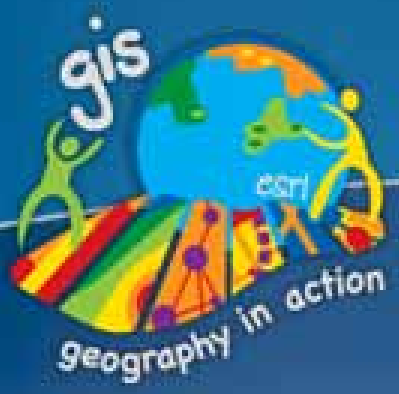


Please!
Turn OFF cell phones
and paging devices

Using Spatial Statistics

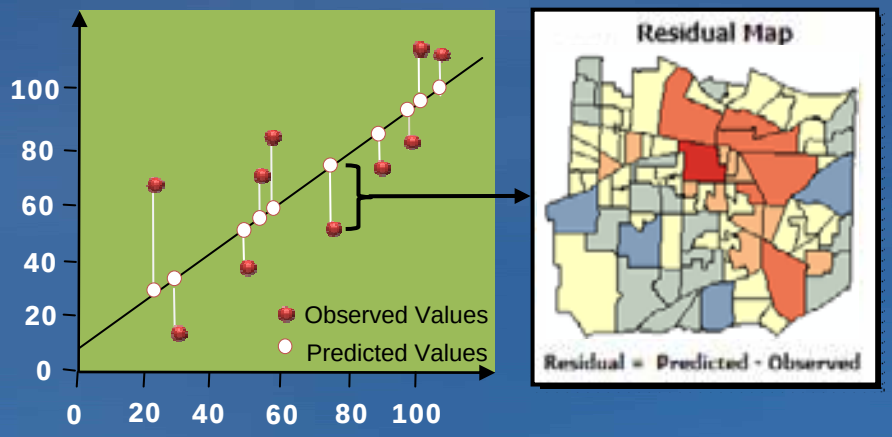
Social Service Applications
Public Safety and Public Health

Lauren Rosenshein

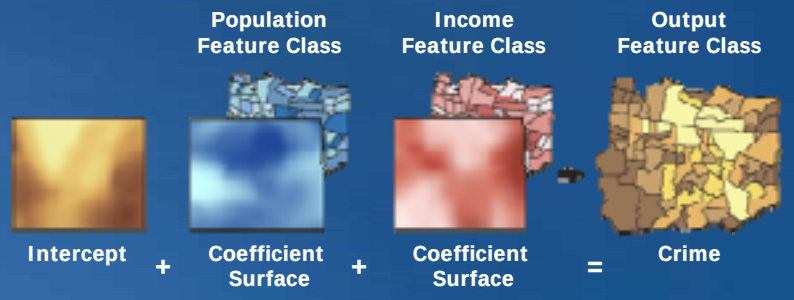
Regression analysis

- Regression analysis allows you to model, examine, and explore spatial relationships, in order to better understand the factors behind observed spatial patterns or to predict outcomes.

