




REGRESSION DIAGNOSTICS

Richard Lee Rogers



The Problem: The Error Term

$$y = b_0 + b_1x + e$$




The Problem: The Error Term

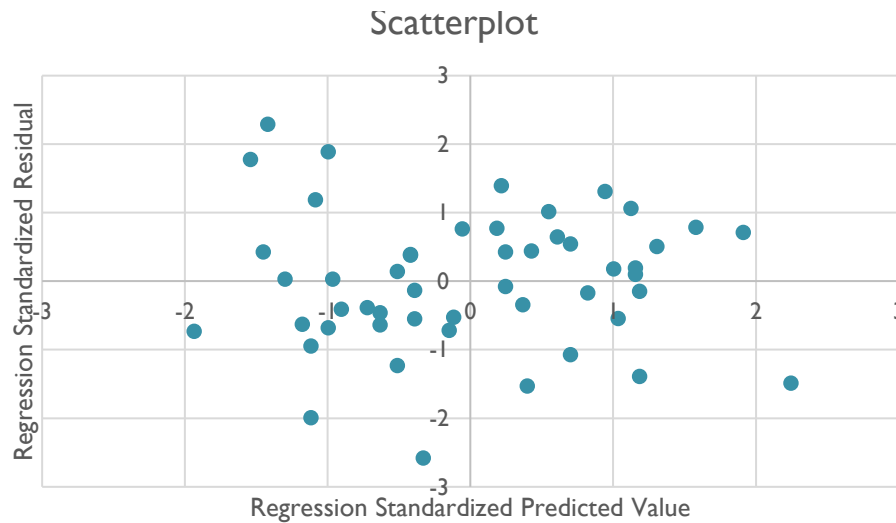
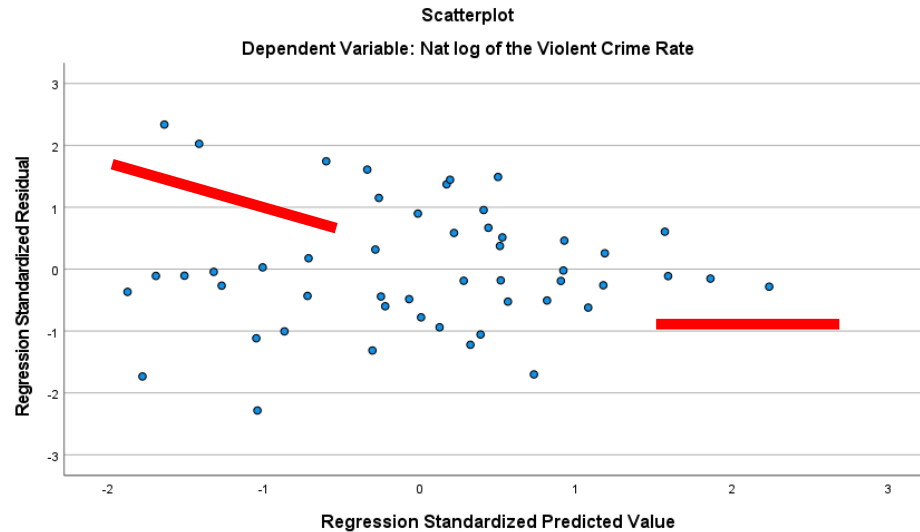
$$y = b_0 + b_1x + e$$



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



Linear Regression: Plots



DEPENDNT

*ZPRED

*ZRESID

*DRESID

*ADJPRED

*SRESID

*SDRESID

Scatter 1 of 1

Previous

Next

Y:

*ZRESID

X:

*ZPRED

Standardized Residual Plots

Histogram

Normal probability plot

Produce all partial plots

Continue

Cancel

Help



#1: DIRECT TO PLOTS



Linear Regression: Plots



DEPENDNT

*ZPRED

*ZRESID

*DRESID

*ADJPRED

*SRESID

*SDRESID

Scatter 1 of 1

Previous

Next

Y:

*ZRESID

X:

*ZPRED

Standardized Residual Plots

Histogram

Normal probability plot

Produce all partial plots

Continue

Cancel

Help



#2: SAVE VALUES TO DATASET

Linear Regression: Save

Predicted values

- Unstandardized
- Standardized
- Adjusted
- S.E. of mean predictions

Residuals

- Unstandardized
- Standardized
- Studentized
- Deleted
- Studentized deleted

Distances

- Mahalanobis
- Cook's
- Leverage values

Influence Statistics

- DfBetas
- Standardized DfBetas
- DfFits
- Standardized DfFits
- Covariance ratios

