



**Please!**  
Turn **OFF** cell phones  
and paging devices



# Regression Analysis for Spatial Data

Lauren M. Scott, PhD and Mark V. Janikas, PhD

Geoprocessing & Spatial Statistics, ESRI Redlands, CA

Presented by Owen Evans, Solutions Engineer, ESRI-WDC

at the

DOT Workshop for Transportation Forecasters

September 2009

# Objectives

## GOALS OF THIS WORKSHOP

- Introduce basic objectives, terminology, and analysis strategies.
- Demonstrate the utility of OLS and GWR regression analysis.
- Outline the challenges of regression for spatial data.
- Present regression analysis diagnostics.
- Provide strategies to help navigate and interpret regression results.
- Highlight resources for learning more about regression analysis.

## DEMO

- Why are people dying young in South Dakota?

# Regression Analysis

## Basics

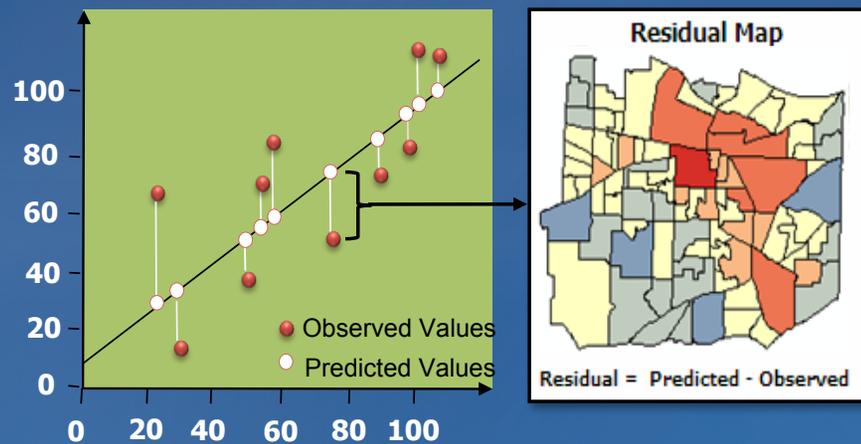
# Regression analysis

- **Regression analysis allows you to:**
  - **Model, examine, and explore spatial relationships**
  - **Better understand the factors behind observed spatial patterns**
  - **Predict outcomes based on that understanding**

# Regression analysis

- Regression analysis allows you to:
  - Model, examine, and explore spatial relationships
  - Better understand the factors behind observed spatial patterns
  - Predict outcomes based on that understanding

## Ordinary Least Squares (OLS)

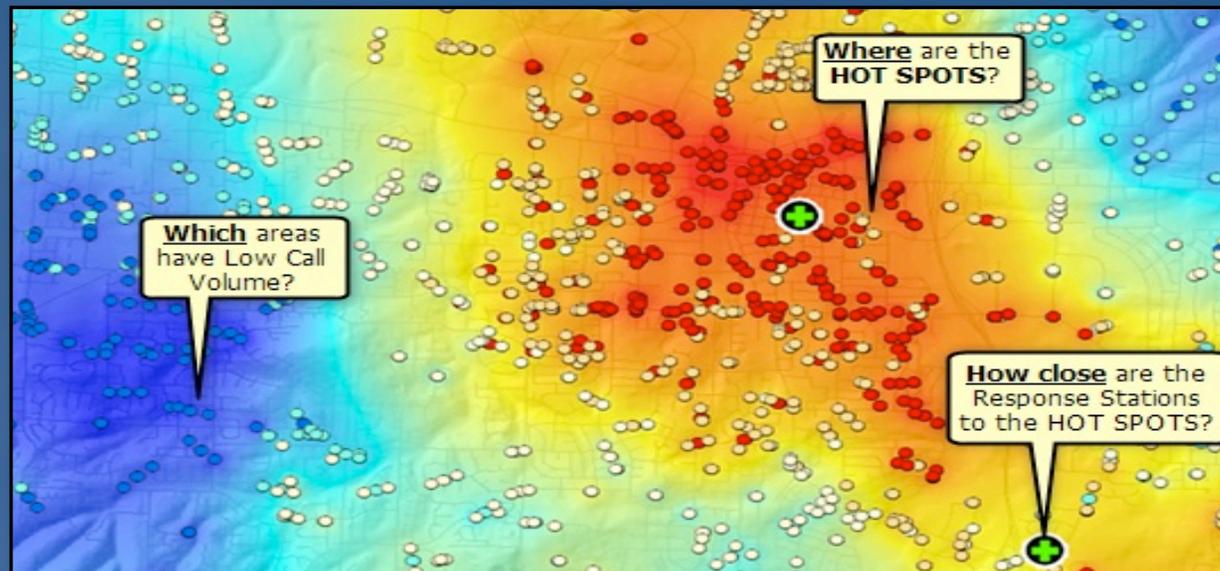


## Geographically Weighted Regression (GWR)



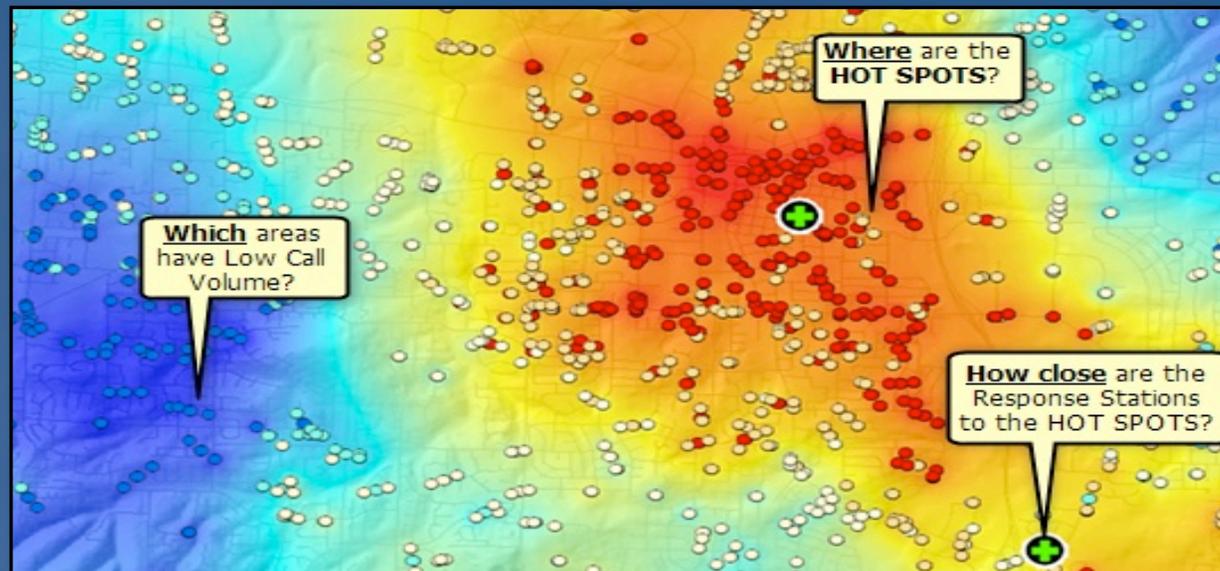
# Finding Where...

- **Pattern analysis (without regression):**
  - Are there places where people persistently die young?
  - Where are test scores consistently high?
  - Where are 911 emergency call hot spots?



# Finding Where vs. Understanding Why

- **Pattern analysis (without regression):**
  - Are there places where people persistently die young?
  - Where are test scores consistently high?
  - Where are 911 emergency call hot spots?



- **Regression analysis:**
  - Why are people persistently dying young?
  - *What factors* contribute to consistently high test scores?
  - *Which variables* effectively predict 911 emergency call volumes?

# Why use regression?

- **Understand key factors**
  - What are the most important habitat characteristics for an endangered bird?



# Why use regression?

- **Understand key factors**
  - What are the most important habitat characteristics for an endangered bird?
- **Predict unknown values**
  - How much rainfall will occur in a given location?



# Why use regression?

- **Understand key factors**
  - What are the most important habitat characteristics for an endangered bird?
- **Predict unknown values**
  - How much rainfall will occur in a given location?
- **Test hypotheses**
  - “Broken Window” Theory: Is there a positive relationship between vandalism and residential burglary?



# Applications

- **Education**
  - Why are literacy rates so low in particular regions?
- **Natural resource management**
  - What are the key variables that explain high forest fire frequency?
- **Ecology**
  - Which environments should be protected, to encourage reintroduction of an endangered species?
- **Transportation**
  - What demographic characteristics contribute to high rates of public transportation usage?
- **Many more...**
  - Business, crime prevention, epidemiology, finances, public safety, public health



# Regression analysis terms and concepts

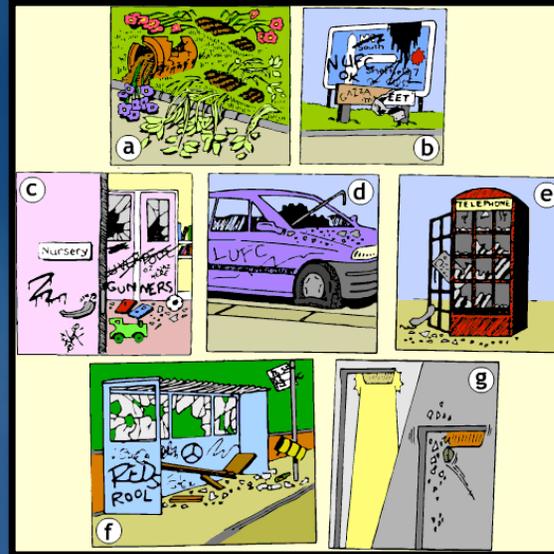


Residential Burglary

# Regression analysis terms and concepts



Residential Burglary



Vandalism



Income



Number of households

# Regression analysis terms and concepts



$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_n X_n + \varepsilon$$

# Regression analysis terms and concepts

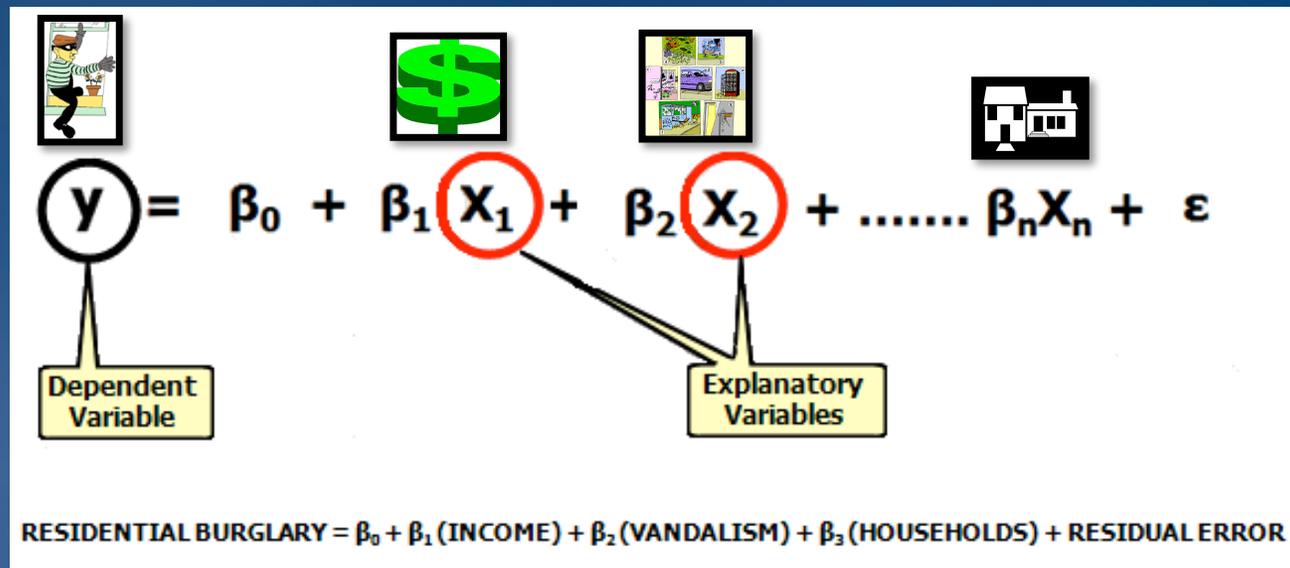
$y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_n X_n + \epsilon$