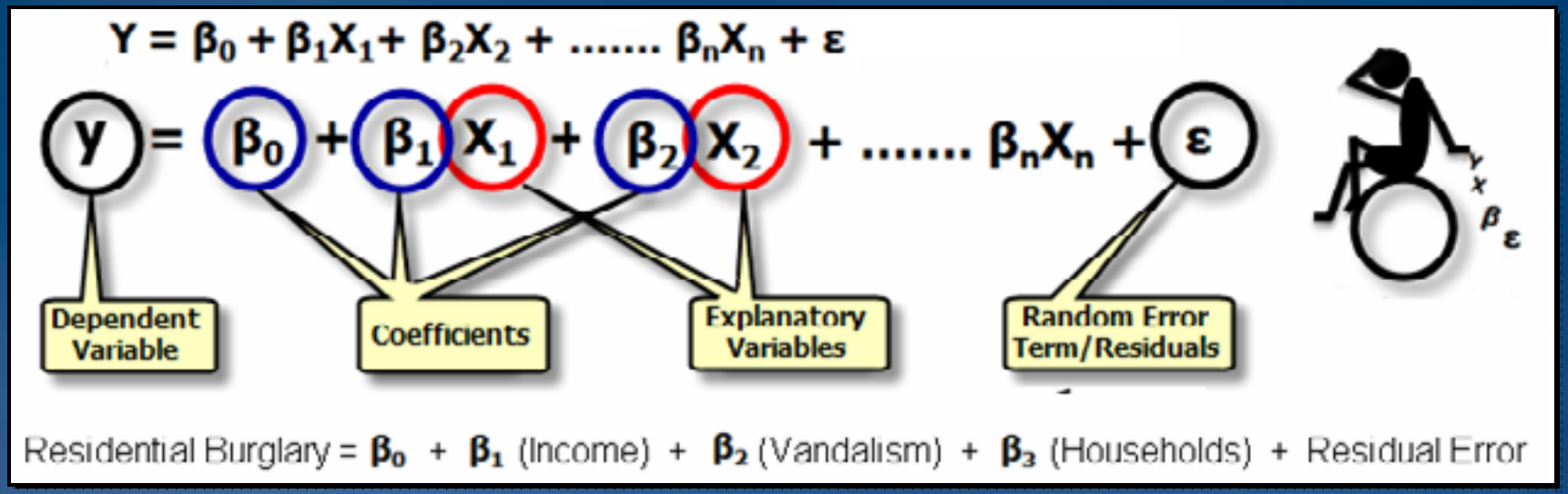
Ordinary Least Square



Geographically Weighted Regression



Regression analysis terms and concepts

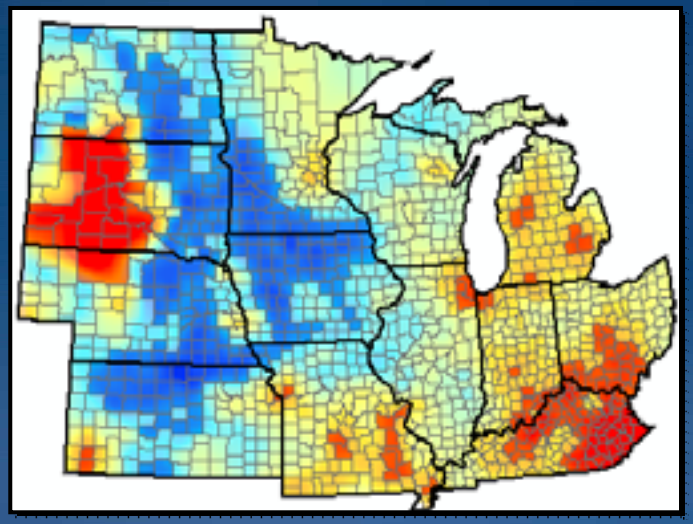


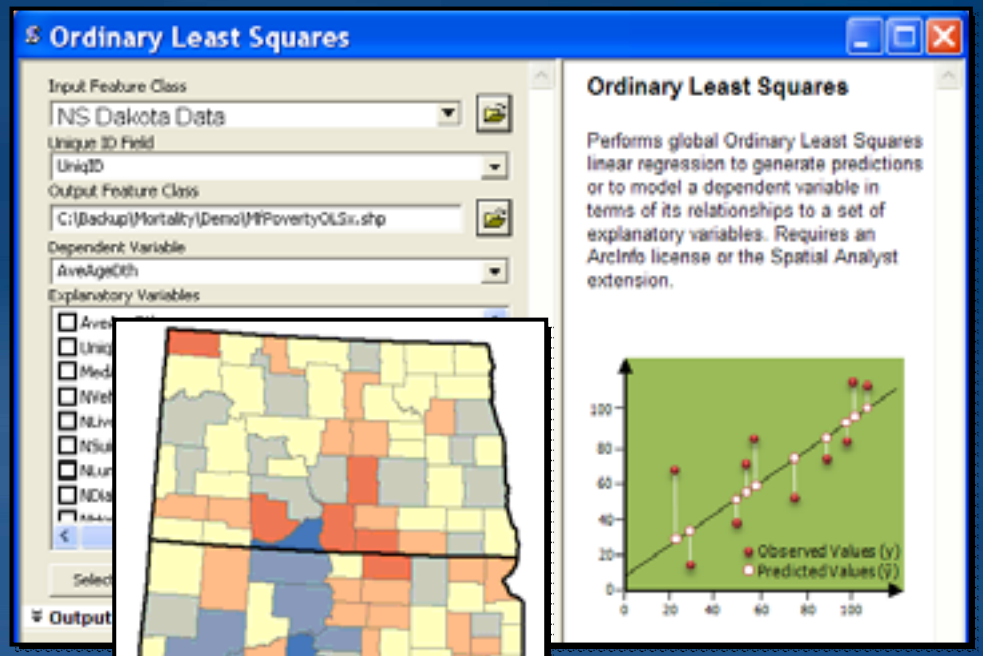
- **Dependent variable (Y):** what you are trying to model or predict (Residential Burglary, for example).
- **Explanatory variables (X):** variables you believe cause or explain the dependent variable (like: income, vandalism, households).
- **Coefficients (β):** values, computed by the regression tool, reflecting explanatory to dependent variable relationships.
- **Residuals (ε):** the portion of the dependent variable that isn't explained by the model; the model under and over predictions.