Prediction Intervals

- Mean Individual
- Confidence Interval: %

Coefficient statistics

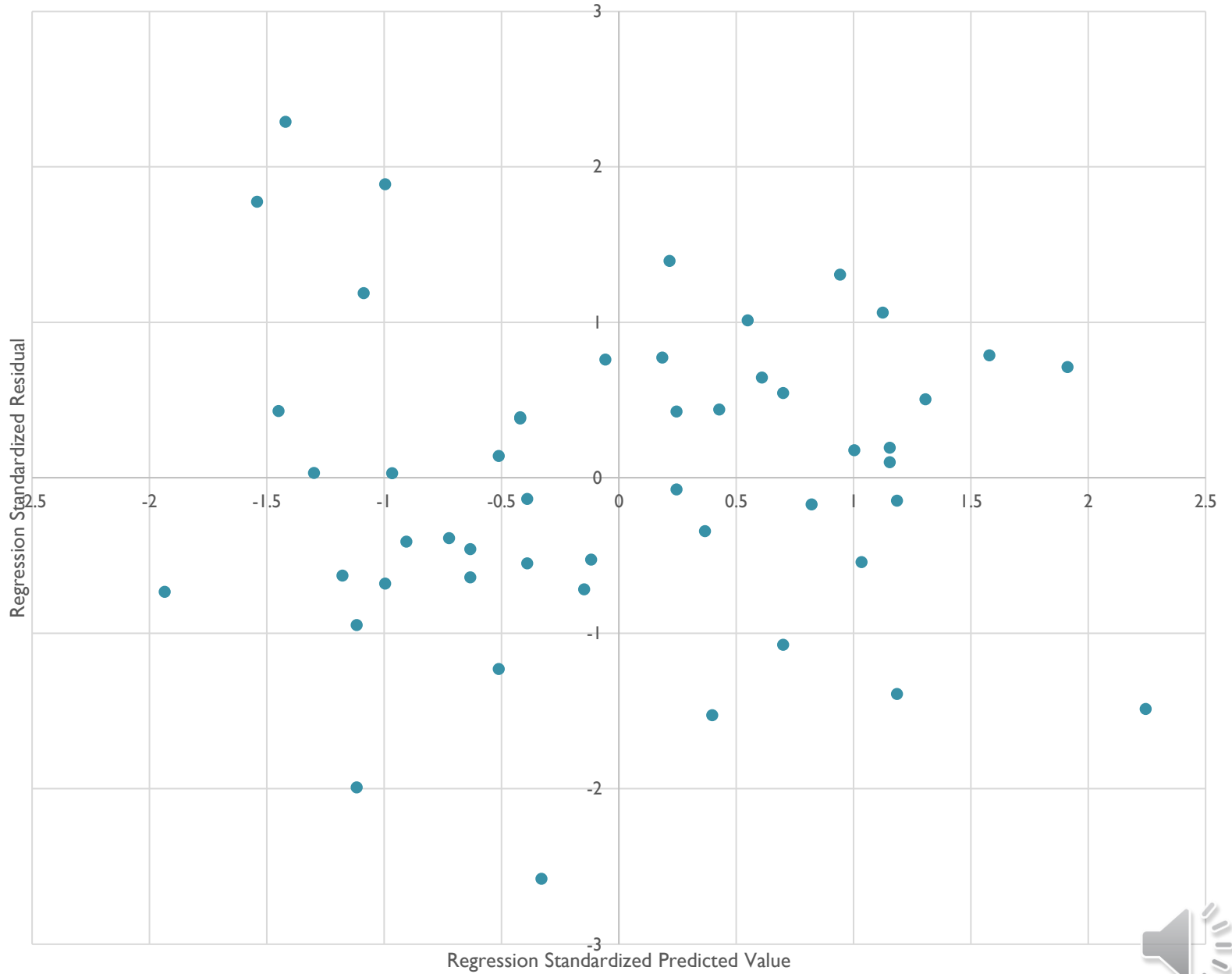
- Create coefficient statistics
- Create a new dataset
 - Dataset name:
- Write a new data file
 - File...

Export model information to XML file

-
- Include the covariance matrix



Scatterplot



Analyze>General Linear Model>Univariate>Options

Univariate: Options

Display

<input type="checkbox"/> Descriptive statistics	<input type="checkbox"/> Homogeneity tests
<input checked="" type="checkbox"/> Estimates of effect size	<input type="checkbox"/> Spread-vs.-level plots
<input checked="" type="checkbox"/> Observed power	<input type="checkbox"/> Residual plots
<input checked="" type="checkbox"/> Parameter estimates	<input type="checkbox"/> Lack-of-fit test
<input type="checkbox"/> Contrast coefficient matrix	<input type="checkbox"/> General estimable function(s)

Heteroskedasticity Tests

<input type="checkbox"/> Modified Breusch-Pagan test Model...	<input type="checkbox"/> F test Model...
<input checked="" type="checkbox"/> Breusch-Pagan test Model...	<input type="checkbox"/> White's test

Parameter estimates with robust standard errors

- HC0
- HC1
- HC2
- HC3
- HC4

Significance level: .05 Confidence intervals are 95.0 %

Continue Cancel Help



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

- Dependent variable: Nat log of the Violent Crime Rate
- Tests the null hypothesis that the variance of the errors does not depend on the values of the independent variables.
- 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), Inpoptot