Dependent Variable

RESIDENTIAL BURGLARY =  $\beta_0 + \beta_1$  (INCOME) +  $\beta_2$  (VANDALISM) +  $\beta_3$  (HOUSEHOLDS) + RESIDUAL ERROR

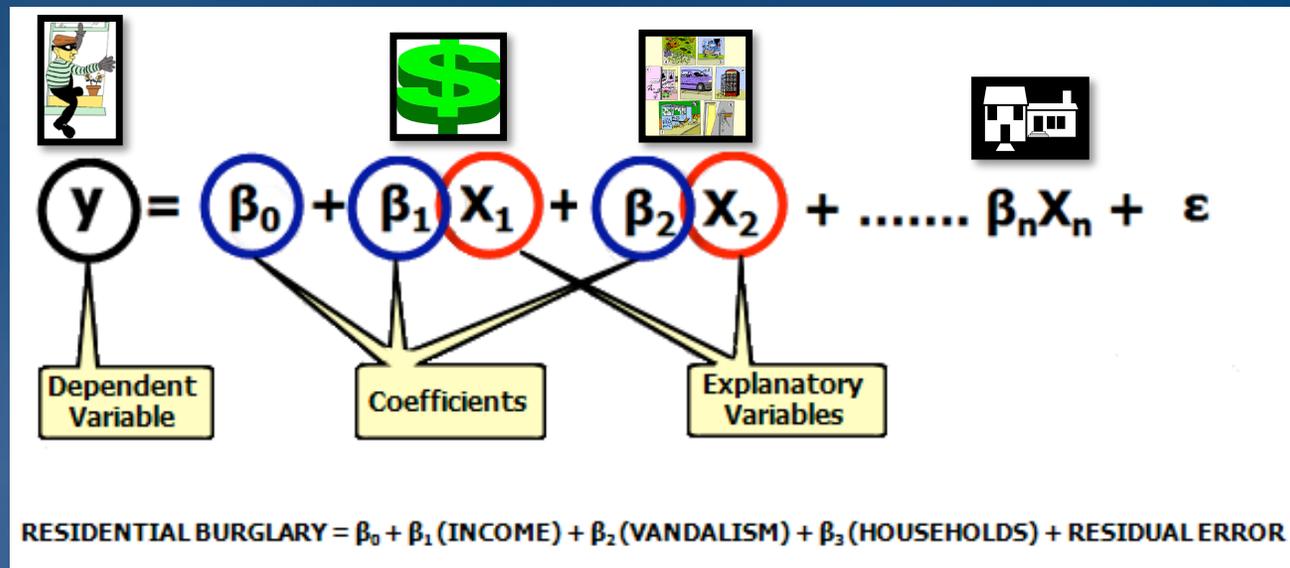
- *Dependent variable* (Y): What you are trying to model or predict (e.g., residential burglary).

# Regression analysis terms and concepts



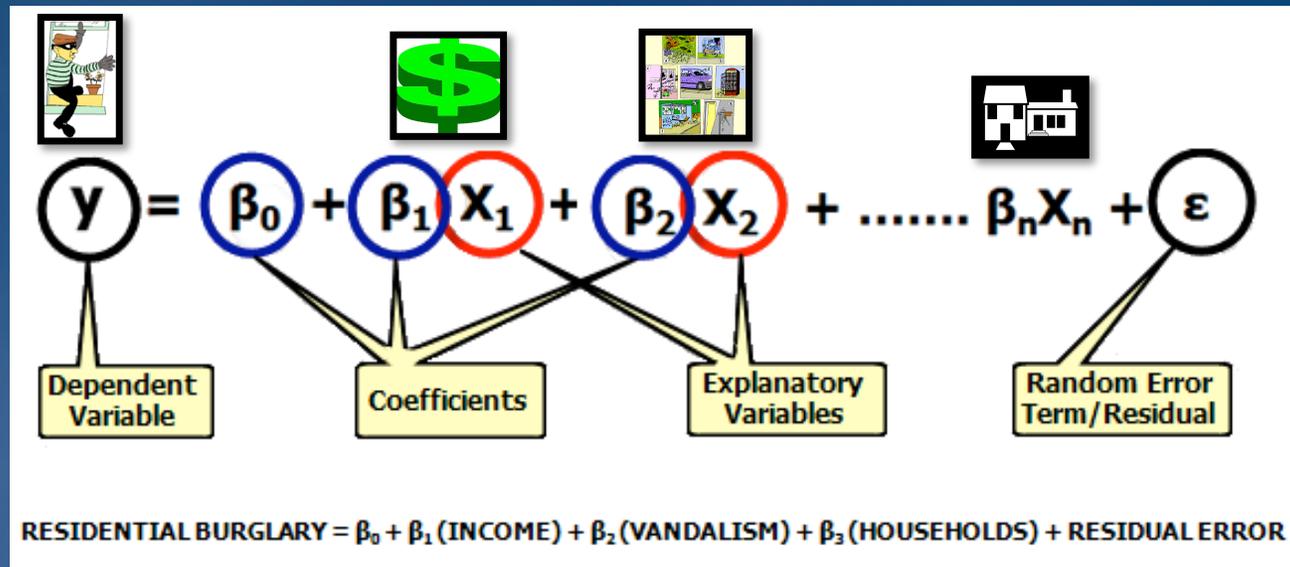
- *Dependent variable (Y)*: What you are trying to model or predict (e.g., residential burglary).
- *Explanatory variables (X)*: Variables you believe cause or explain the dependent variable (e.g., income, vandalism, number of households).

# Regression analysis terms and concepts



- *Dependent variable (Y)*: What you are trying to model or predict (e.g., residential burglary).
- *Explanatory variables (X)*: Variables you believe cause or explain the dependent variable (e.g., income, vandalism, number of households).
- *Coefficients ( $\beta$ )*: Values, computed by the regression tool, reflecting the relationship between explanatory variables and the dependent variable.

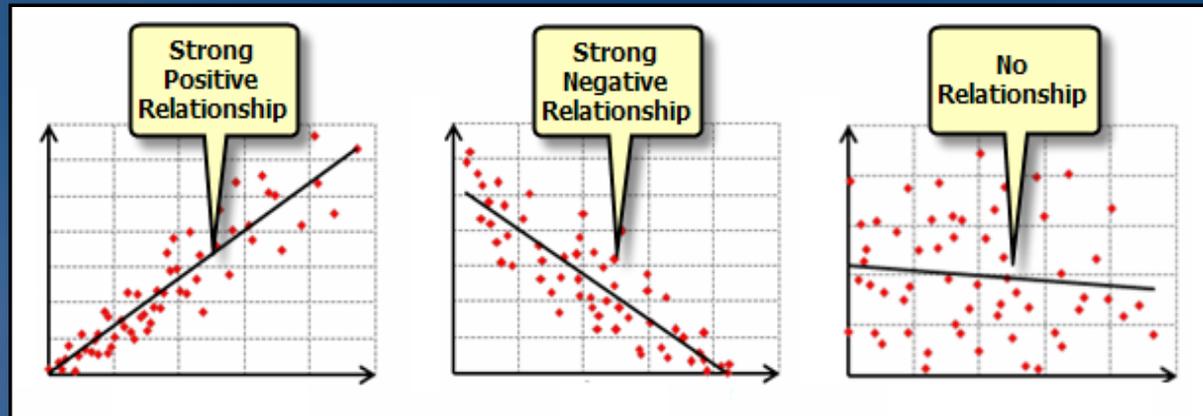
# Regression analysis terms and concepts



- *Dependent variable (Y)*: What you are trying to model or predict (e.g., residential burglary).
- *Explanatory variables (X)*: Variables you believe cause or explain the dependent variable (e.g., income, vandalism, number of households).
- *Coefficients ( $\beta$ )*: Values, computed by the regression tool, reflecting the relationship between explanatory variables and the dependent variable.
- *Residuals ( $\epsilon$ )*: The portion of the dependent variable that isn't explained by the model; the model under- and over-predictions.

# Regression model coefficients

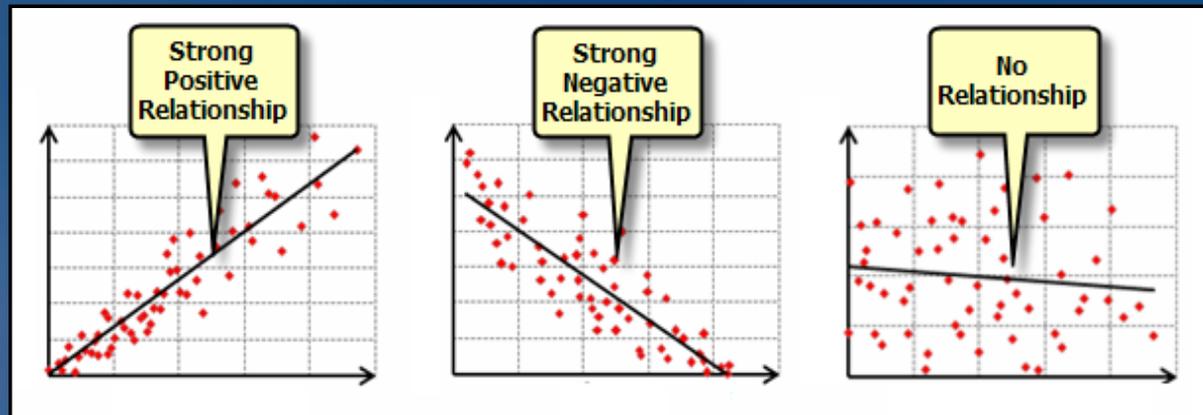
- Coefficient sign (+/-) and magnitude reflect each explanatory variable's relationship to the dependent variable



Summary of OLS Results									
Variable	Coefficient	StdError	t-Statistic	Probability	Robust_SE	Robust_t	Robust_Pr	VIF	[1]
Intercept	1.625506	0.132184	12.297312	0.000000*	0.141894	11.455814	0.000000*	-----	
INCOME	-0.000030	0.000003	-9.714560	0.000000*	0.000003	-10.200131	0.000000*	1.312936	
VANDALISM	0.133712	0.007818	17.103675	0.000000*	0.013133	10.181417	0.000000*	1.306447	
HOUSEHOLD	0.012425	0.001213	10.245426	0.000000*	0.002155	5.764924	0.000000*	1.315309	
LOWERCITY	0.136569	0.122861	1.111567	0.266500	0.112888	1.209771	0.226562	1.318318	

# Regression model coefficients

- Coefficient sign (+/-) and magnitude reflect each explanatory variable's relationship to the dependent variable



**Intercept**            1.625506  
**INCOME**                -0.000030  
**VANDALISM**            0.133712  
**HOUSEHOLDS**          0.012425  
**LOWER CITY**            0.136569

The asterisk \* indicates the explanatory variable is statistically significant

Summary of OLS Results								
Variable	Coefficient	StdError	t-Statistic	Probability	Robust_SE	Robust_t	Robust_Pr	VIF [1]
Intercept	1.625506	0.132184	12.297312	0.000000*	0.141894	11.455814	0.000000*	-----
INCOME	-0.000030	0.000003	-9.714560	0.000000*	0.000003	-10.200131	0.000000*	1.312936
VANDALISM	0.133712	0.007818	17.103675	0.000000*	0.013133	10.181417	0.000000*	1.306447
HOUSEHOLD	0.012425	0.001213	10.245426	0.000000*	0.002155	5.764924	0.000000*	1.315309
LOWERCITY	0.136569	0.122861	1.111567	0.266500	0.112888	1.209771	0.226562	1.318318

# **Building a Regression Model:**

## **OLS**

### **Ordinary Least Squares Regression**

# Building a global OLS regression model

1. Choose your dependent variable (Y).

# Building a global OLS regression model

1. Choose your dependent variable (Y).
2. Identify potential explanatory variables (X).

# Building a global OLS regression model

1. Choose your dependent variable ( $Y$ ).
2. Identify potential explanatory variables ( $X$ ).
3. Explore those explanatory variables.