DEMO

Mortality Data Analysis

Use OLS to test hypotheses





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

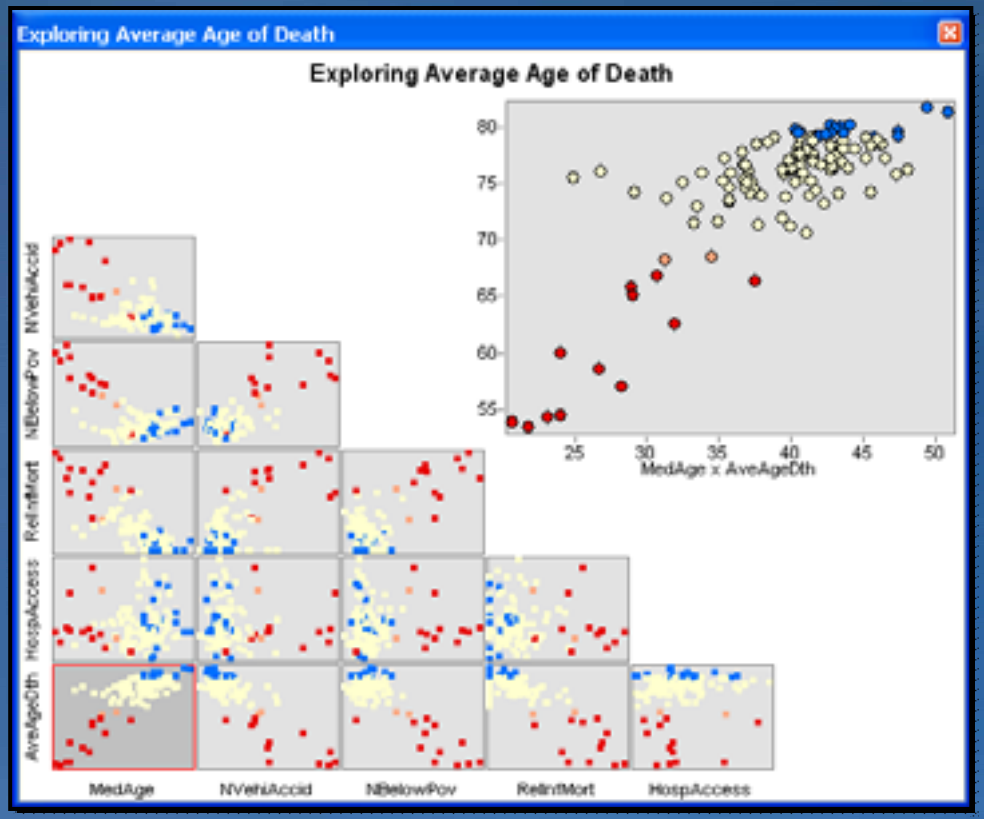
Poverty rates explain 66% of the variation in the average age of death

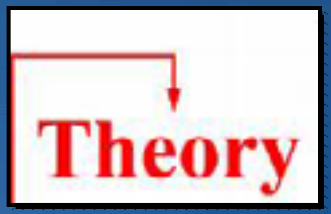
dependent variable: Adjusted R-Squared [2]: 0.659

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, often tedious, trial and error, process)





Interpreting OLS results

- **Use the notes on interpretation as a guide to understanding OLS model output.**