b. Dependent Variable: Inviolent

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: Inviolent

b. Predictors: (Constant), Inpoptot

Coefficients^a

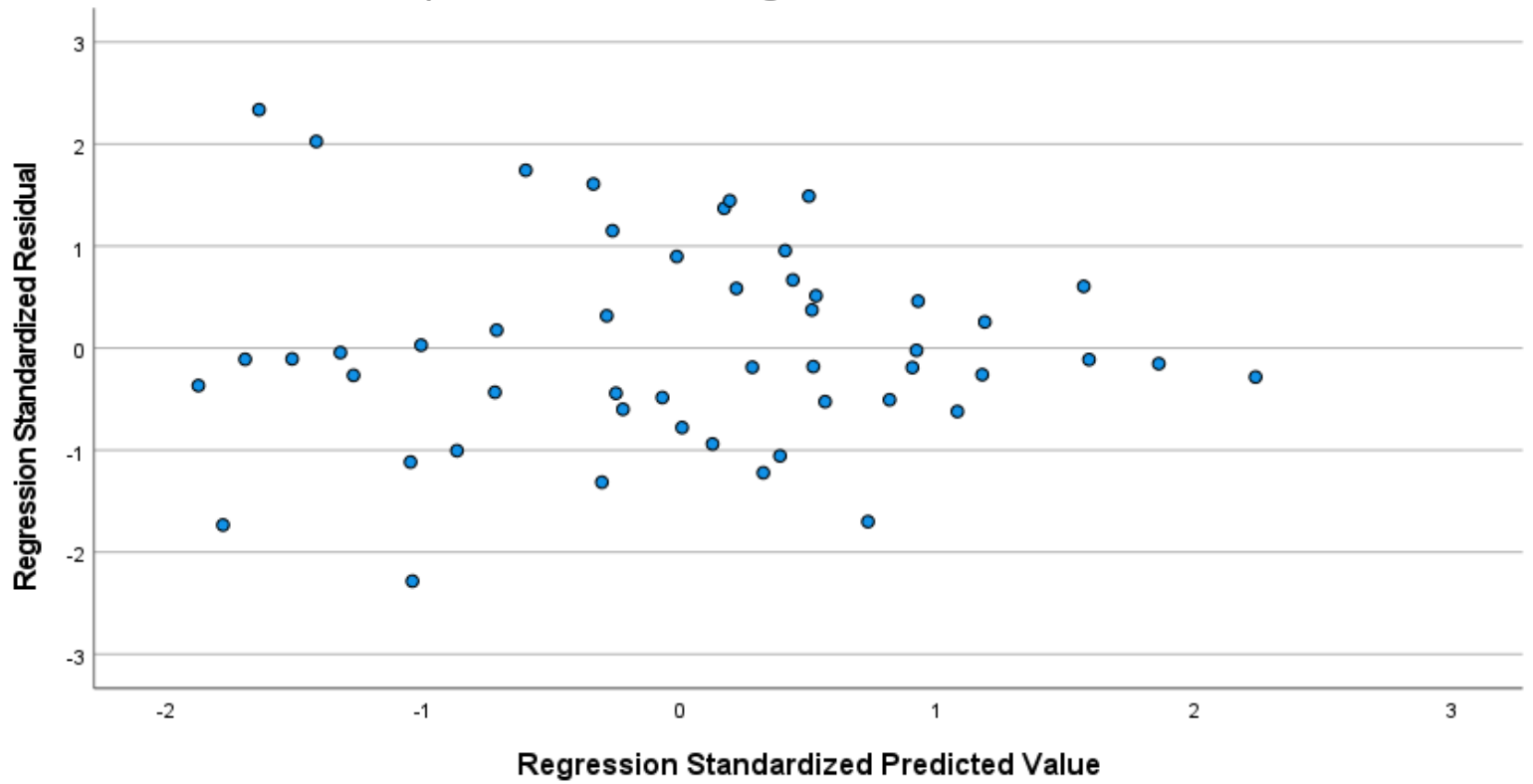
Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	3.636	.773		4.705	.000
	Inpoptot	.143	.051	.375	2.803	.007