# Building a global OLS regression model

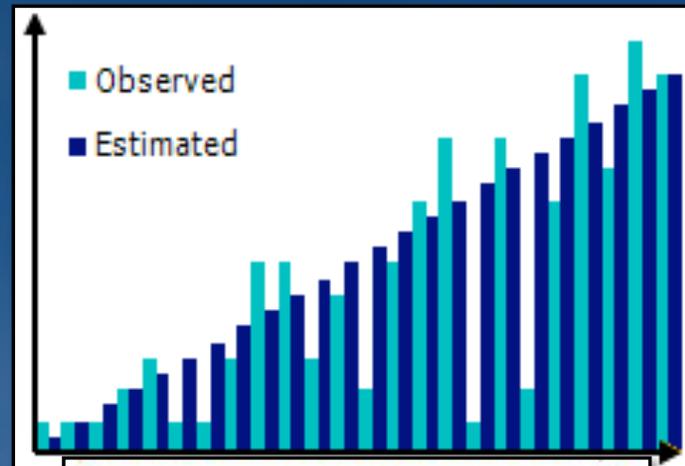
1. Choose your dependent variable (Y).
2. Identify potential explanatory variables (X).
3. Explore those explanatory variables.
4. Run OLS regression with different combinations of explanatory variables, until you find a properly specified model.

# Building a global OLS regression model

1. Choose your dependent variable (Y).
2. Identify potential explanatory variables (X).
3. Explore those explanatory variables.
4. Run OLS regression with different combinations of explanatory variables, until you find a properly specified model.

**How do we know we have a properly specified model?**

# Regression analysis output



The R squared value is a measure of how well the estimated values explain the variance in the observed values

## Summary of OLS Results

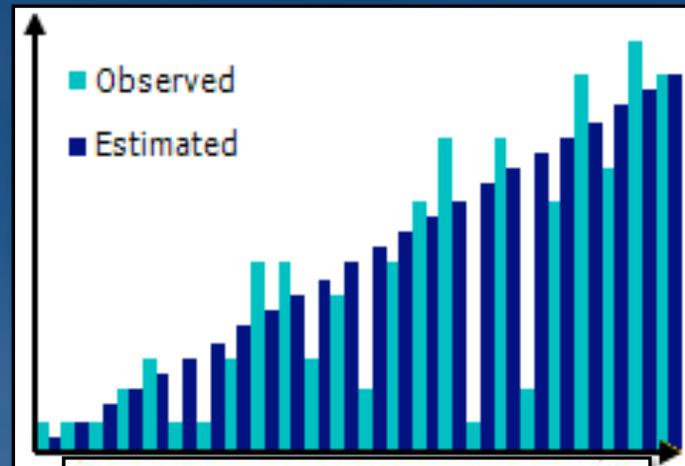
Variable	Coefficient	StdError	t-Statistic	Probability	Robust_SE	Robust_t	Robust_Pr	VIF [1]
Intercept	1.588164	0.127854	12.421710	0.000000*	0.138934	11.431081	0.000000*	-----
INCOME	-0.000029	0.000003						
VANDALISM	0.133469	0.007815						
HOUSEHOLD	0.012620	0.001200						

**Adjusted R-Squared [2]: 0.37407**

## OLS Diagnostics

Number of Observations:	1482	Number of Variables:	4
Degrees of Freedom:	1478	Akaike's Information Criterion (AIC) [2]:	5813.121
Multiple R-Squared [2]:	0.375339	<b>Adjusted R-Squared [2]:</b>	<b>0.374071</b>
Joint F-Statistic [3]:	296.027584	Prob(>F), (3,1478) degrees of freedom:	0.000000*
Joint Wald Statistic [4]:	308.309782	Prob(>chi-squared), (3) degrees of freedom:	0.000000*
Koenker (BP) Statistic [5]:	232.194311	Prob(>chi-squared), (3) degrees of freedom:	0.000000*
Jarque-Bera Statistic [6]:	835.125479	Prob(>chi-squared), (2) degrees of freedom:	0.000000*

# Regression analysis output



The R squared value is a measure of how well the estimated values explain the variance in the observed values

## Summary of OLS Results

Variable	Coefficient	StdError	t-Statistic	Probability	Robust_SE	Robust_t	Robust_Pr	VIF [1]
Intercept	1.588164	0.127854	12.421710	0.000000*	0.138934	11.431081	0.000000*	-----
INCOME	-0.000029	0.000003						
VANDALISM	0.133469	0.007815						
HOUSEHOLD	0.012620	0.001200						

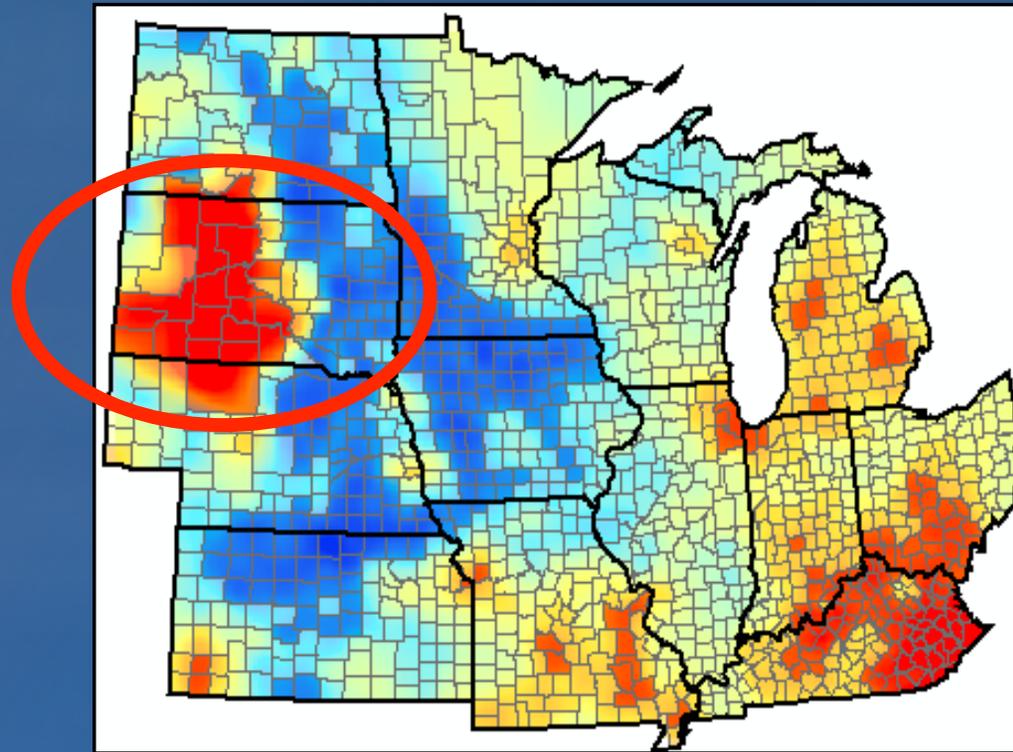
**Adjusted R-Squared [2]: 0.37407**  
**Akaike's Information Criterion (AIC) [2]: 5813.121**

## OLS Diagnostics

Number of Observations:	1482	Number of Variables:	4
Degrees of Freedom:	1478	Akaike's Information Criterion (AIC) [2]:	5813.121
Multiple R-Squared [2]:	0.375339	Adjusted R-Squared [2]:	0.374071
Joint F-Statistic [3]:	296.027584	Prob(>F), (3,1478) degrees of freedom:	0.000000*
Joint Wald Statistic [4]:	308.309782	Prob(>chi-squared), (3) degrees of freedom:	0.000000*
Koenker (BP) Statistic [5]:	232.194311	Prob(>chi-squared), (3) degrees of freedom:	0.000000*
Jarque-Bera Statistic [6]:	835.125479	Prob(>chi-squared), (2) degrees of freedom:	0.000000*

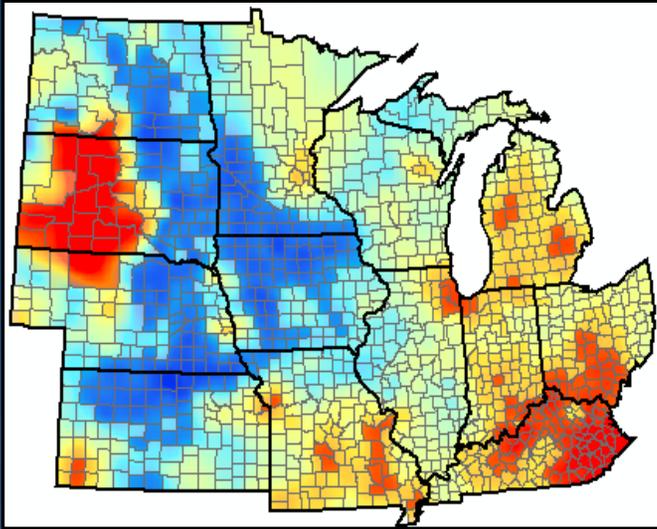
# Demo

# OLS Regression



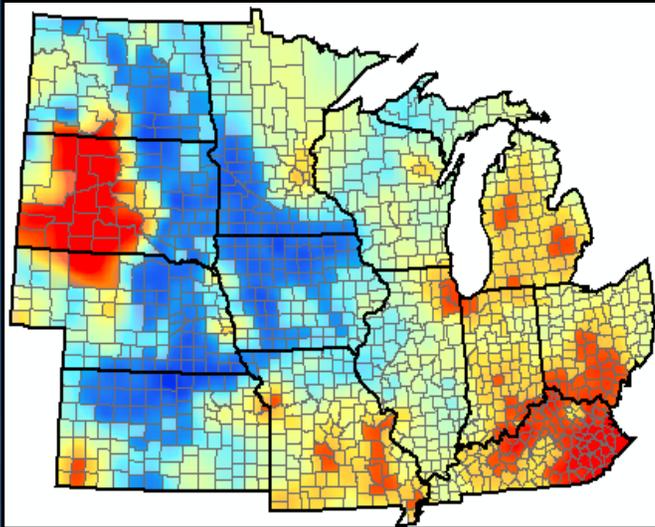
Why are people dying young in South Dakota?

# Use OLS to test hypotheses

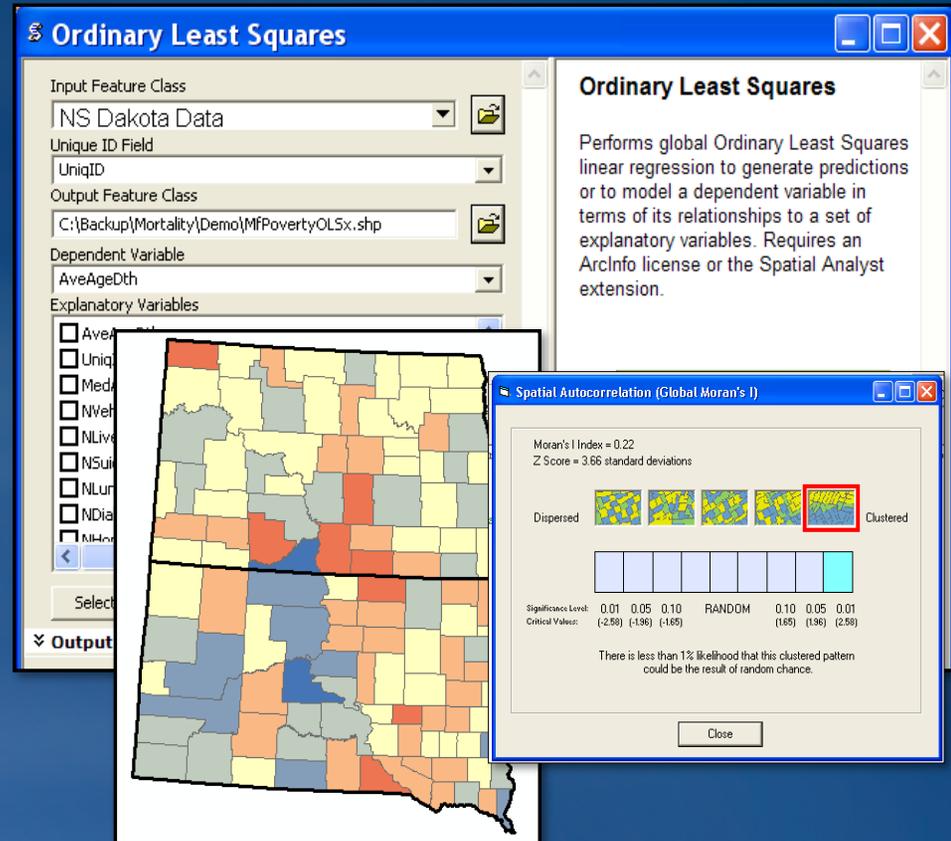


**Why are people dying young in South Dakota?  
Do economic factors explain this spatial pattern?**

# Use OLS to test hypotheses



Why are people dying young in South Dakota?  
Do economic factors explain this spatial pattern?