Summary of OLS Results

Variable	Coefficient	StdError	t-Statistic	Probability	Robust_SE	Robust_t	Robust_Pr	VIF [1]
Intercept	85.614389	0.921428	92.914876	0.000000*	0.899381	95.192600	0.000000*	--------
NVEHIACCID	-141.498860	9.974846	-14.185569	0.000000*	16.098865	-8.789368	0.000000*	1.389034
NSUICIDE	-163.173481	18.124402	-9.002972	0.000000*	31.296397	-5.213810	0.000001*	1.370010
NLUNGCANC	-42.888108	12.783969	-3.354835	0.001087*	15.026004	-2.854259	0.005126*	1.043282
NDIABETES	-55.951298	13.371354	-4.184415	0.000060*	17.842851	-3.135782	0.002186*	1.163340

OLS Diagnostics

Number of Observations:	119	Number of Variables:	5
Degrees of Freedom:	114	Akaike's Information Criterion (AIC) [2]:	538.4782
Multiple R-Squared [2]:	0.852540	Adjusted R-Squared [2]:	0.847366
Joint F-Statistic [3]:	164.772345	Prob(>F), (4,114) degrees of freedom:	0.000000*
Joint Wald Statistic [4]:	233.496820	Prob(>chi-squared), (4) degrees of freedom:	0.000000*
Koenker (BP) Statistic [5]:	41.367715	Prob(>chi-squared), (4) degrees of freedom:	0.000000*
Jarque-Bera Statistic [6]:	4.452889	Prob(>chi-squared), (2) degrees of freedom:	0.107911

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.

Coefficient significance

- Look for statistically significant explanatory variables.
- Consult the robust probabilities when the Koenker test is statistically significant

```
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*	0.000000*
0.001219*	0.005990*
0.000035*	0.001994*
0.079514	0.067555

Summary of OLS Results

Variable	Coefficient	StdError	t-Statistic	Probability	Robust_SE	Robust_t	Robust_Pr	VIF [1]
Intercept	85.374381	0.922955	92.501086	0.000000*	0.942427	90.589905	0.000000*	--------
NVEHIACCID	-140.428069	9.901408	-14.182636	0.000000*	15.617472	-8.991728	0.000000*	1.394241
NSUICIDE	-163.284992	17.957444	-9.092886	0.000000*	28.922397	-5.645625	0.000000*	1.370027
NLUNGCANC	-42.082857	12.674300	-3.320330	0.001219*	15.022877	-2.801252	0.005990*	1.044628
NDIABETES	-57.520555	13.277747	-4.332102	0.000035*	18.168554	-3.165940	0.001994*	1.168553
HOSPACCESS	0.276706	0.156376	1.769493	0.079514	0.149919	1.845711	0.067555	1.009638

OLS Diagnostics

Number of Observations:	119	Number of Variables:	6
Degrees of Freedom:	113	Akaike's Information Criterion (AIC) [2]:	537.2257
Multiple R-Squared [2]:	0.856515	Adjusted R-Squared [2]:	0.850167
Joint F-Statistic [3]:	134.908289	Prob(>F), (5,113) degrees of freedom:	0.000000*
Joint Wald Statistic [4]:	267.278746	Prob(>chi-squared), (5) degrees of freedom:	0.000000*
Koenker (BP) Statistic [5]:	38.994033	Prob(>chi-squared), (5) degrees of freedom:	0.000000*
Jarque-Bera Statistic [6]:	2.140918	Prob(>chi-squared), (2) degrees of freedom:	0.342851

Koenker(BP) Statistic [5]: 38.994033 Prob(>chi-squared),(5) degrees of freedom: 0.00000*

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.**
  - What did one regression coefficient say to the other regression coefficient?

    …I'm partial to you!

```
                    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.
```

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

VIF
2.351229
1.556498
1.051207
1.400358
3.232363

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]:	524.9762
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

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]: 524.976
Adjusted R-Squared [2]: 0.864823

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

                              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
```

Model significance

- The Joint F-Statistic and Joint Wald Statistic measure overall model significance.
- Consult the Joint Wald statistic when the Koenker test is statistically significant.

```
                    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.
```

Joint F-Statistic [3]:	151.985705 Prob(>F), (4,113) degrees of freedom: 0.000000*
Joint Wald Statistic [4]:	496.057428 Prob(>chi-sq), 5 degrees of freedom: 0.000000*
Koenker (BP) Statistic [5]:	21.590491 Prob(>chi-sq), 5 degrees of freedom: 0.000626*