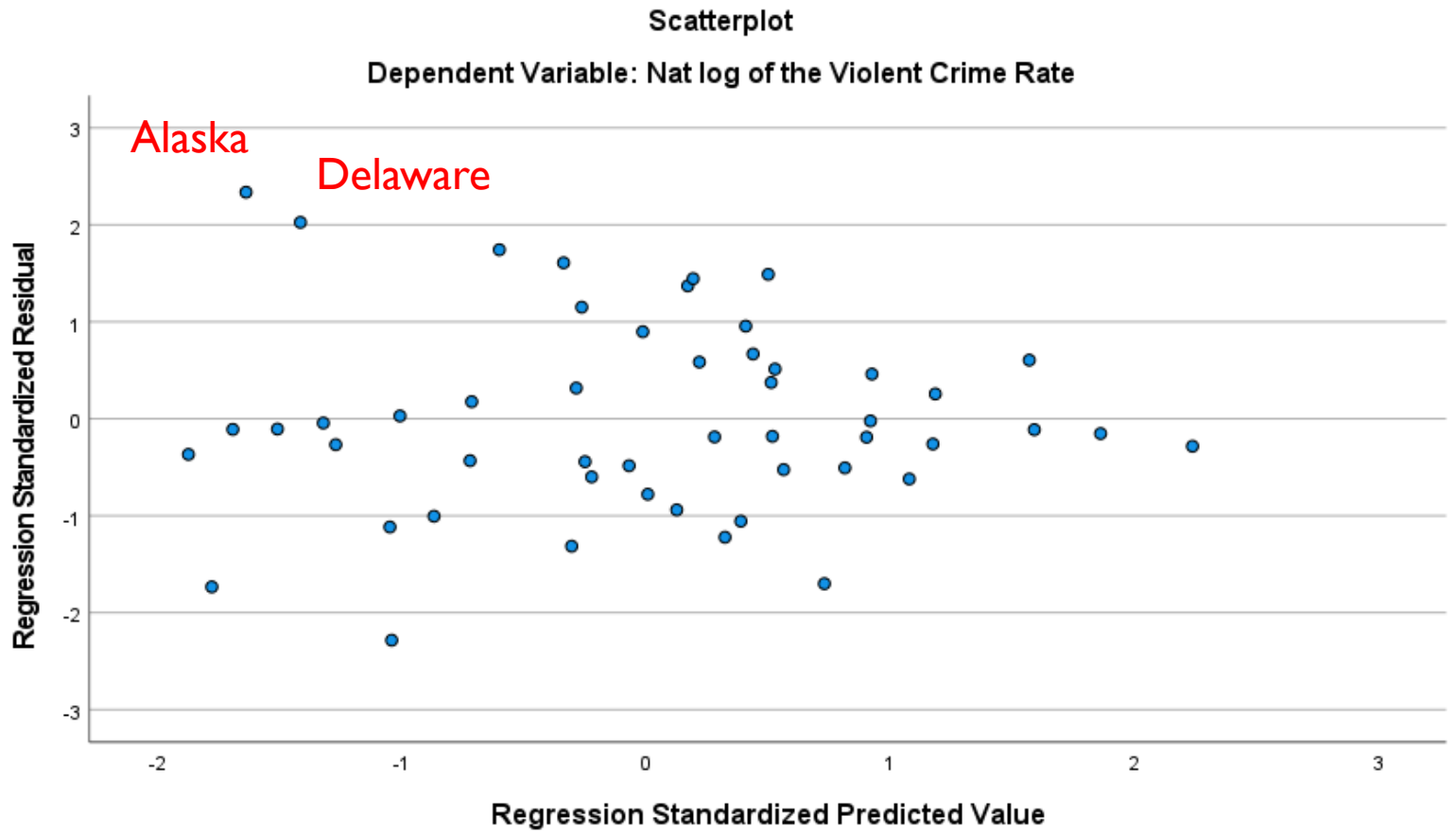
a. Dependent Variable: Inviolent



Scatterplot

Dependent Variable: Nat log of the Violent Crime Rate





Influence and Leverage Statistics

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



Analyze > Regression > Linear > Save

Linear Regression: Save

Unstandardized
 Standardized
 Adjusted
 S.E. of mean predictions

Unstandardized
 Standardized
 Studentized
 Deleted
 Studentized deleted

Mahalanobis
 Cook's
 Leverage values

DfBetas
 Standardized DfBetas
 DfFits
 Standardized DfFits
 Covariance ratios

Prediction Intervals
 Mean Individual

Confidence Interval: 95 %

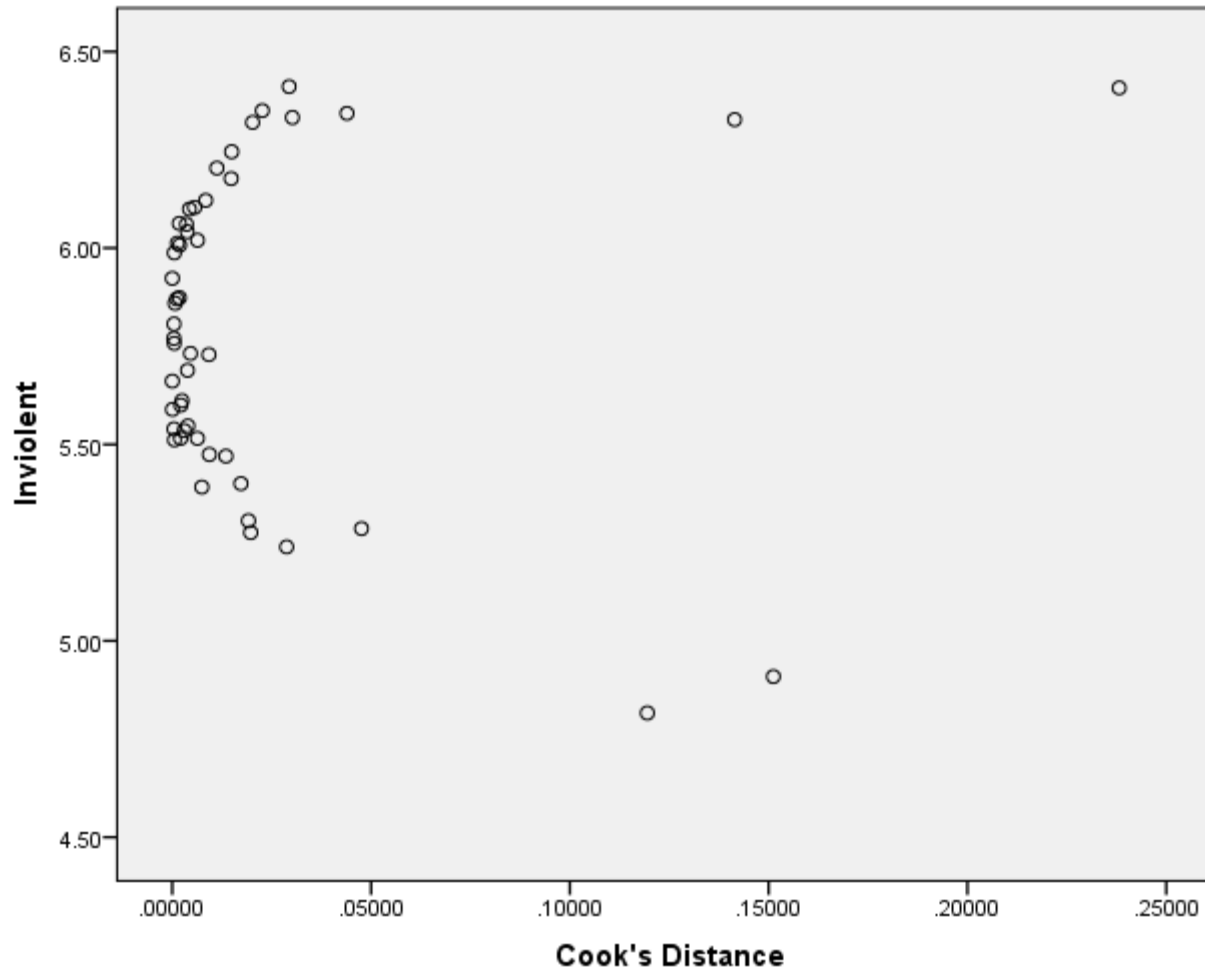
Coefficient statistics
 Create coefficient statistics
 Create a new dataset
Dataset name:
 Write a new data file
File...