**Ordinary Least Squares**

Input Feature Class: NS Dakota Data  
Unique ID Field: UniqID  
Output Feature Class: C:\Backup\Mortality\Demo\MFPovertyOLSx.shp  
Dependent Variable: AveAgeDth  
Explanatory Variables:  AveAgeDth,  UniqID,  Med,  NVer,  NLive,  NSur,  NLun,  NDi,  NIn

**Spatial Autocorrelation (Global Moran's I)**

Moran's I Index = 0.22  
Z Score = 3.66 standard deviations

Dispersed      Clustered

Significance Level: 0.01 0.05 0.10 RANDOM 0.10 0.05 0.01  
Critical Values: (-2.58) (-1.96) (-1.65) (1.65) (1.96) (2.58)

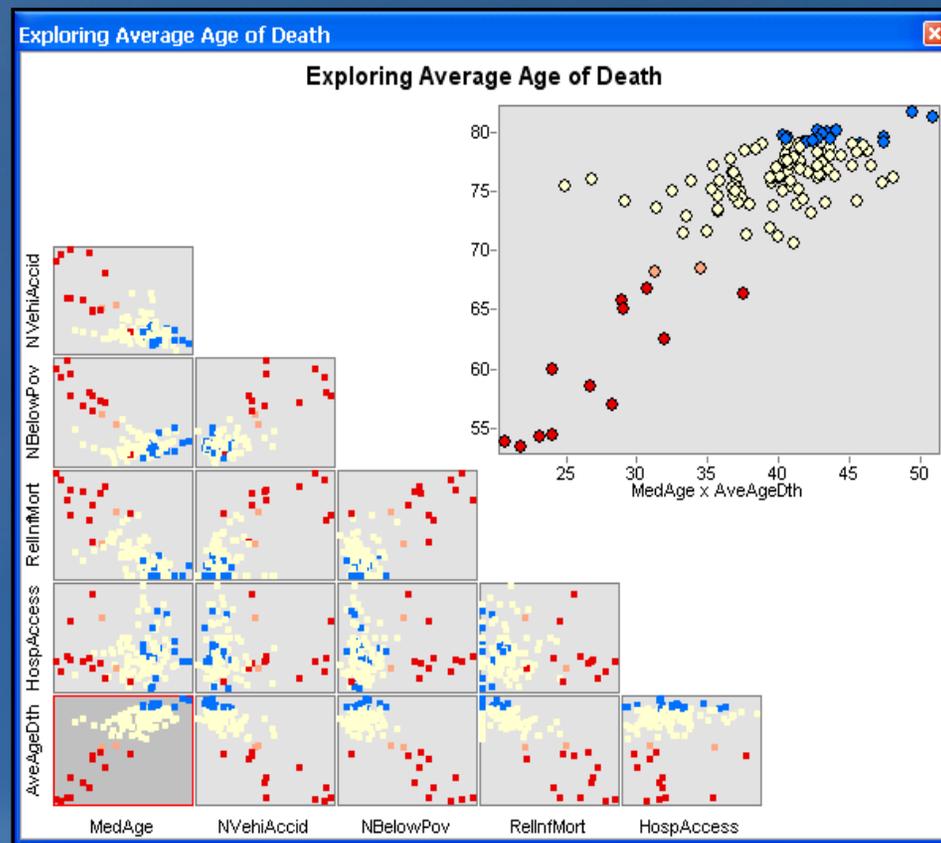
There is less than 1% likelihood that this clustered pattern could be the result of random chance.

Poverty rates explain 66% of the variation in the average age of death  
dependent variable: **Adjusted R-Squared [2]: 0.659**

However, significant spatial autocorrelation among model residuals indicates important explanatory variables are missing from the model.

# Build a multivariate regression model

- Explore variable relationships using the scatterplot matrix
- Consult theory and field experts
- Look for spatial variables
- Run OLS (this is an iterative, trial and error, process)



# Check OLS results

1

Coefficients have the expected sign.



2

No redundancy among model explanatory variables.



3

Coefficients are statistically significant.



Summary of OLS Results								
Variable	Coefficient	StdError	t-Statistic	Probability	Robust_SE	Robust_t	Robust_Pr	VIF [1]
Intercept	86.082979	0.875151	98.363521	0.000000*	0.813152	105.863324	0.000000*	-----
NVEHIACCID	-110.520016	12.213013	-9.049366	0.000000*	14.544464	-7.598769	0.000000*	2.351229
NSUICIDE	-138.221155	18.180324	-7.602788	0.000000*	29.800993	-4.638139	0.000011*	1.556498
NLUNGCANC	-47.045741	12.076316	-3.895703	0.000172*	13.536130	-3.475568	0.000732*	1.051207
NDIABETES	-33.429850	13.805975	-2.421405	0.017044*	14.732174	-2.269173	0.025148*	1.400358
NBELOWPOV	-14.408804	3.633873	-3.965137	0.000134*	4.125643	-3.492499	0.000692*	3.232363

OLS Diagnostics			
Number of Observations:	119	Number of Variables:	6
Degrees of Freedom:	113	Akaike's Information Criterion (AIC) [2]:	524.97620
Multiple R-Squared [2]:	0.870551	Adjusted R-Squared [2]:	0.864823
Joint F-Statistic [3]:	151.985705	Prob(>F), (5,113) degrees of freedom:	0.000000*
Joint Wald Statistic [4]:	496.057428	Prob(>chi-squared), (5) degrees of freedom:	0.000000*
Koenker (BP) Statistic [5]:	21.590491	Prob(>chi-squared), (5) degrees of freedom:	0.000626*
Jarque-Bera Statistic [6]:	4.207198	Prob(>chi-squared), (2) degrees of freedom:	0.122017

4

Residuals are normally distributed.



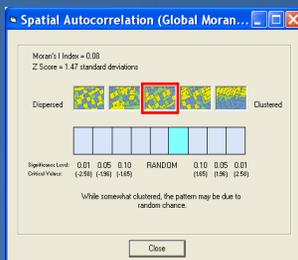
5

Strong Adjusted R-Square value.



6

Residuals are not spatially autocorrelated.



# Online help is ... helpful!

Contents | Index | Favorites | Search

- [-] Spatial Statistics toolbox
  - [+] An overview of the Spatial Sta
  - [+] Modeling spatial relationships
  - [+] Spatial Statistics toolbox sam
  - [+] What is a Z Score? What is a
  - [+] Analyzing Patterns toolset
  - [+] Mapping Clusters toolset
  - [+] Measuring Geographic Distrib
  - [-] Modeling Spatial Relationship
    - [+] An overview of the Modelin
    - [+] Spatial weights
    - [+] Regression analysis basic



## Regression analysis basics

 [Related topics](#)

The spatial statistics toolbox provides effective tools for quantifying spatial patterns. Using the Hot Spot Analysis tool, for example, you can ask questions like:

1. Are there places in the United States where people are persistently dying young?
2. Where are the hot spots for crime, 911 emergency calls (see graphic below), or fires?
3. Where do we find a higher than expected proportion of traffic accidents in a city?

Contents | Index | Favorites | Search

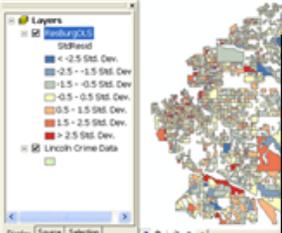
- [+] Modeling spatial relationships
- [+] Spatial Statistics toolbox sam
- [+] What is a Z Score? What is a
- [+] Analyzing Patterns toolset
- [+] Mapping Clusters toolset
- [+] Measuring Geographic Distrib
- [-] Modeling Spatial Relationship
  - [+] An overview of the Modelin
  - [+] Spatial weights
  - [+] Regression analysis basic
  - [+] Tools
    - [+] Generate Network Spat
    - [+] Generate Spatial Weigt
    - [+] Geographically Weight
    - [+] Ordinary Least Squares



## Interpreting OLS results

Output generated from the OLS Regression tool includes:

- Output feature class.



<b>Common Regression Problems, Consequences, and Solutions</b>		
<p>Omitted explanatory variables (misspecification).</p>	<p>When key explanatory variables are missing from a regression model, coefficients and their associated p-values cannot be trusted.</p>	<p>Map and examine <a href="#">OLS residuals</a> and <a href="#">GWR coefficients</a>, or run <a href="#">Hot Spot Analysis</a> on OLS regression residuals to see if this provides clues about possible missing variables.</p>
<p>Non-linear relationships. <a href="#">View an illustration.</a></p>	<p>OLS and GWR are both linear models. If the relationship between any of the explanatory variables and the dependent variable is non-linear, the resultant model will perform poorly.</p>	<p>Use the <a href="#">scatterplot matrix</a> graphic to elucidate the relationships among all variables in the model. Pay careful attention to relationships involving the dependent variable. Curvilinearity can often be remedied by transforming the variables. <a href="#">View an illustration.</a> Alternatively, use a non-linear regression method.</p>
<p>Data Outliers. <a href="#">View an illustration.</a></p>	<p>Influential outliers can pull modeled regression relationships away from</p>	<p>Use the <a href="#">scatterplot matrix</a> and other graphing tools to examine extreme data values. Correct or remove</p>

# Coefficient significance

- Look for statistically significant explanatory variables.

Notes on Interpretation

\* Statistically significant at the 0.05 level.

[1] Large VIF (> 7.5, for example) indicates explanatory variable redundancy.  
 [2] Measure of model fit/performance.  
 [3] Significant p-value indicates overall model significance.  
 [4] Significant p-value indicates robust overall model significance.  
 [5] Significant p-value indicates biased standard errors; use robust estimates.  
 [6] Significant p-value indicates residuals deviate from a normal distribution.

\* Statistically significant at the 0.05 level.



Summary of OLS Results								
Variable	Coefficient	StdError	t-Statistic	Probability	Robust_SE	Robust_t	Robust_Pr	VIF [1]
Intercept	86.082979	0.875151	98.363521	0.000000*	0.813152	105.863324	0.000000*	-----
NVEHIACCID	-110.520016	12.213013	-9.049366	0.000000*	14.544464	-7.598769	0.000000*	2.351229
NSUICIDE	-138.221155	18.180324	-7.602788	0.000000*	29.800993	-4.638139	0.000011*	1.556498
NLUNGCANC	-47.045741	12.076316	-3.895703	0.000172*	13.536130	-3.475568	0.000732*	1.051207
NDIABETES	-33.429850	13.805975	-2.421405	0.017044*	14.732174	-2.269173	0.025148*	1.400358
NBELOWPOV	-14.408804	3.633873	-3.965137	0.000134*	4.125643	-3.492499	0.000692*	3.232363

OLS Diagnostics			
Number of Observations:	119	Number of Variables:	6
Degrees of Freedom:	113	Akaike's Information Criterion (AIC) [2]:	524.97620
Multiple R-Squared [2]:	0.870551	Adjusted R-Squared [2]:	0.864823
Joint F-Statistic [3]:	151.985705	Prob(>F), (5,113) degrees of freedom:	0.000000*
Joint Wald Statistic [4]:	496.057428	Prob(>chi-squared), (5) degrees of freedom:	0.000000*
Koenker (BP) Statistic [5]:	21.590491	Prob(>chi-squared), (5) degrees of freedom:	0.000626*
Jarque-Bera Statistic [6]:	4.207198	Prob(>chi-squared), (2) degrees of freedom:	0.122017

# Coefficient significance

- Look for statistically significant explanatory variables.

Notes on Interpretation

\* Statistically significant at the 0.05 level.

[1] Large VIF (> 7.5, for example) indicates explanatory variable redundancy.  
 [2] Measure of model fit/performance.  
 [3] Significant p-value indicates overall model significance.  
 [4] Significant p-value indicates robust overall model significance.  
 [5] Significant p-value indicates biased standard errors; use robust estimates.  
 [6] Significant p-value indicates residuals deviate from a normal distribution.

\* Statistically significant at the 0.05 level.

