Beyond Where: Modeling Spatial Relationships Using Regression Analysis

Lauren Bennett
Flora Vale
Ordinary Least Squares
Geographically Weighted Regression
demo
set the context
What is regression analysis?

- Model, examine, and explore spatial relationships
- Better understand factors behind observed spatial patterns
- Predict outcomes based on that understanding
a statistical process for estimating the relationships between variables
Why use regression analysis?

- Explore correlations
- Predict unknown values
- Understand key factors
Divorce Rate in Maine vs Per Capita Consumption of Margarine

Correlation: 0.992558
Number People Who Drowned by Falling into a Swimming-Pool vs Number of Nicolas Cage Films

Correlation: 0.666004
I used to think correlation implied causation.

Then I took a statistics class. Now I don't.

Sounds like the class helped.

Well, maybe.

http://xkcd.com/552/
Why use regression analysis?

- Explore correlations
- Predict unknown values
- Understand key factors
But, wait!
My data is messy…
Dealing with Missing or Suppressed Data

Real world data is messy!
Interpolating Missing Data for Polygons

- Missing data
- Suppressed data
- Cast your data to different geometry
- Design your geometry to fit your analysis

Areal Interpolation
People with Access to Potable Water in Kenya
Ready!
Back to business.
\[ y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n + \epsilon \]
Dependent Variable

\[ y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n + \epsilon \]
Dependent Variable

\[ y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n + \epsilon \]

What are you trying to predict or understand?
Explanatory Variables

\[ y = B_0 + B_1 X_1 + B_2 X_2 + \ldots + B_n X_n + \varepsilon \]
Explanatory Variables

\[ y = B_0 + B_1X_1 + B_2X_2 + \ldots + B_nX_n + e \]

Variables you believe to cause or explain the dependent variable
Coefficients

\[ y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n + \varepsilon \]
Represent the strength and type of relationship that $X$ has to $y$. 

$y = B_0 + B_1 X_1 + B_2 X_2 + \ldots + B_n X_n + \varepsilon$
Coefficients

\[ y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n + \epsilon \]

Positive relationship - as obesity rates rise, diabetes rates also rise
Coefficients

\[ y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n + \varepsilon \]

**Negative relationship** - as foreclosure rates rise, home values drop
No relationship - the value for X is not correlated with the value for y
Residual

\[ y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n + \epsilon \]
Residual

\[ y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n + \epsilon \]

Model over and under predictions
Residual

\[ y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n + \epsilon \]
y = β₀ + β₁X₁ + β₂X₂ + ... + βₙXₙ + ε

- predicted value
- observed value
Residual

\[ y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n + \epsilon \]

\( \{ \) over prediction
Residual

$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n + \varepsilon$

\[
\text{difference between the observed value and the predicted value} = \varepsilon
\]
Demo
building a model
Ordinary Least Squares Regression
Finding a Properly Specified Model
Finding a Properly Specified Model

the 6 checks

\[ y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n + \epsilon \]
$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n + \epsilon$$

Your coefficients should support your hypothesis
Each explanatory variable should have the relationship we expect

<table>
<thead>
<tr>
<th>Variable</th>
<th>Coefficient [a]</th>
<th>StdError</th>
<th>t-Statistic</th>
<th>Probability [b]</th>
<th>Robust_SE</th>
<th>Robust_t</th>
<th>Robust_Pr [b]</th>
<th>VIF [c]</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>9947.038596</td>
<td>44.461950</td>
<td>223.720250</td>
<td>0.000000*</td>
<td>43.232352</td>
<td>230.083219</td>
<td>0.000000*</td>
<td>-------</td>
</tr>
<tr>
<td>HOSPBEDSD</td>
<td>453.383027</td>
<td>94.410182</td>
<td>4.807088</td>
<td>0.000006*</td>
<td>95.160978</td>
<td>4.769161</td>
<td>0.000007*</td>
<td>1.789789</td>
</tr>
<tr>
<td>EVANDMAND</td>
<td>0.399674</td>
<td>0.036395</td>
<td>10.981628</td>
<td>0.000000*</td>
<td>0.041993</td>
<td>9.517526</td>
<td>0.000000*</td>
<td>3.476159</td>
</tr>
<tr>
<td>IMAGINGD</td>
<td>0.359023</td>
<td>0.165592</td>
<td>2.168123</td>
<td>0.032424*</td>
<td>0.161761</td>
<td>2.219469</td>
<td>0.028620*</td>
<td>3.044556</td>
</tr>
<tr>
<td>HOUSTOND</td>
<td>-0.001145</td>
<td>0.000115</td>
<td>-9.999029</td>
<td>0.000000*</td>
<td>0.000104</td>
<td>-10.976894</td>
<td>0.000000*</td>
<td>1.329077</td>
</tr>
<tr>
<td>PQI10D</td>
<td>1.041130</td>
<td>0.220645</td>
<td>4.718573</td>
<td>0.000009*</td>
<td>0.213298</td>
<td>4.881106</td>
<td>0.000005*</td>
<td>1.528101</td>
</tr>
</tbody>
</table>
Do the variables that should have a positive relationship have a positive coefficient?