Export model information to XML file
 Browse...
 Include the covariance matrix

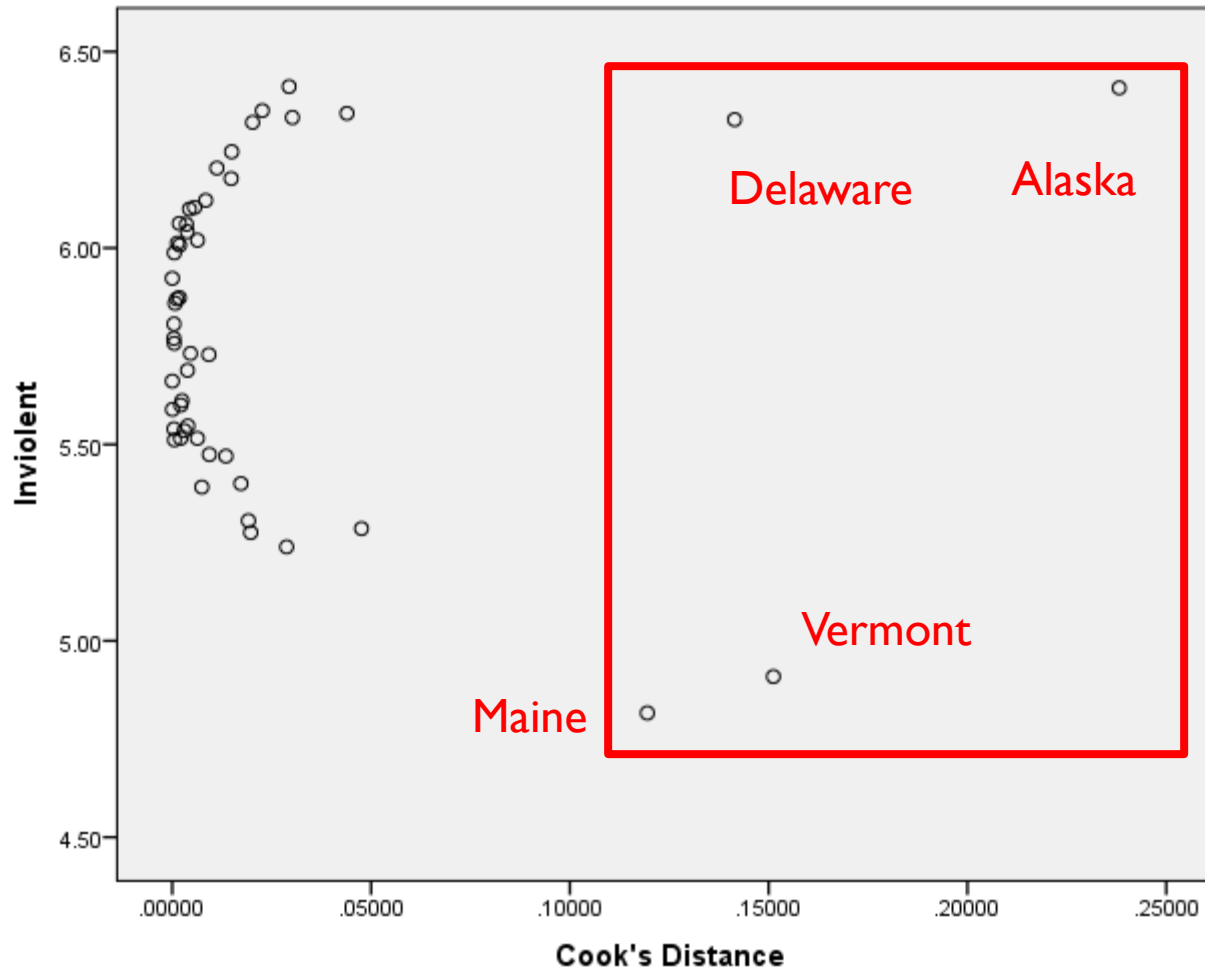
Continue Cancel Help