Probability	Robust_Prob
0.000000*	0.000000*
0.000000*	0.000000*
0.000000*	0.000011*
0.000172*	0.000732*
0.017044*	0.025148*
0.000134*	0.000692*

Summary of OLS Results								
Variable	Coefficient	StdError	t-Statistic	Probability	Robust_SE	Robust_t	Robust_Pr	VIF [1]
Intercept	86.082979	0.875151	98.363521	0.000000*	0.813152	105.863324	0.000000*	-----
NVEHIACCID	-110.520016	12.213013	-9.049366	0.000000*	14.544464	-7.598769	0.000000*	2.351229
NSUICIDE	-138.221155	18.180324	-7.602788	0.000000*	29.800993	-4.638139	0.000011*	1.556498
NLUNGCANC	-47.045741	12.076316	-3.895703	0.000172*	13.536130	-3.475568	0.000732*	1.051207
NDIABETES	-33.429850	13.805975	-2.421405	0.017044*	14.732174	-2.269173	0.025148*	1.400358
NBELOWPOV	-14.408804	3.633873	-3.965137	0.000134*	4.125643	-3.492499	0.000692*	3.232363

OLS Diagnostics

Number of Observations: 119      Number of Variables: 6

Degrees of Freedom: 113      Akaike's Information Criterion (AIC) [2]: 524.97620

Multiple R-Squared [2]: 0.870551      Adjusted R-Squared [2]: 0.864823

Joint F-Statistic [3]: 151.985705      Prob(>F), (5,113) degrees of freedom: 0.000000\*

Joint Wald Statistic [4]: 496.057428      Prob(>chi-squared), (5) degrees of freedom: 0.000000\*

Koenker (BP) Statistic [5]: 21.590491      Prob(>chi-squared), (5) degrees of freedom: 0.000626\*

Jarque-Bera Statistic [6]: 4.207198      Prob(>chi-squared), (2) degrees of freedom: 0.122017

Koenker(BP) Statistic [5]: 38.994033 Prob(>chi-squared), (5) degrees of freedom: 0.000626\*

# Check redundancy in explanatory variables

- **Multicollinearity**
  - Term used to describe the phenomenon when two or more of the variables in your model are highly correlated
- **Variance Inflation Factor (VIF)**
  - Detects the severity of multicollinearity
  - Explanatory variables with a  $VIF > 7.5$  should be removed one by one from the model

# Multicollinearity

- Find a set of explanatory variables that have low VIF values.
- In a strong model, each explanatory variable gets at a different facet of the dependent variable.

Notes on Interpretation

\* Statistically significant at the 0.05 level.

[1] Large VIF (> 7.5, for example) indicates explanatory variable redundancy.

[2] Measure of model fit/performance.

[3] Significant p-value indicates overall model significance.

[4] Significant p-value indicates robust overall model significance.

[5] Significant p-value indicates biased standard errors; use robust estimates.

[6] Significant p-value indicates residuals deviate from a normal distribution.



VIF

-----  
 2.351229  
 1.556498  
 1.051207  
 1.400358  
 3.232363

[1] Large VIF (> 7.5, for example) indicates explanatory variable redundancy.

Summary of OLS Results								
Variable	Coefficient	StdError	t-Statistic	Probability	Robust_SE	Robust_t	Robust_Pr	VIF [1]
Intercept	86.082979	0.875151	98.363521	0.000000*	0.813152	105.863324	0.000000*	-----
NVEHIACCID	-110.520016	12.213013	-9.049366	0.000000*	14.544464	-7.598769	0.000000*	2.351229
NSUICIDE	-138.221155	18.180324	-7.602788	0.000000*	29.800993	-4.638139	0.000011*	1.556498
NLUNGCANC	-47.045741	12.076316	-3.895703	0.000172*	13.536130	-3.475568	0.000732*	1.051207
NDIABETES	-33.429850	13.805975	-2.421405	0.017044*	14.732174	-2.269173	0.025148*	1.400358
NBELOWPOV	-14.408804	3.633873	-3.965137	0.000134*	4.125643	-3.492499	0.000692*	3.232363

# Model performance

- Compare models by looking for the lowest AIC value.
  - As long as the dependent variable remains fixed, the AIC value for different OLS/GWR models are comparable
- Look for a model with a high Adjusted R-Squared value.

## Notes on Interpretation

\* Statistically significant at the 0.05 level.  
[1] Large VIF (> 7.5, for example) indicates explanatory variable redundancy.  
[2] Measure of model fit/performance.  
[3] Significant p-value indicates overall model significance.  
[4] Significant p-value indicates robust overall model significance.  
[5] Significant p-value indicates biased standard errors; use robust estimates.  
[6] Significant p-value indicates residuals deviate from a normal distribution.

[2] Measure of model fit/performance.

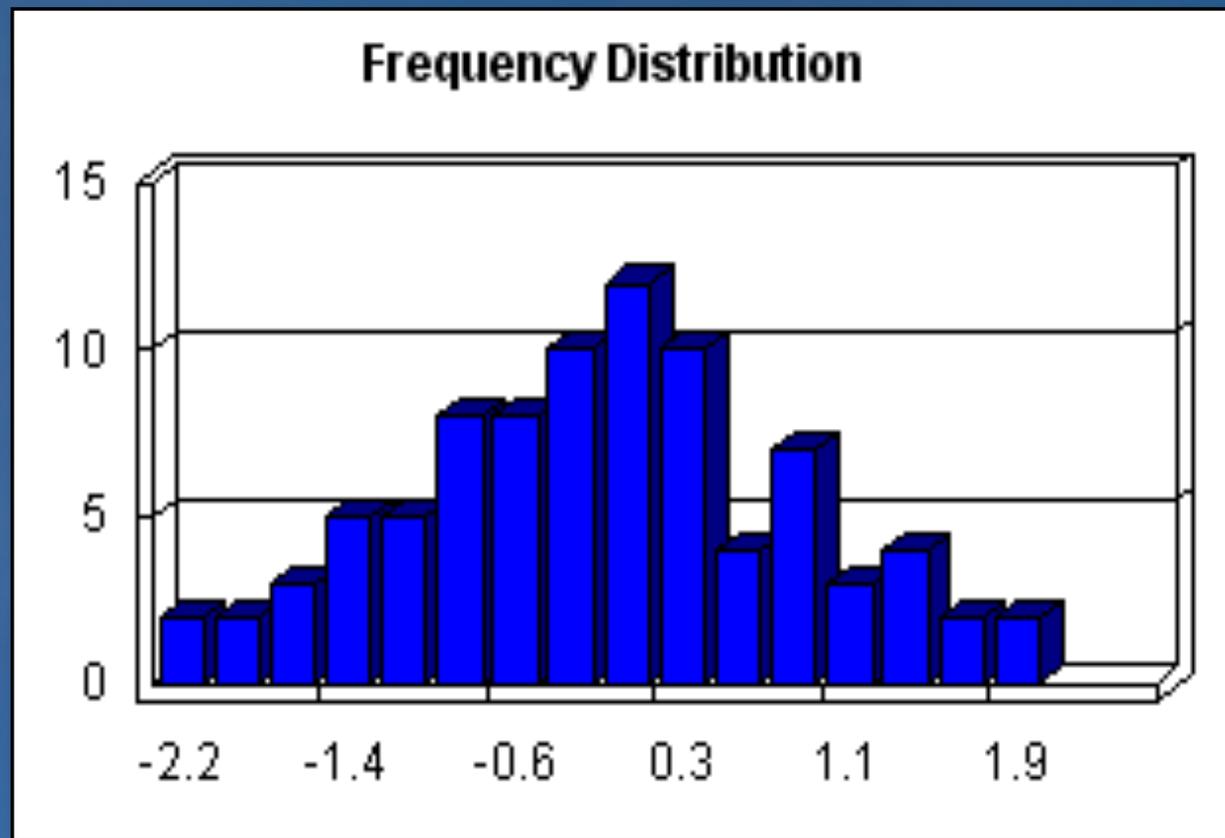


**Akaike's Information Criterion (AIC) [2]: 524.9762**  
**Adjusted R-Squared [2]: 0.864823**

		OLS Diagnostics	
Number of Observations:	119	Number of Variables:	6
Degrees of Freedom:	113	Akaike's Information Criterion (AIC) [2]:	524.9762
Multiple R-Squared [2]:	0.870551	Adjusted R-Squared [2]:	0.864823
Joint F-Statistic [3]:	151.985705	Prob(>F), (5,113) degrees of freedom:	0.000000*
Joint Wald Statistic [4]:	496.057428	Prob(>chi-squared), (5) degrees of freedom:	0.000000*
Koenker (BP) Statistic [5]:	21.590491	Prob(>chi-squared), (5) degrees of freedom:	0.000626*
Jarque-Bera Statistic [6]:	4.207198	Prob(>chi-squared), (2) degrees of freedom:	0.122017

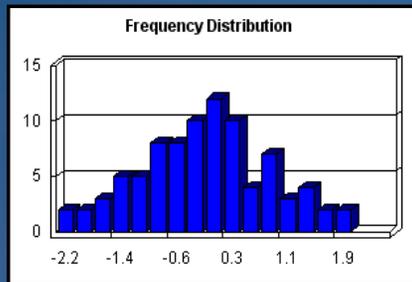
# Checking for model bias

- The residuals of a good model should be normally distributed with a mean of zero
- The Jarque-Bera test checks model bias



# Model bias

- When the Jarque-Bera test is statistically significant:
  - The model is biased
  - Results are *not* reliable
  - Often indicates that a key variable is missing from the model



Notes on Interpretation

\* Statistically significant at the 0.05 level.

[1] Large VIF (> 7.5, for example) indicates explanatory variable redundancy.

[2] Measure of model fit/performance.

[3] Significant p-value indicates overall model significance.

[4] Significant p-value indicates robust overall model significance.

[5] Significant p-value indicates biased standard errors; use robust estimates.

[6] Significant p-value indicates residuals deviate from a normal distribution.

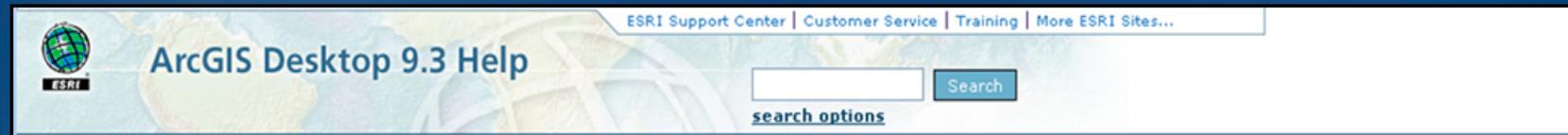
[6] Significant p-value indicates residuals deviate from a normal distribution.

Jarque-Bera Statistic [6]: 4.207198 Prob(>chi-sq), (2) degrees of freedom: 0.122017

OLS Diagnostics			
Number of Observations:	119	Number of Variables:	6
Degrees of Freedom:	113	Akaike's Information Criterion (AIC) [2]:	524.9762
Multiple R-Squared [2]:	0.870551	Adjusted R-Squared [2]:	0.864823
Joint F-Statistic [3]:	151.985705	Prob(>F), (5,113) degrees of freedom:	0.000000*
Joint Wald Statistic [4]:	496.057428	Prob(>chi-squared), (5) degrees of freedom:	0.000000*
Koenker (BP) Statistic [5]:	21.590491	Prob(>chi-squared), (5) degrees of freedom:	0.000626*
Jarque-Bera Statistic [6]:	4.207198	Prob(>chi-squared), (2) degrees of freedom:	0.122017



# How regression models go bad... ...and how to fix them!



There will be times, however, when the missing variables are too complex to model, impossible to quantify, or too difficult to measure. In these cases, you may be able to move to GWR or to another spatial regression method to get a well-specified model.