<table>
<thead>
<tr>
<th>Variable</th>
<th>Coefficient [a]</th>
<th>StdError</th>
<th>t-Statistic</th>
<th>Probability [b]</th>
<th>Robust_SE</th>
<th>Robust_t</th>
<th>Robust_Pr [b]</th>
<th>VIF [c]</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>9947.038596</td>
<td>44.461950</td>
<td>223.720250</td>
<td>0.000000*</td>
<td>43.232352</td>
<td>230.083219</td>
<td>0.000000*</td>
<td>-------</td>
</tr>
<tr>
<td>HOSPBEDSD</td>
<td>453.838027</td>
<td>94.410182</td>
<td>4.807088</td>
<td>0.000006*</td>
<td>95.160978</td>
<td>4.769161</td>
<td>0.000007*</td>
<td>1.789789</td>
</tr>
<tr>
<td>EVANDMAND</td>
<td>0.399674</td>
<td>0.036395</td>
<td>10.981628</td>
<td>0.000000*</td>
<td>0.041993</td>
<td>9.517526</td>
<td>0.000000*</td>
<td>3.476159</td>
</tr>
<tr>
<td>IMAGINGD</td>
<td>0.359023</td>
<td>0.165592</td>
<td>2.168123</td>
<td>0.032424*</td>
<td>0.161761</td>
<td>2.219469</td>
<td>0.028620*</td>
<td>3.044556</td>
</tr>
<tr>
<td>HOUSTOND</td>
<td>-0.001145</td>
<td>0.000115</td>
<td>-9.999029</td>
<td>0.000000*</td>
<td>0.000104</td>
<td>-10.976894</td>
<td>0.000000*</td>
<td>1.329077</td>
</tr>
<tr>
<td>PQI10D</td>
<td>1.041130</td>
<td>0.220645</td>
<td>4.718573</td>
<td>0.000009*</td>
<td>0.213298</td>
<td>4.881106</td>
<td>0.000005*</td>
<td>1.528101</td>
</tr>
</tbody>
</table>
Check # 2

Make sure each explanatory variable is statistically significant

<table>
<thead>
<tr>
<th>Variable</th>
<th>Coefficient [a]</th>
<th>StdError</th>
<th>t-Statistic</th>
<th>Probability [b]</th>
<th>Robust SE</th>
<th>Robust t</th>
<th>Robust Pr [b]</th>
<th>VIF [c]</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>9947.038596</td>
<td>44.461950</td>
<td>223.720250</td>
<td>0.000000*</td>
<td>43.232352</td>
<td>230.083219</td>
<td>0.000000*</td>
<td>--------</td>
</tr>
<tr>
<td>HOSPBEDSD</td>
<td>453.838027</td>
<td>94.410182</td>
<td>4.807088</td>
<td>0.000006*</td>
<td>95.160978</td>
<td>4.769161</td>
<td>0.000007*</td>
<td>1.789789</td>
</tr>
<tr>
<td>EVANDMAND</td>
<td>0.399674</td>
<td>0.036395</td>
<td>10.981628</td>
<td>0.000000*</td>
<td>0.041993</td>
<td>9.517526</td>
<td>0.000000*</td>
<td>3.476159</td>
</tr>
<tr>
<td>IMAGINGD</td>
<td>0.359023</td>
<td>0.165592</td>
<td>2.168123</td>
<td>0.032424*</td>
<td>0.161761</td>
<td>2.219469</td>
<td>0.028620*</td>
<td>3.044556</td>
</tr>
<tr>
<td>HOUSTOND</td>
<td>-0.001145</td>
<td>0.000115</td>
<td>-9.999029</td>
<td>0.000000*</td>
<td>0.000104</td>
<td>-10.976894</td>
<td>0.000000*</td>
<td>1.329077</td>
</tr>
<tr>
<td>PQI10D</td>
<td>1.041130</td>
<td>0.220645</td>
<td>4.718573</td>
<td>0.000009*</td>
<td>0.213298</td>
<td>4.881106</td>
<td>0.000005*</td>
<td>1.528101</td>
</tr>
</tbody>
</table>
Does every explanatory variable have an asterisk?