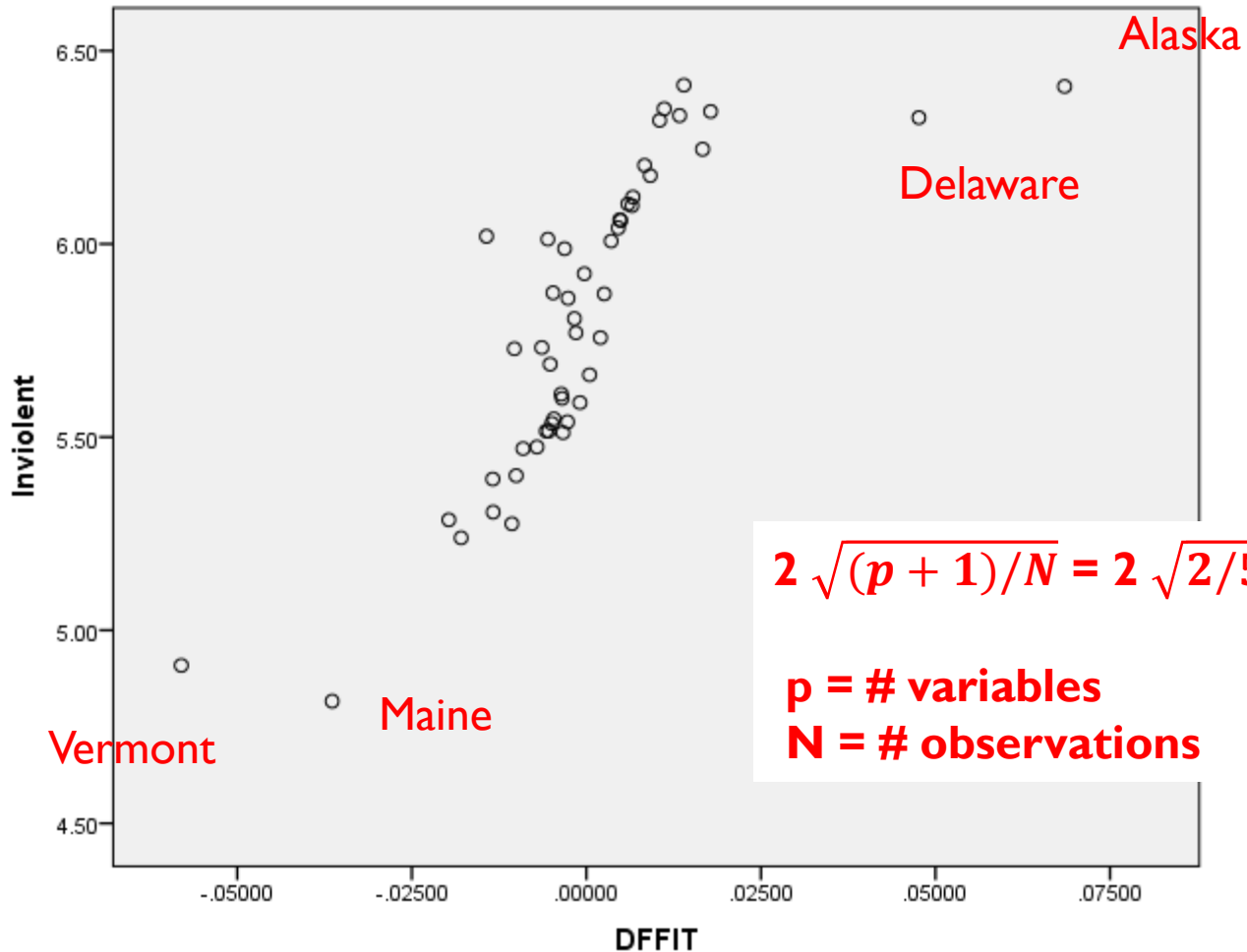
LNVIOLENT on Cook's D



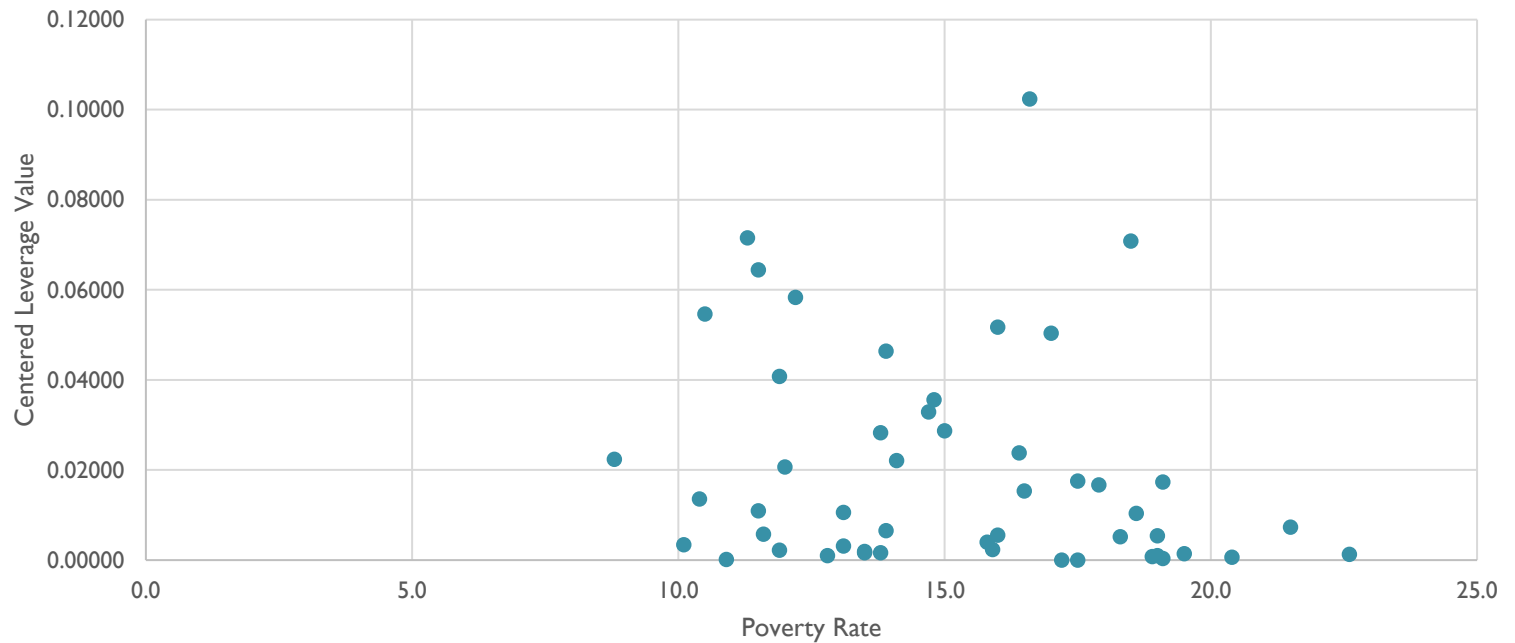
LNVIOLENT on Cook's D



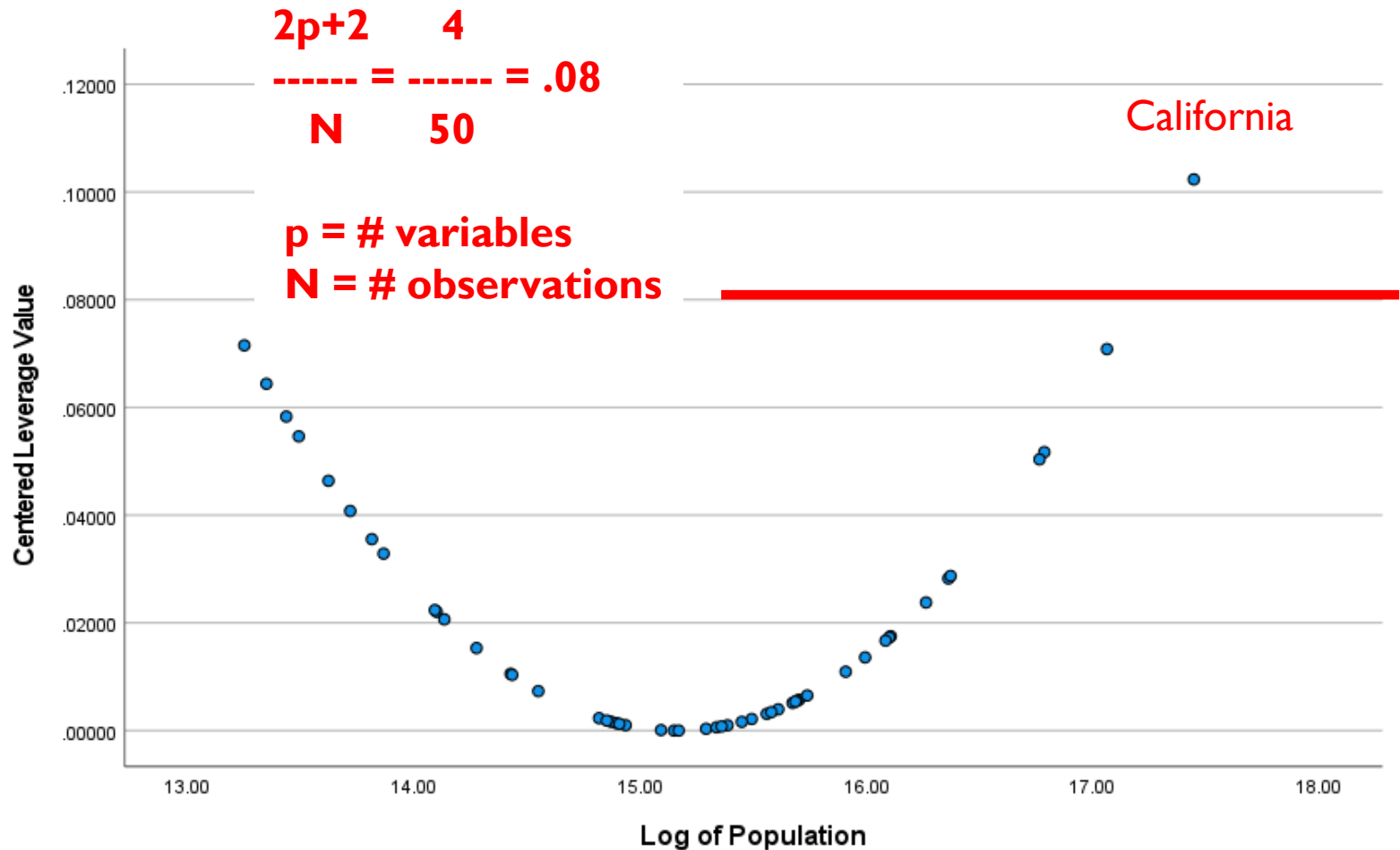
LNVIOLENT on DFFITS



LNVIOLENT on Centered Leverage Statistic



LNVIOLENT on Centered Leverage Statistic



A Question

Which influence statistic is the best?