The following table lists common problems with regression models and the tools available in ArcGIS to help address them:

## Common Regression Problems, Consequences, and Solutions

Omitted explanatory variables (misspecification).	When key explanatory variables are missing from a regression model, coefficients and their associated p-values cannot be trusted.	Map and examine <a href="#">OLS residuals</a> and <a href="#">GWR coefficients</a> , or run <a href="#">Hot Spot Analysis</a> on OLS regression residuals to see if this provides clues about possible missing variables.
Nonlinear relationships. <a href="#">View an illustration.</a>	OLS and GWR are both linear models. If the relationship between any of the explanatory variables and the dependent variable is nonlinear, the resultant model will perform poorly.	Use the <a href="#">scatterplot matrix</a> graphic to elucidate the relationships among all variables in the model. Pay careful attention to relationships involving the dependent variable. Curvilinearity can often be remedied by transforming the variables. <a href="#">View an illustration.</a> Alternatively, use a nonlinear regression method.
Data outliers. <a href="#">View an illustration.</a>	Influential outliers can pull modeled regression relationships away from their true best fit, biasing regression coefficients.	Use the <a href="#">scatterplot matrix</a> and other graphing tools to examine extreme data values. Correct or remove outliers if they represent errors. When outliers are correct/valid values, they cannot/should not be removed. Run the regression with and without the outliers to see how much they are affecting your results.

# Check OLS results

1

Coefficients have the expected sign.



2

No redundancy among model explanatory variables.



3

Coefficients are statistically significant.



Summary of OLS Results								
Variable	Coefficient	StdError	t-Statistic	Probability	Robust_SE	Robust_t	Robust_Pr	VIF [1]
Intercept	86.082979	0.875151	98.363521	0.000000*	0.813152	105.863324	0.000000*	-----
NVEHIACCID	-110.520016	12.213013	-9.049366	0.000000*	14.544464	-7.598769	0.000000*	2.351229
NSUICIDE	-138.221155	18.180324	-7.602788	0.000000*	29.800993	-4.638139	0.000011*	1.556498
NLUNGCANC	-47.045741	12.076316	-3.895703	0.000172*	13.536130	-3.475568	0.000732*	1.051207
NDIABETES	-33.429850	13.805975	-2.421405	0.017044*	14.732174	-2.269173	0.025148*	1.400358
NBELOWPOV	-14.408804	3.633873	-3.965137	0.000134*	4.125643	-3.492499	0.000692*	3.232363

OLS Diagnostics			
Number of Observations:	119	Number of Variables:	6
Degrees of Freedom:	113	Akaike's Information Criterion (AIC) [2]:	524.97620
Multiple R-Squared [2]:	0.870551	Adjusted R-Squared [2]:	0.864823
Joint F-Statistic [3]:	151.985705	Prob(>F), (5,113) degrees of freedom:	0.000000*
Joint Wald Statistic [4]:	496.057428	Prob(>chi-squared), (5) degrees of freedom:	0.000000*
Koenker (BP) Statistic [5]:	21.590491	Prob(>chi-squared), (5) degrees of freedom:	0.000626*
Jarque-Bera Statistic [6]:	4.207198	Prob(>chi-squared), (2) degrees of freedom:	0.122017

4

Residuals are normally distributed.



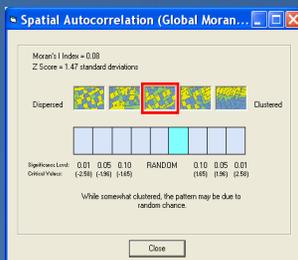
5

Strong Adjusted R-Square value.



6

Residuals are not spatially autocorrelated.



# Exploring Spatial Variation:

## **GWR**

**Geographically Weighted  
Regression**

# Global vs. local regression models

- **OLS**
  - Global regression model
  - One equation, calibrated using data from all features
  - Relationships are fixed

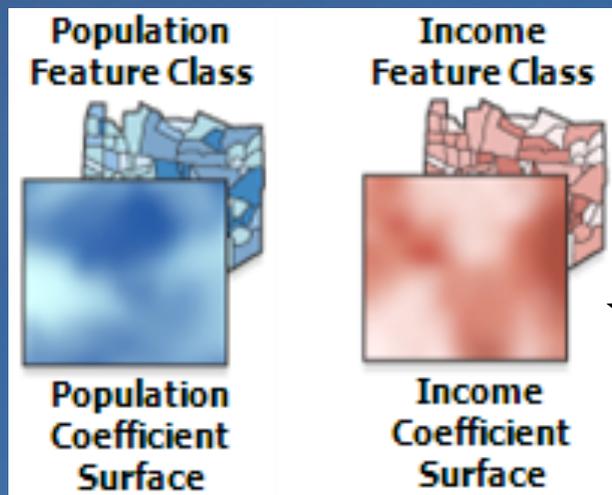
# Global vs. local regression models

- OLS

- Global regression model
- One equation, calibrated using data from all features
- Relationships are fixed

- GWR

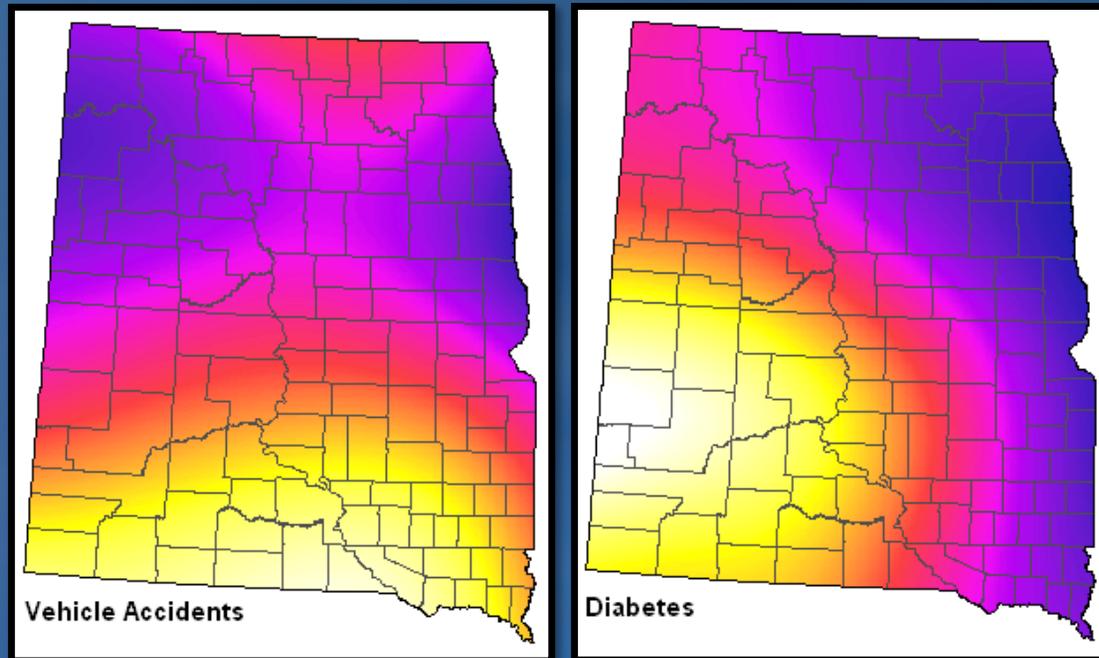
- Local regression model
- One equation for every feature (e.g., county, block group, intersection)
- Equations are calibrated using data from nearby features
- Relationships are allowed to vary across the study area



For each explanatory variable, GWR creates a coefficient surface showing you *where* relationships are strongest.

# GWR

## Demonstration



Exploring regional variation

# Running GWR

- **GWR is a local spatial regression model**
  - Modeled relationships are allowed to vary
- **GWR variables are the same as OLS, except:**
  - **Do not include spatial regime (dummy) variables**
  - **Do not include variables with little value variation**

**Geographically Weighted...**

Input feature class: NS Dakota Data

Dependent variable: AveAgeDth

Explanatory variable(s):

- NvehiAccid
- NSuicide
- NLungCanc
- NDiabetes
- NBelowPov

Output feature class: C:\Backup\Mortality\Demo\GWROutputFC.shp

Kernel type: FIXED

Bandwidth method: AICc

Distance (optional):

Number of neighbors (optional): 30

Weights (optional):

**Additional Parameters (Optional)**

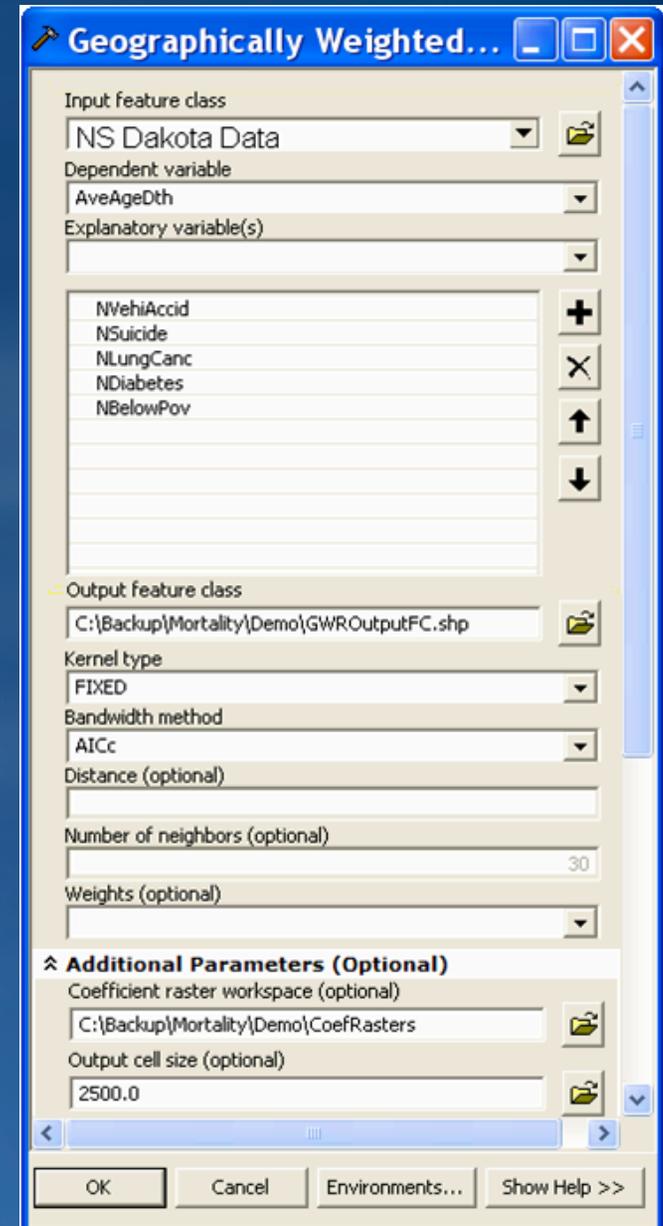
Coefficient raster workspace (optional): C:\Backup\Mortality\Demo\CoefRasters

