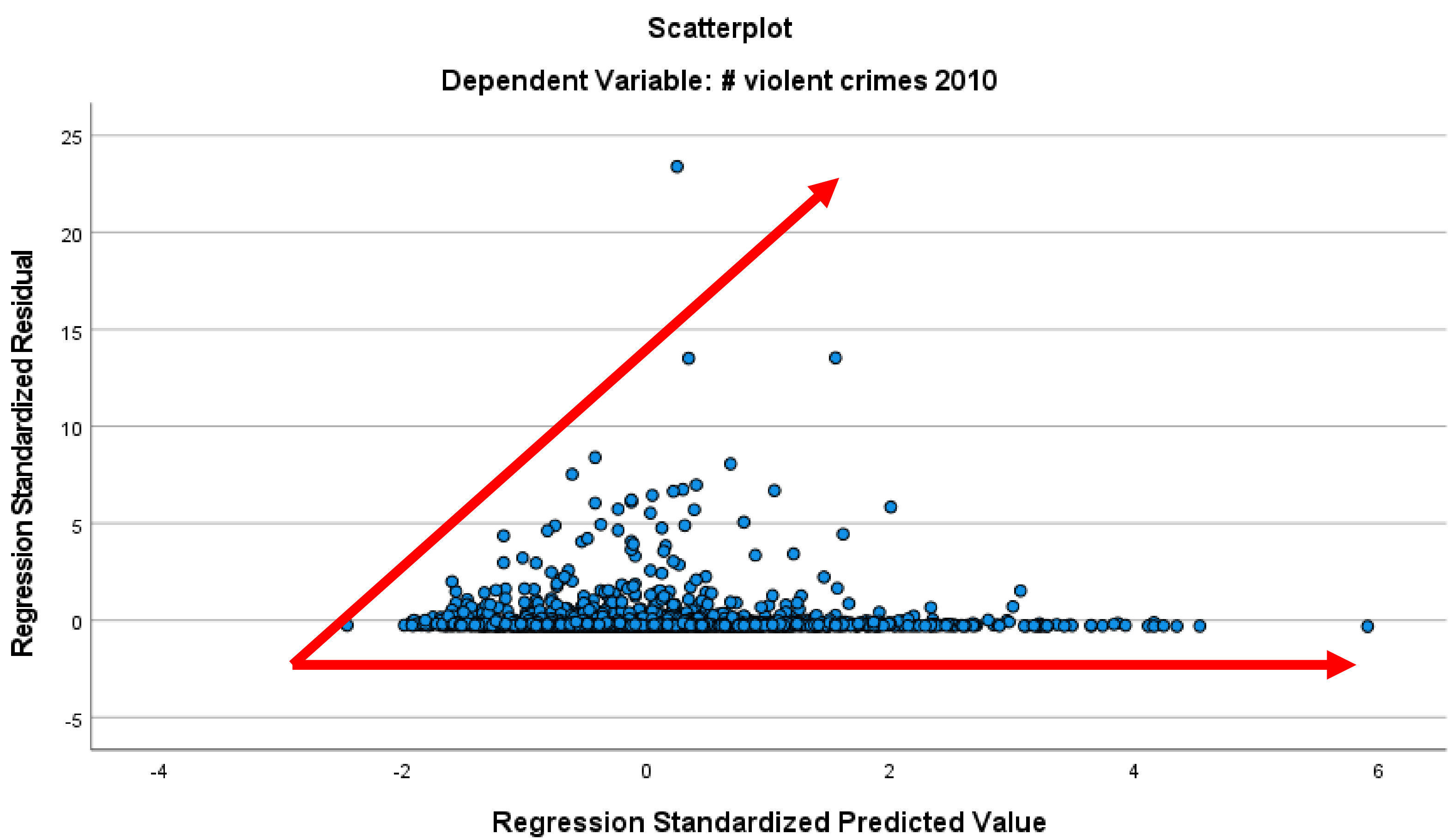
- Outliers
- Clusters of points
- Critical mass (population)
- Non-normality
- Counts

Residuals on Population



Possible Sources of Heteroskedasticity

- Outliers
- Clusters of points
- Critical mass (population)
- Non-normality
- Counts


Scatterplot
Dependent Variable: # violent crimes 2010
Regression Standardized Residual
25
20
15
10
5
0
-5
-4
-2
0
2
4
6
Regression Standardized Predicted Value

What to Do?