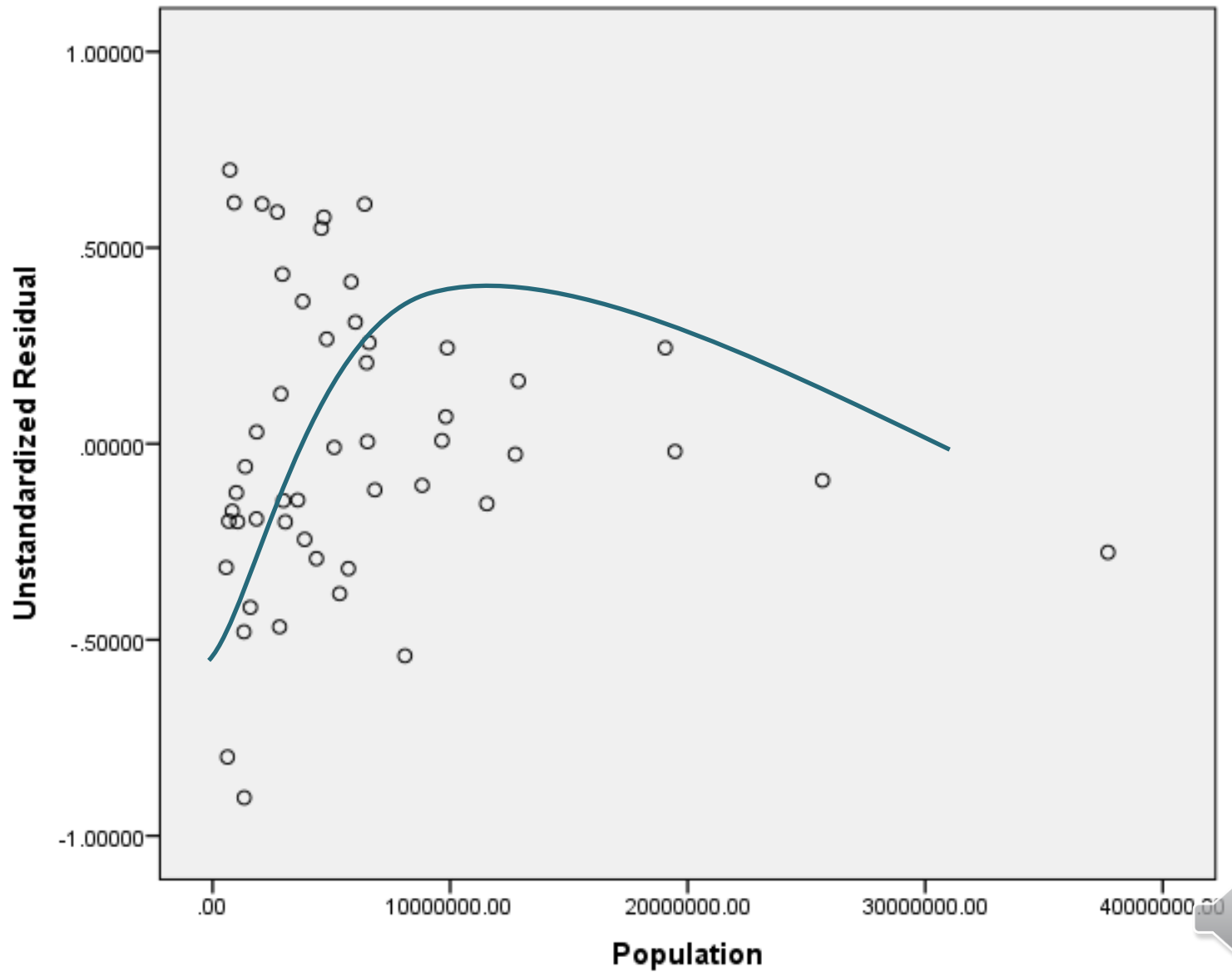


Possible Sources of Heteroskedasticity

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



Residuals on Population



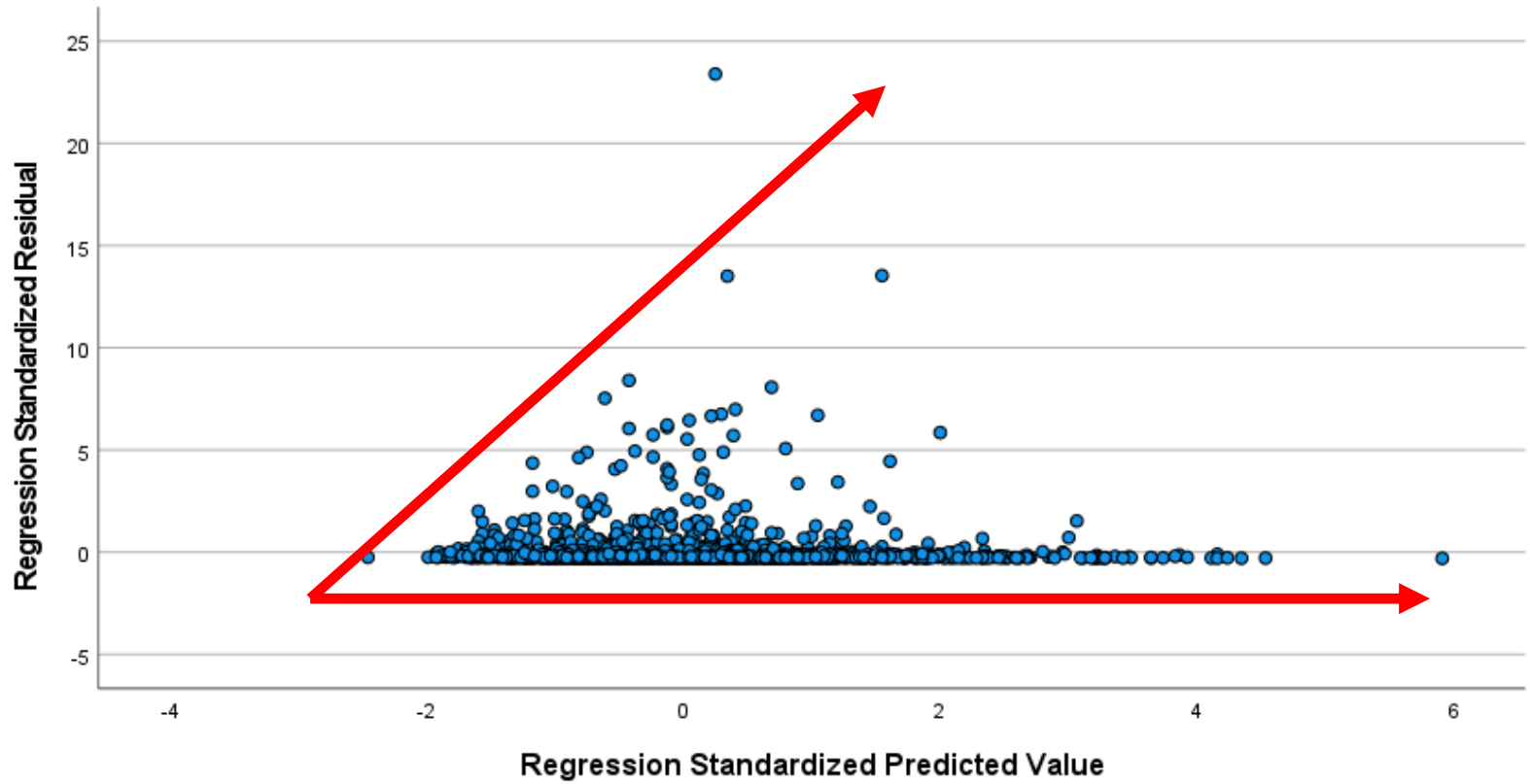
Possible Sources of Heteroskedasticity

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



Scatterplot

Dependent Variable: # violent crimes 2010



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

Linear Regression

State Name [StateName]
State Abbrev [StateAbbrev]
Region Name [RegionName]
Region [Region]
Northeast [Northeast]
Midwest [Midwest]
South [South]
West [West]
Violent Crime Rate [ViolentCrimeRate]
Imprisonment Rate [ImprisonmentRate]
Poverty Rate [PovertyRate]
Population [Population]
Nat log of Population [LnPop]
Population Density [PopulationDensity]
Urban Percent [UrbanPercent]
Black Percent [BlackPercent]
Death Penalty [DeathPenalty]

Dependent:
Nat log of the Violent Crime Rate [LnViolentR]

Block 1 of 1

Previous Next

Independent(s):
Nat log of Population [LnPop]
Poverty Rate [PovertyRate]

Method: Enter

Selection Variable: Rule...

Case Labels:

WLS Weight:

OK Paste Reset Cancel Help

Statistics...
Plots...
Save...
Options...
Style...
Bootstrap...



Remember: Sometimes statistics is more like art than science.