<table>
<thead>
<tr>
<th>Variable</th>
<th>Coefficient [a]</th>
<th>StdError</th>
<th>t-Statistic</th>
<th>Probability [b]</th>
<th>Robust_SE</th>
<th>Robust_t</th>
<th>Robust_Pr [b]</th>
<th>VIF [c]</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>9947.038596</td>
<td>44.461950</td>
<td>223.720250</td>
<td>0.000000*</td>
<td>43.232352</td>
<td>230.083219</td>
<td>0.000000*</td>
<td>--------</td>
</tr>
<tr>
<td>HOSPBEDSD</td>
<td>453.838027</td>
<td>94.410182</td>
<td>4.807088</td>
<td>0.000006*</td>
<td>95.160978</td>
<td>4.769161</td>
<td>0.000007*</td>
<td>1.789789</td>
</tr>
<tr>
<td>EVANDMAND</td>
<td>0.399674</td>
<td>0.036395</td>
<td>10.981628</td>
<td>0.000000*</td>
<td>0.041993</td>
<td>9.517526</td>
<td>0.000000*</td>
<td>3.476159</td>
</tr>
<tr>
<td>IMAGINGD</td>
<td>0.359023</td>
<td>0.165592</td>
<td>2.168123</td>
<td>0.032424*</td>
<td>0.161761</td>
<td>2.219469</td>
<td>0.028620*</td>
<td>3.044556</td>
</tr>
<tr>
<td>HOUSTOND</td>
<td>-0.001145</td>
<td>0.000115</td>
<td>-9.999029</td>
<td>0.000000*</td>
<td>0.000104</td>
<td>-10.976894</td>
<td>0.000000*</td>
<td>1.329077</td>
</tr>
<tr>
<td>PQI10D</td>
<td>1.041130</td>
<td>0.220645</td>
<td>4.718573</td>
<td>0.0000009*</td>
<td>0.213298</td>
<td>4.881106</td>
<td>0.000005*</td>
<td>1.528101</td>
</tr>
</tbody>
</table>
Your residuals should not be clustered in location or in value.
Check # 3

Test the clustering of your residuals using the Spatial Autocorrelation tool
Check # 3

Are the residuals spatially random?

\[
y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n + \varepsilon
\]
Check # 4

Verify that residuals are normally distributed using the Jarque-Bera test

- Jarque-Bera Statistic \[ g \]: 1.591597
- Prob(>chi-squared), (2) degrees of freedom: 0.451221
Make sure the Jarque-Bera statistic does NOT have an asterisk!

<table>
<thead>
<tr>
<th>Input Features:</th>
<th>Hot Spot Study Area</th>
<th>Dependent Variable:</th>
<th>TOTAL_COSTS_2010</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of Observations:</td>
<td>110</td>
<td>Akaike’s Information Criterion (AICc) [d]:</td>
<td>1672.966945</td>
</tr>
<tr>
<td>Multiple R-Squared [d]:</td>
<td>0.870841</td>
<td>Adjusted R-Squared [d]:</td>
<td>0.864631</td>
</tr>
<tr>
<td>Joint F-Statistic [e]:</td>
<td>140.241872</td>
<td>Prob(&gt;F), (5,104) degrees of freedom:</td>
<td>0.000000*</td>
</tr>
<tr>
<td>Joint Wald Statistic [e]:</td>
<td>556.069919</td>
<td>Prob(&gt;chi-squared), (5) degrees of freedom:</td>
<td>0.000000*</td>
</tr>
<tr>
<td>Koenker (BP) Statistic [f]:</td>
<td>27.470483</td>
<td>Prob(&gt;chi-squared), (5) degrees of freedom:</td>
<td>0.0000046*</td>
</tr>
<tr>
<td>Jarque-Bera Statistic [g]:</td>
<td>1.591597</td>
<td>Prob(&gt;chi-squared), (2) degrees of freedom:</td>
<td>0.451221</td>
</tr>
</tbody>
</table>
Check # 5