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

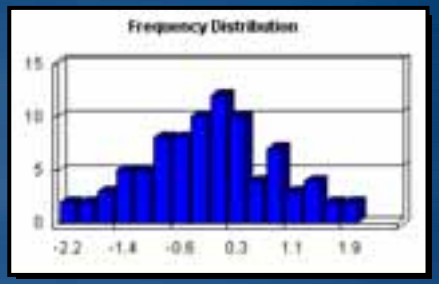
Number of Observations:	119	Number of Variables:	6
Degrees of Freedom:	113	Akaike's Information Criterion (AIC) [2]:	524.9762
Multiple R-Squared [2]:	0.870551	Adjusted R-Squared [2]:	0.864823
Joint F-Statistic [3]:	151.985705	Prob(>F), (5,113) degrees of freedom:	0.000000*
Joint Wald Statistic [4]:	496.057428	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

Model bias

```
                    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.

- **When the Jarque-Bera test is statistically significant:**
  - **the model is biased**
  - **results are *not* reliable**
  - **often this indicates that a key variable is missing from the model**



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

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

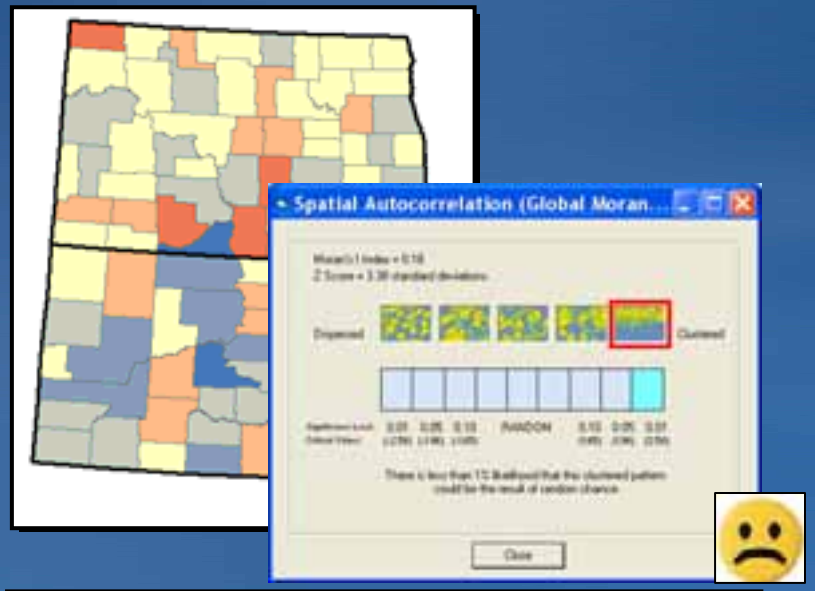
OLS Diagnostics

Number of Observations:	119	Number of Variables:	6
Degrees of Freedom:	113	Akaike's Information Criterion (AIC) [2]:	524.9762
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

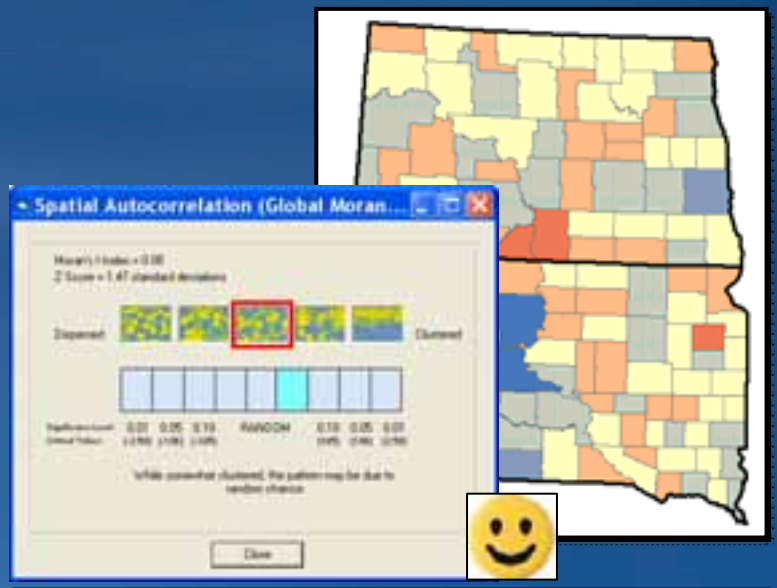
Spatial Autocorrelation

WARNING 000851: Use the Spatial Autocorrelation (Moran's I) Tool to ensure residuals are not spatially autocorrelated.