Output cell size (optional): 2500.0

OK Cancel Environments... Show Help >>

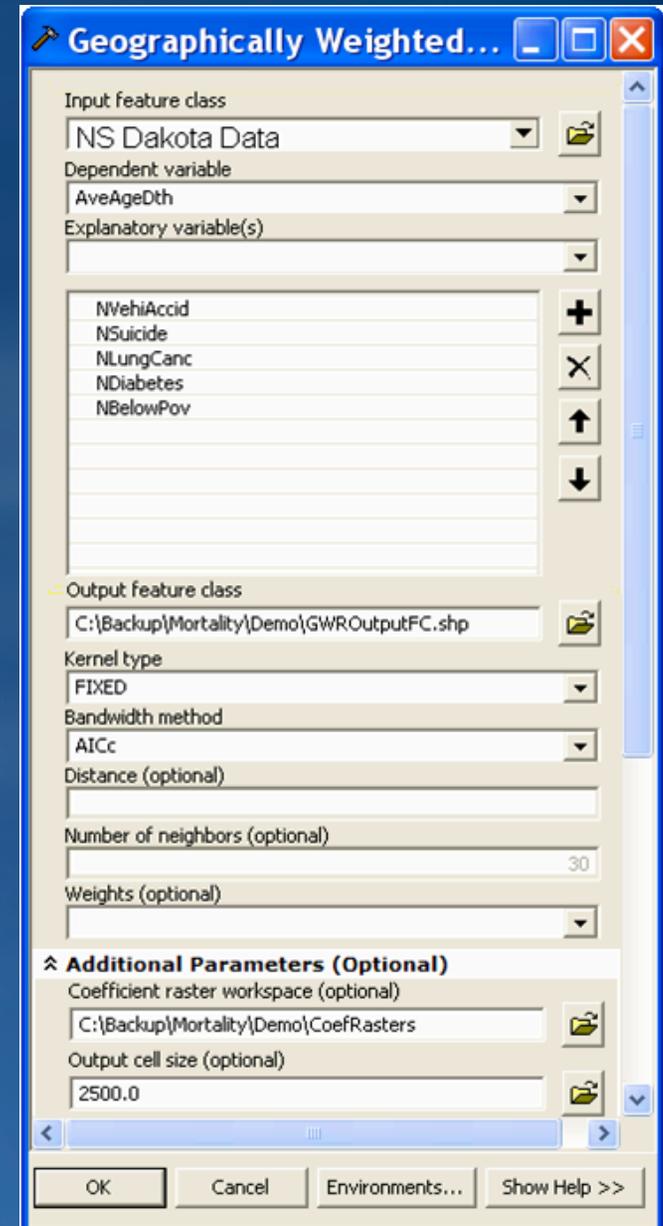
# Defining *local*

- GWR constructs an equation for each feature



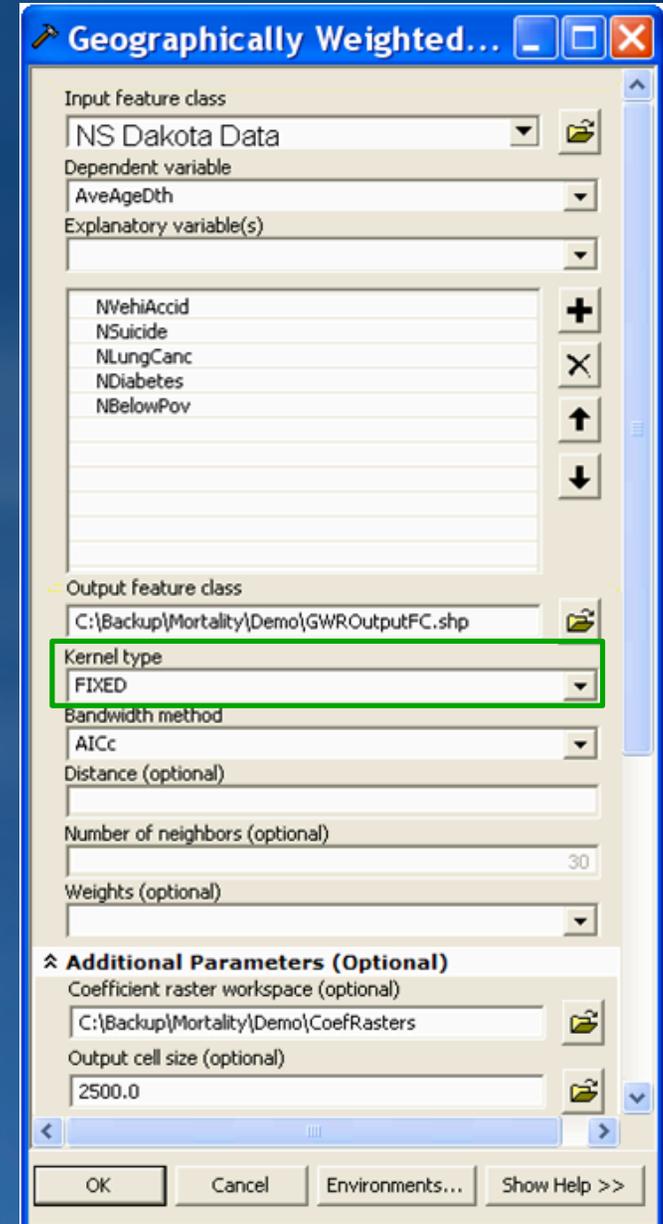
# Defining *local*

- GWR constructs an equation for each feature
- Coefficients are estimated using nearby feature values



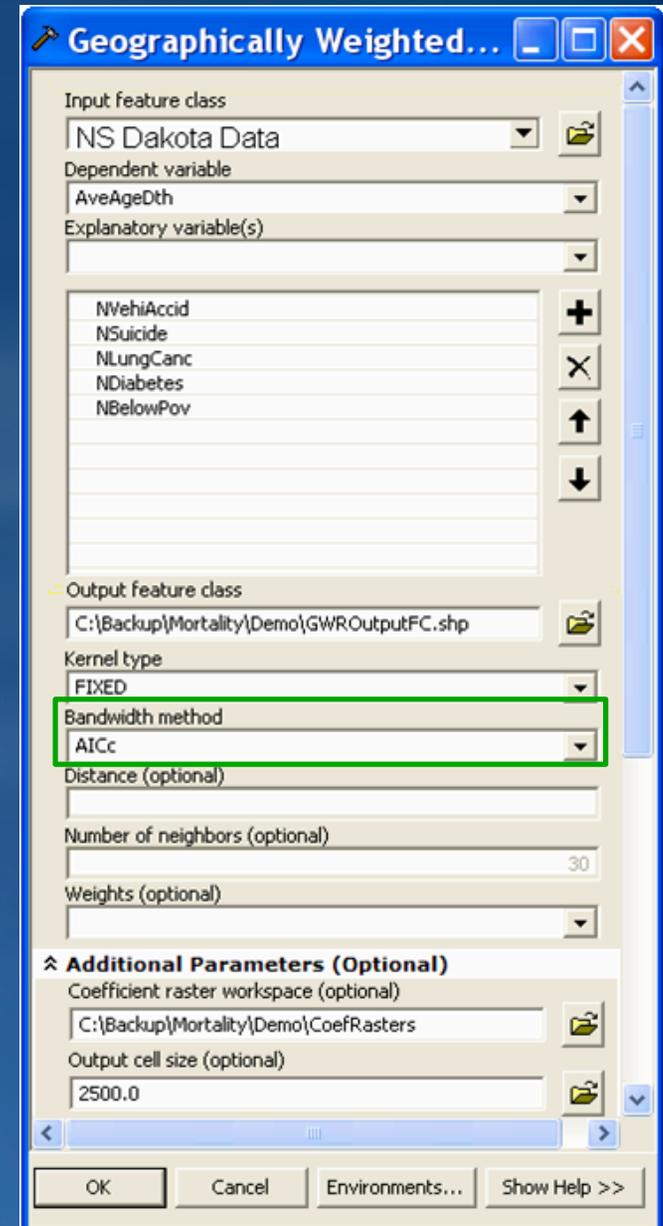
# Defining *local*

- GWR constructs an equation for each feature
- Coefficients are estimated using nearby feature values
- GWR requires a definition for *nearby*
  - Kernel type
    - Fixed: *Nearby* is determined by a fixed distance band
    - Adaptive: *Nearby* is determined by a fixed number of neighbors



# Defining *local*

- GWR constructs an equation for each feature
- Coefficients are estimated using nearby feature values
- GWR requires a definition for *nearby*
  - Kernel type
    - Fixed: *Nearby* is determined by a fixed distance band
    - Adaptive: *Nearby* is determined by a fixed number of neighbors
  - Bandwidth method
    - AIC or Cross Validation (CV): GWR will find the optimal distance or optimal number of neighbors
    - Bandwidth parameter: User-provided distance or user-provided number of neighbors



# Interpreting GWR results

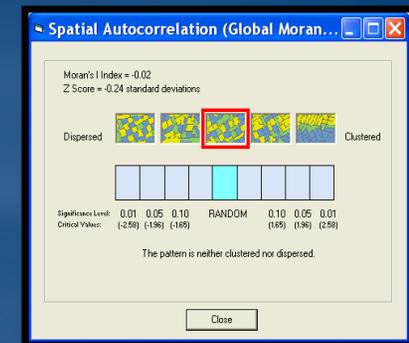
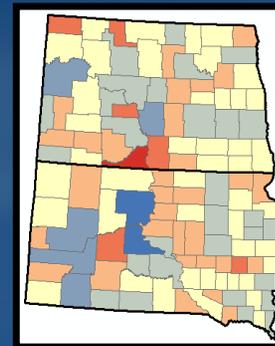
- **Compare GWR  $R^2$  and AIC values to OLS  $R^2$  and AIC values**
  - The better model has a lower AIC and a high  $R^2$ .

Bandwidth	: 2e+005
ResidualSquares	: 327.57434924067235
EffectiveNumber	: 30.68145239098456
Sigma	: 1.9258789406364027
AICc	: 518.280903017286
R2	: 0.9183016622131718
R2Adjusted	: 0.8908450815844072

# Interpreting GWR results

- Compare GWR  $R^2$  and AIC values to OLS  $R^2$  and AIC values
  - The better model has a lower AIC and a high  $R^2$ .
- Residual maps show model under- and over-predictions. They shouldn't be clustered.

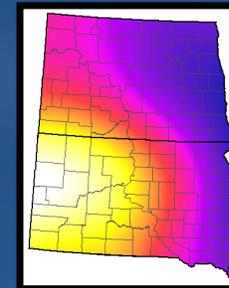
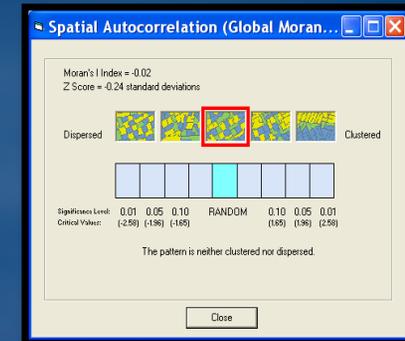
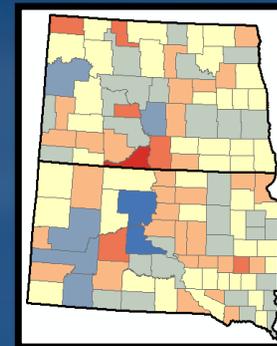
```
Bandwidth : 2e+005
ResidualSquares : 327.57434924067235
EffectiveNumber : 30.68145239098456
Sigma : 1.9258789406364027
AICc : 518.280903017286
R2 : 0.9183016622131718
R2Adjusted : 0.8908450815844072
```



# Interpreting GWR results

- Compare GWR  $R^2$  and AIC values to OLS  $R^2$  and AIC values
  - The better model has a lower AIC and a high  $R^2$ .
- Residual maps show model under- and over-predictions. They shouldn't be clustered.
- Coefficient maps show how modeled relationships vary across the study area.

```
Bandwidth : 2e+005
ResidualSquares : 327.57434924067235
EffectiveNumber : 30.68145239098456
Sigma : 1.9258789406364027
AICc : 518.280903017286
R2 : 0.9183016622131718
R2Adjusted : 0.8908450815844072
```

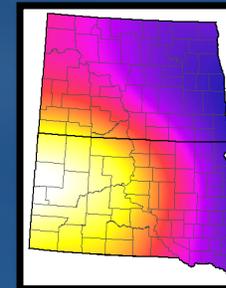
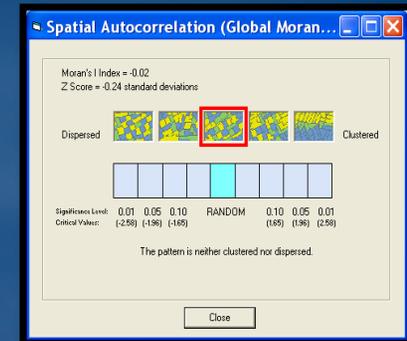
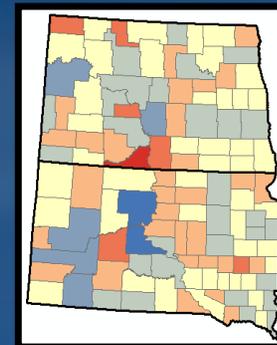


# Interpreting GWR results

- Compare GWR  $R^2$  and AIC values to OLS  $R^2$  and AIC values
  - The better model has a lower AIC and a high  $R^2$ .
- Residual maps show model under- and over-predictions. They shouldn't be clustered.
- Coefficient maps show how modeled relationships vary across the study area.
- Model predictions, residuals, standard errors, coefficients, and condition numbers are written to the output feature class.