- Leave unchanged
- Transform or untransform a variable
- Change statistical procedures
- Remove outliers
- Add a weight

Huber Weights

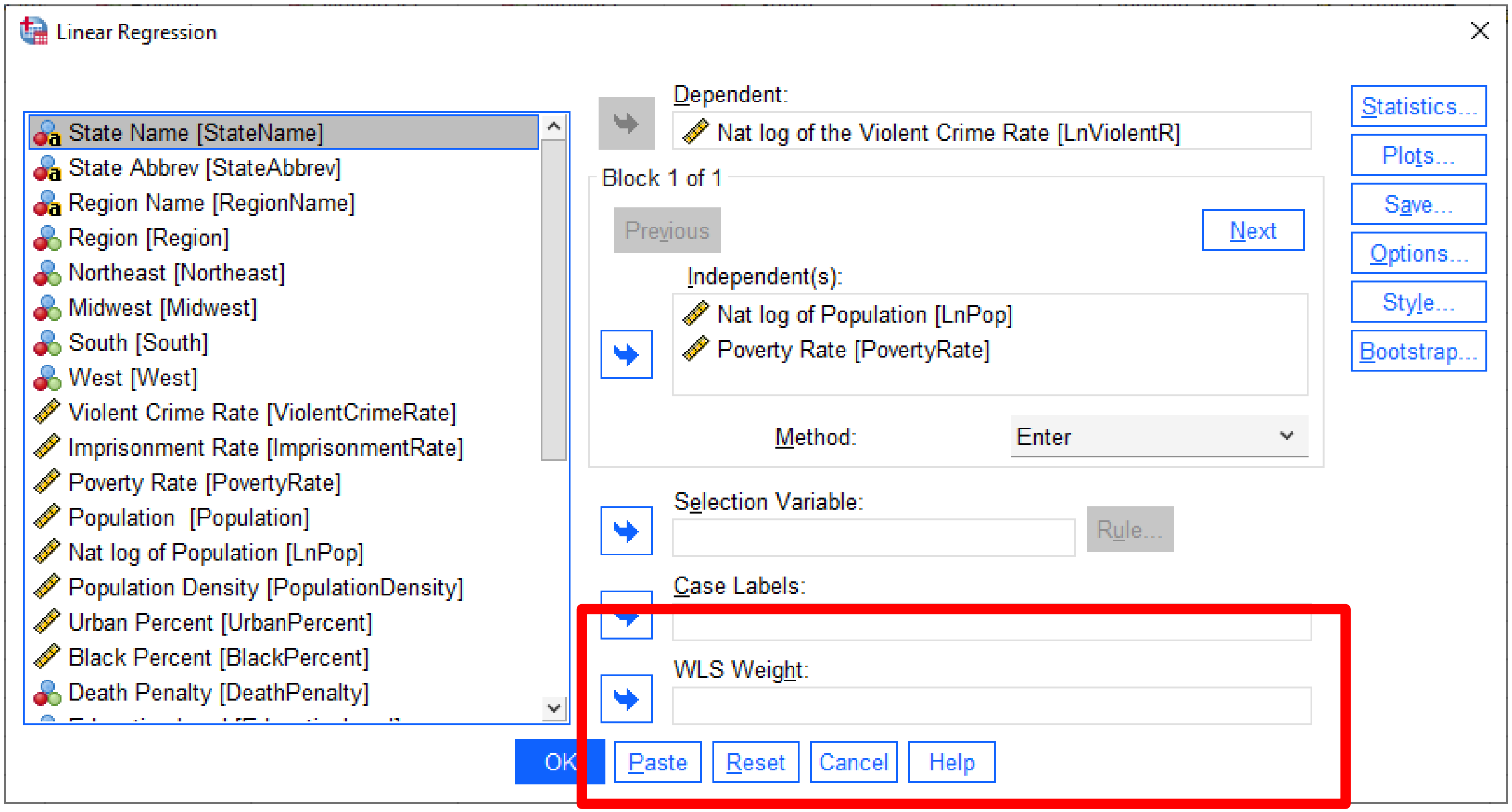
Let $k=1.345\sigma$

Determine the weight

If $|e_i| \leq k$: Weight = 1

If $|e_i| > k$: Weight = $k/|e_i|$

Weighted Least Squares Regression



Remember: Sometimes statistics is more like art than science.