Error code:	000851: Use the Spatial Autocorrelation (Moran's I) Tool to ensure residuals are not spatially autocorrelated.
Description:	Results from regression analysis are only trustworthy when the model and data meet the assumptions/limitations of that method. Statistically significant spatial autocorrelation in the regression residuals indicates misspecification (a key missing explanatory variable). Results are invalid when a model is misspecified.
Solution:	Run the Spatial Autocorrelation (Moran's I) tool on the regression residuals in the output feature class. If the Z score indicates spatial autocorrelation is statistically significant, map the residuals and perhaps run hot spot analysis on the residuals to see if the spatial pattern of over and under predictions provides clues about missing key variables from the model. If you cannot identify the key missing variables, results of the regression are invalid and you should consider using a spatial regression method designed to deal with spatial autocorrelation in the error term. When spatial autocorrelation in OLS residuals is due to non-stationary spatial processes, use Geographically Weighted Regression instead of OLS.



Statistically significant clustering of under and over predictions.



Random spatial pattern of under and over predictions.

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

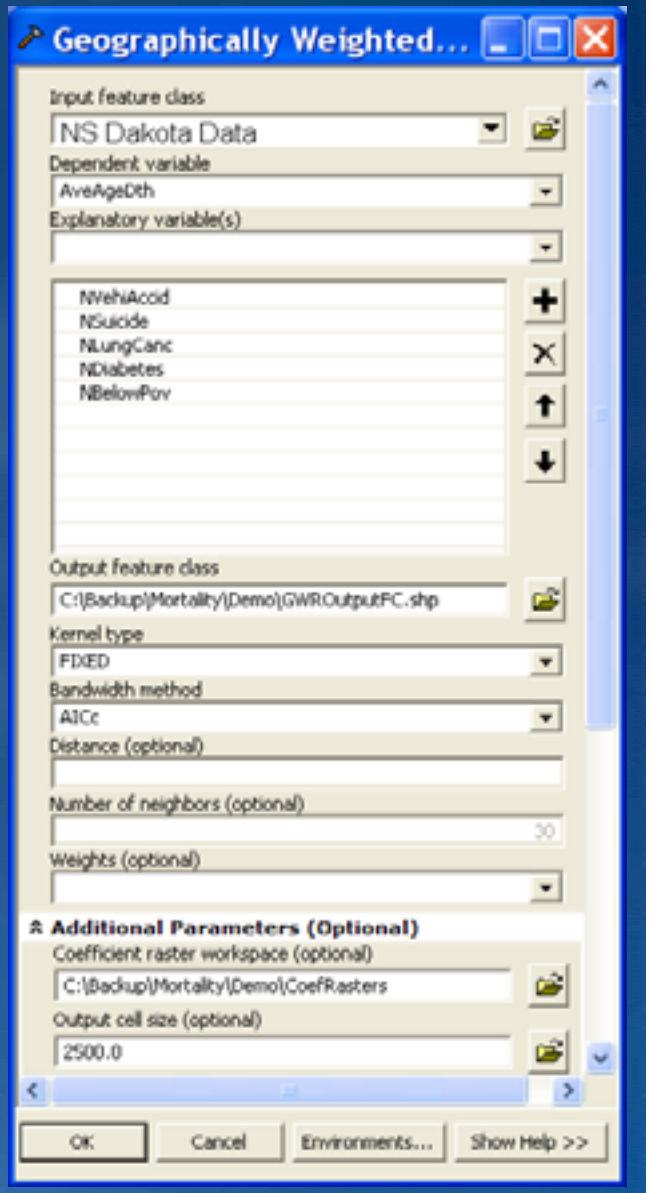
OLS Diagnostics

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. Relationships do not vary significantly across the study area. ✓

Run Geographically Weighted Regression (GWR)

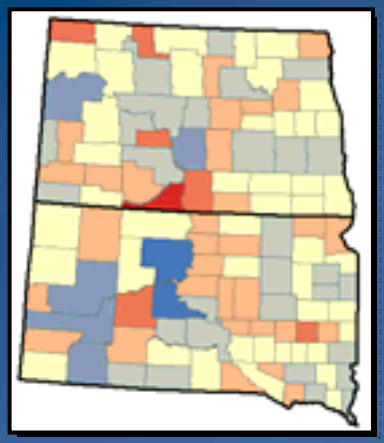
- GWR is a **local, spatial, regression** model
  - Global Regression methods, like OLS, break down when the strength of model relationships vary across the study area
- GWR variables are the same as OLS, except:
  - **Do not include spatial regime** (dummy) **variables**
  - **Do not include variables with little value variation**
- Selecting a bandwidth and kernel
  - Fixed or Adaptive
  - AIC, Cross Validation (CV), bandwidth parameter
    - Condition numbers



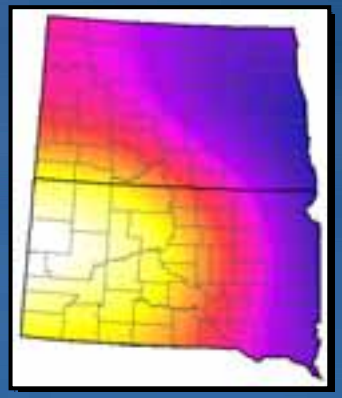
Interpreting GWR results

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

Compare GWR R2 and AIC values to OLS R2 and AIC values. The better model has a lower AIC and a high R2.



Residual maps show model under and over predictions. They shouldn't be clustered.



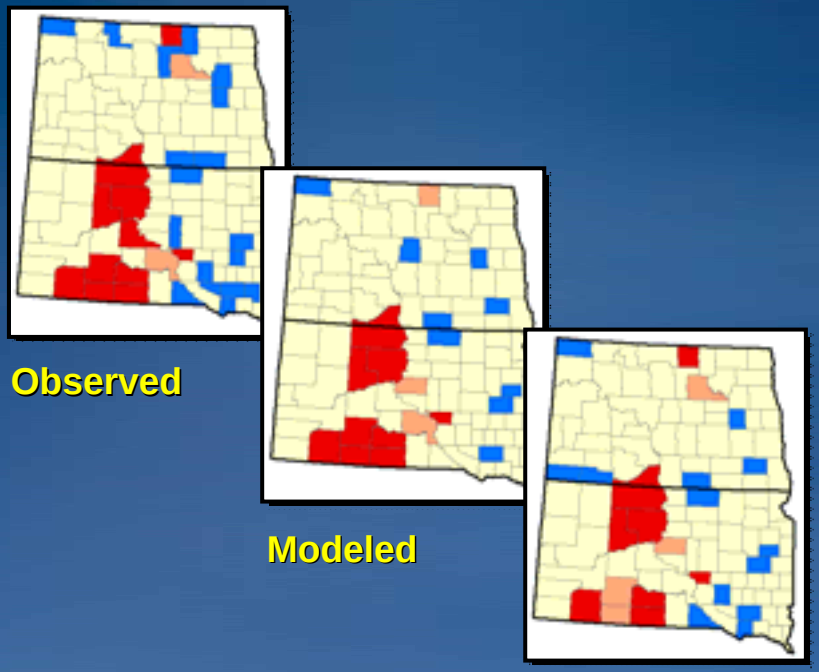
Coefficient maps show how modeled relationships vary across the study area.

	Observed	Cond	LocalR2	Predicted	Intercept	C1_NVehiAc	C2_NSuicid	C3_NLungCa
▶	78.419998	12.613701	0.881075	78.510341	85.562468	-79.980532	-148.284402	-37.899139