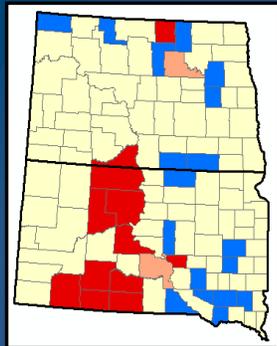
```

Bandwidth           : 2e+005
ResidualSquares    : 327.57434924067235
EffectiveNumber     : 30.68145239098456
Sigma               : 1.9258789406364027
AICc                : 518.280903017286
R2                  : 0.9183016622131718
R2Adjusted          : 0.8908450815844072
    
```

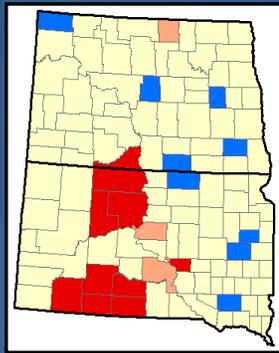


	Observed	Cond	LocalR2	Predicted	Intercept	C1_IVehiAc
▶	78.419998	12.613701	0.881075	78.510341	85.562468	-79.980532
	76.5	14.048718	0.834124	77.920484	85.36851	-78.018965
	68.209999	12.25915	0.847111	68.964384	85.920939	-80.417789
	73.190002	12.515339	0.857244	72.182168	84.312503	-65.634913
	78.290001	11.758007	0.836626	77.979938	85.876829	-79.211314
	78.230003	12.612641	0.874848	79.159514	84.765217	-77.439124

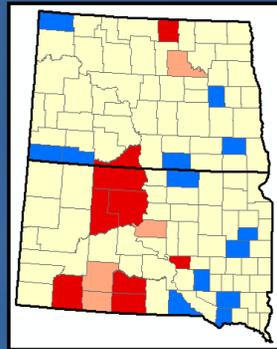
# GWR prediction



Observed



Modeled



Predicted

### Geographically Weighted Regression

Input feature class: DakotaData.shp  
Dependent variable: AveAgeDth  
Explanatory variable(s): NWhiAccid, NSuicide, NLungCanc, NDiabetes, NBelowPov

Output feature class: C:\Backup\Mortality\Demo\MFVSLDGWR.shp  
Kernel type: FIXED  
Bandwidth method: BANDWIDTH PARAMETER  
Distance (optional): 200000  
Number of neighbors (optional): 30  
Weights (optional):

**Additional Parameters (Optional)**  
Coefficient raster workspace (optional):  
Output cell size (optional): 2467.599155  
Prediction locations (optional): DakotaData.shp  
Prediction explanatory variable(s) (optional): NWhiAccid, NSuicide, NLungCanc, FutureDiab, NBelowPov  
Output prediction feature class (optional): C:\Backup\Mortality\Demo\AvAgDthFYrPred.shp

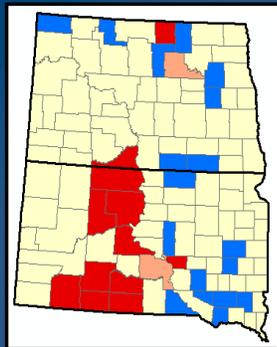
#### Geographically Weighted Regression

Performs GWR, a local form of linear regression used to model spatially varying relationships. Requires an ArcInfo, Spatial Analyst, or Geostatistical Analyst License.

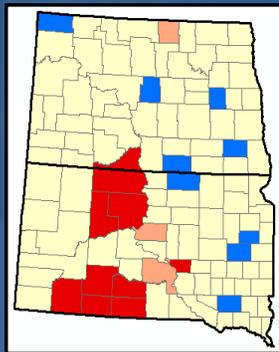
$$\beta_0 + \beta_1 \text{Population} + \beta_2 \text{Income} = \text{Crime}$$

# GWR prediction

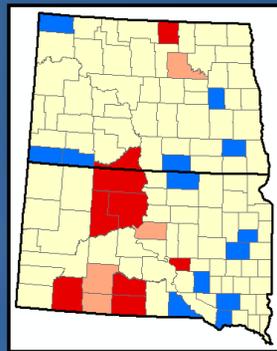
Calibrate the GWR model using known values for the dependent variable and all of the explanatory variables.



Observed



Modeled



Predicted

**Geographically Weighted Regression**

Input feature class: DakotaData.shp

Dependent variable: AveAgeDth

Explanatory variable(s): NWehiAccid, NSuicide, NLungCanc, NDiabetes, NBelowPov

Output feature class: C:\Backup\Mortality\Demo\MFVSLDGWR.shp

Kernel type: FIXED

Bandwidth method: BANDWIDTH PARAMETER

Distance (optional): 200000

Number of neighbors (optional): 30

Weights (optional):

**Additional Parameters (Optional)**

Coefficient raster workspace (optional):

Output cell size (optional): 2467.599155

Prediction locations (optional):

Prediction explanatory variable(s) (optional): NWehiAccid, NSuicide, NLungCanc, FutureDiab, NBelowPov

Output prediction feature class (optional): C:\Backup\Mortality\Demo\AvAgDthFYrPred.shp

**Geographically Weighted Regression**

Performs GWR, a local form of linear regression used to model spatially varying relationships. Requires an ArcInfo, Spatial Analyst, or Geostatistical Analyst License.

$\beta_0$

+

$\beta_1$  Population

+

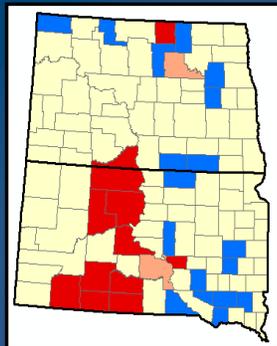
$\beta_2$  Income

=

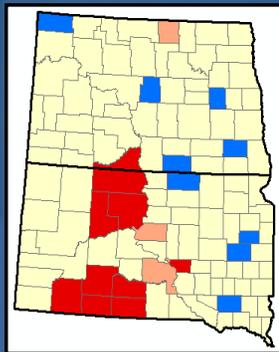
Crime

# GWR prediction

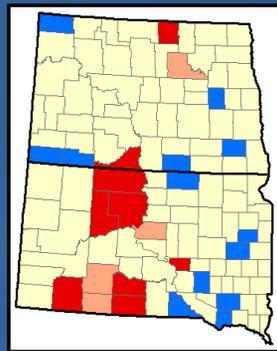
Calibrate the GWR model using known values for the dependent variable and all of the explanatory variables.



Observed



Modeled



Predicted

Provide a feature class of prediction locations containing values for all of the explanatory variables.

**Geographically Weighted Regression**

Input feature class: DakotaData.shp

Dependent variable: AveAgeDth

Explanatory variable(s): NWehiAccid, NSuicide, NLungCanc, NDiabetes, NBelowPov

Output feature class: C:\Backup\Mortality\Demo\MFVSLDGWR.shp

Kernel type: FIXED

Bandwidth method: BANDWIDTH PARAMETER

Distance (optional): 200000

Number of neighbors (optional): 30

Weights (optional):

**Additional Parameters (Optional)**

Coefficient raster workspace (optional):

Output cell size (optional): 2467.599155

Prediction locations (optional): DakotaData.shp

Prediction explanatory variable(s) (optional): NWehiAccid, NSuicide, NLungCanc, FutureDiab, NBelowPov

Output prediction feature class (optional): C:\Backup\Mortality\Demo\AvAgDthFYrPred.shp

**Geographically Weighted Regression**

Performs GWR, a local form of linear regression used to model spatially varying relationships. Requires an ArcInfo, Spatial Analyst, or Geostatistical Analyst License.

$B_0$

+

$B_1$  Population

+

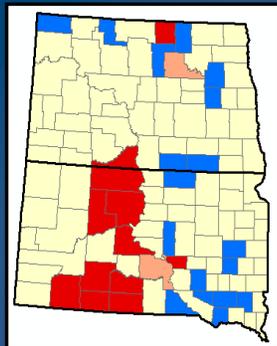
$B_2$  Income

=

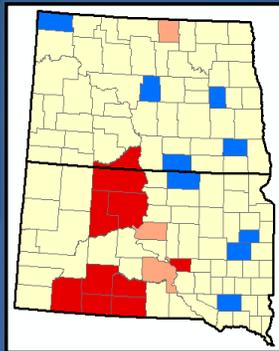
Crime

# GWR prediction

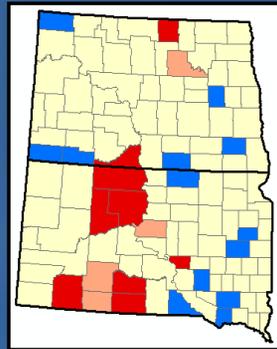
Calibrate the GWR model using known values for the dependent variable and all of the explanatory variables.



Observed



Modeled



Predicted

Provide a feature class of prediction locations containing values for all of the explanatory variables.

GWR will create an output feature class with the computed predictions.

**Geographically Weighted Regression**

Input feature class: DakotaData.shp  
Dependent variable: AveAgeDth  
Explanatory variable(s): NWhiAccid, NSuicide, NLungCanc, NDiabetes, NBelowPov

Output feature class: C:\Backup\Mortality\Demo\MFVSLDGWR.shp  
Kernel type: FIXED  
Bandwidth method: BANDWIDTH PARAMETER  
Distance (optional): 200000  
Number of neighbors (optional): 30  
Weights (optional):

**Additional Parameters (Optional)**  
Coefficient raster workspace (optional):  
Output cell size (optional): 2467.599155  
Prediction locations (optional): DakotaData.shp  
Prediction explanatory variable(s) (optional): NWhiAccid, NSuicide, NLungCanc, FutureDiab, NBelowPov

Output prediction feature class (optional): C:\Backup\Mortality\Demo\AvAgDthFYrPred.shp

**Geographically Weighted Regression**

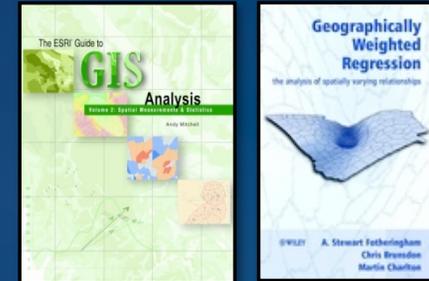
Performs GWR, a local form of linear regression used to model spatially varying relationships. Requires an ArcInfo, Spatial Analyst, or Geostatistical Analyst License.

$B_0$  +  $B_1$  Population +  $B_2$  Income = Crime

# Review

- **Regression is...**
  - Used to model relationships, understand patterns, and predict outcomes.
- **OLS is...**
  - A global regression model.
  - The starting point for regression analysis.
  - A way to find which variables are most important to the model.
  - Preceded by research and exploratory data analysis
  - An iterative process.
  - Only trusted if certain validation checks are passed.
- **GWR is...**
  - A local regression model.
  - Helpful when relationships in the data vary across space.
  - Useful to make predictions based on your local model.

# Resources for learning more...



- [The ESRI Guide to GIS Analysis, Vol. 2](#)
- [Geographically Weighted Regression](#), by Fotheringham, Brundson, and Charlton
- Virtual campus free web seminar (search for “regression”)  
<http://campus.esri.com/>
- Several articles can be found in the Spring 2009 issue of *ArcUser magazine* (view it online)
- *GEOconnexion International Magazine*, Dec/Jan 2009
- ArcGIS 9.3/9.3.1 Web Help (<http://webhelp.esri.com>)
  - Regression Analysis Basics
  - Interpreting OLS Results
  - Interpreting GWR Results

**Watch for updates**

- GP Resource Center

This is a screenshot of the ArcGIS Desktop 9.3 Help page. The page title is 'ArcGIS Desktop 9.3 Help'. The main content area is titled 'Regression analysis basics' and includes a 'You are here' breadcrumb trail: 'Geoprocessing tool reference &gt; Spatial Statistics toolbox &gt; Modeling Spatial Relationships toolset'. The page also shows a 'Release 9.3' label and a 'Last modified July 17, 2008' date. There are links for 'E-mail This Topic', 'Print This Topic', and 'Give Us Feedback'. A search bar is visible at the top right of the page.

This is a screenshot of the ESRI Resource Center for ArcGIS Desktop. The page title is 'ESRI Resource Center for ArcGIS Desktop'. The page includes a search bar, a 'Login' button, and a navigation menu with links for 'Home', 'Content', 'Web Help', 'For Developers', 'Support', and 'Other Resource Centers'. The ESRI logo is visible in the top left corner.

# Demonstration: Sharing Regression Analysis Results

## The “FedStat” Concept

video link:

[http://www.esri.com/company/  
stimulus\\_recovery.html](http://www.esri.com/company/stimulus_recovery.html)

# THANK YOU!

For more information:

*Rick Ayers, ESRI-Vienna*  
*DOT Account Manager*  
*RAyers@esri.com*