	76.5	14.048718	0.834124	77.920484	85.36851	-78.018965	-175.587799	-69.818161
	68.209999	12.25915	0.847111	68.964384	85.920939	-80.417789	-153.086076	-64.406605
	73.190002	12.515339	0.857244	72.182168	84.312503	-65.634913	-161.053375	-25.380057
	78.290001	11.758007	0.836626	77.979938	85.876829	-79.211314	-149.618999	-46.828649
	78.230003	12.612641	0.874848	79.159514	84.765217	-77.439124	-150.304914	-26.940229

Model predictions, residuals, standard errors, coefficients, and condition numbers are written to the output feature class.

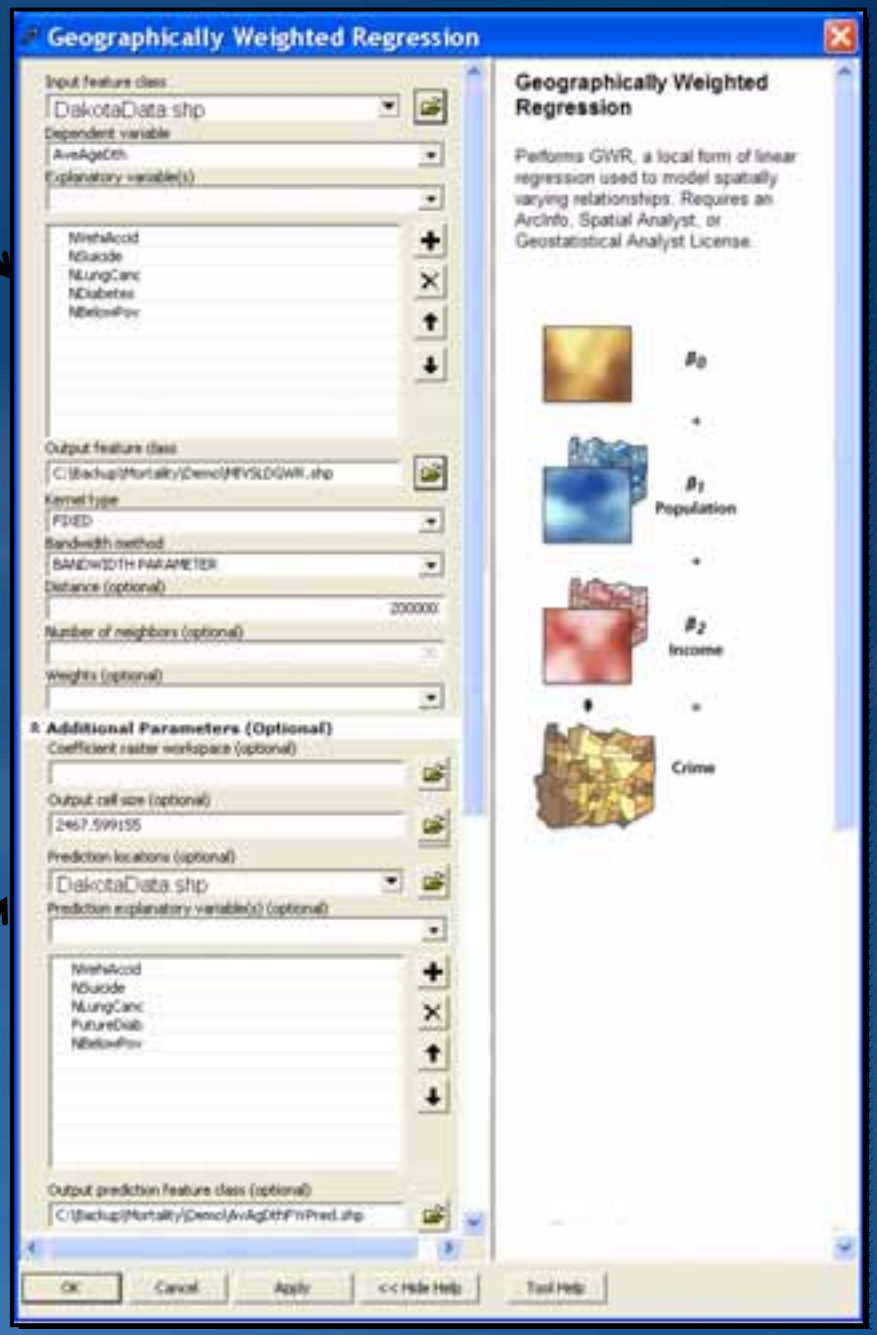
GWR prediction

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



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.



Resources for learning more…

- The ESRI Guide to GIS Analysis, Vol. 2
- Geographically Weighted Regression, by Fotheringham, Brundson, and Charlton
- 911 emergency call analysis demo:
  http://www.esri.com/software/arcgis/arcinfo/about/demos.html
- Virtual campus free web seminar
  http://campus.esri.com/
- Articles (keyword search: “Spatial Statistics”)
  http://www.esri.com/news/arcuser/0405/ss_crimestats1of2.html
- ArcGIS 9.3 Web Help:
  - Regression Analysis Basics
  - Interpreting OLS Results
  - Interpreting GWR Results

  Watch for updates
- GP Resource Center
- LScott@ESRI.com

QUESTIONS?

Lrosenshein@ESRI.com