Each variable should tell a different part of the story

<table>
<thead>
<tr>
<th>Variable</th>
<th>Coefficient [a]</th>
<th>StdError</th>
<th>t-Statistic</th>
<th>Probability [b]</th>
<th>Robust_SE</th>
<th>Robust_t</th>
<th>Robust_Pr [b]</th>
<th>VIF [c]</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>9947.038596</td>
<td>44.461950</td>
<td>223.720250</td>
<td>0.000000*</td>
<td>43.232352</td>
<td>230.083219</td>
<td>0.000000*</td>
<td>-------</td>
</tr>
<tr>
<td>HOSPBEDSD</td>
<td>453.838027</td>
<td>94.410182</td>
<td>4.807088</td>
<td>0.000006*</td>
<td>95.160978</td>
<td>4.769161</td>
<td>0.000007*</td>
<td>1.789789</td>
</tr>
<tr>
<td>EVANDMAND</td>
<td>0.399674</td>
<td>0.036395</td>
<td>10.981628</td>
<td>0.000000*</td>
<td>0.041993</td>
<td>9.517526</td>
<td>0.000000*</td>
<td>3.476159</td>
</tr>
<tr>
<td>IMAGINGD</td>
<td>0.359023</td>
<td>0.165592</td>
<td>2.168123</td>
<td>0.032424*</td>
<td>0.161761</td>
<td>2.219469</td>
<td>0.028620*</td>
<td>3.044556</td>
</tr>
<tr>
<td>HOUSTOND</td>
<td>-0.001145</td>
<td>0.000115</td>
<td>-9.999029</td>
<td>0.000000*</td>
<td>0.000104</td>
<td>-10.976894</td>
<td>0.000000*</td>
<td>1.329077</td>
</tr>
<tr>
<td>PQI10D</td>
<td>1.041130</td>
<td>0.220645</td>
<td>4.718573</td>
<td>0.000009*</td>
<td>0.213298</td>
<td>4.881106</td>
<td>0.000005*</td>
<td>1.528101</td>
</tr>
</tbody>
</table>
Check # ⁵  

Are all my VIF values lower than 7.5?

<table>
<thead>
<tr>
<th>Variable</th>
<th>Coefficient [a]</th>
<th>StdError</th>
<th>t-Statistic</th>
<th>Probability [b]</th>
<th>Robust_SE</th>
<th>Robust_t</th>
<th>Robust_Pr [b]</th>
<th>VIF [c]</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>9947.038596</td>
<td>44.461950</td>
<td>223.720250</td>
<td>0.000000*</td>
<td>43.232352</td>
<td>230.083219</td>
<td>0.000000*</td>
<td>--------</td>
</tr>
<tr>
<td>HOSPBEDSD</td>
<td>453.838027</td>
<td>94.410182</td>
<td>4.807088</td>
<td>0.000006*</td>
<td>95.160978</td>
<td>4.769161</td>
<td>0.000007*</td>
<td>1.789789</td>
</tr>
<tr>
<td>EVANDMAND</td>
<td>0.399674</td>
<td>0.036395</td>
<td>10.981628</td>
<td>0.000000*</td>
<td>0.041993</td>
<td>9.517526</td>
<td>0.000000*</td>
<td>3.476159</td>
</tr>
<tr>
<td>IMAGINGD</td>
<td>0.359023</td>
<td>0.165592</td>
<td>2.168123</td>
<td>0.032424*</td>
<td>0.161761</td>
<td>2.219469</td>
<td>0.028620*</td>
<td>3.044556</td>
</tr>
<tr>
<td>HOUSTOND</td>
<td>-0.001145</td>
<td>0.000115</td>
<td>-9.999029</td>
<td>0.000000*</td>
<td>0.000104</td>
<td>-10.976894</td>
<td>0.000000*</td>
<td>1.329077</td>
</tr>
<tr>
<td>PQI10D</td>
<td>1.041130</td>
<td>0.220645</td>
<td>4.718573</td>
<td>0.000009*</td>
<td>0.213298</td>
<td>4.881106</td>
<td>0.000005*</td>
<td>1.528101</td>
</tr>
</tbody>
</table>
## Evaluate model performance

<table>
<thead>
<tr>
<th>Input Features</th>
<th>Hot Spot Study Area</th>
<th>Dependent Variable: TOTAL_COSTS_2010</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of Observations:</td>
<td>110</td>
<td>1672.966945</td>
</tr>
<tr>
<td>Multiple R-Squared [d]:</td>
<td>0.870841</td>
<td></td>
</tr>
<tr>
<td>Joint F-Statistic [e]:</td>
<td>140.241872</td>
<td></td>
</tr>
<tr>
<td>Joint Wald Statistic [e]:</td>
<td>556.069919</td>
<td></td>
</tr>
<tr>
<td>Koenker (BP) Statistic [f]:</td>
<td>27.470483</td>
<td></td>
</tr>
<tr>
<td>Jarque-Bera Statistic [g]:</td>
<td>1.591597</td>
<td></td>
</tr>
</tbody>
</table>

### Regression Statistics

- Akaike's Information Criterion (AICc) [d]: 1672.966945
- Adjusted R-Squared [d]: 0.864631
  
- Prob(>F), (5,104) degrees of freedom: 0.000000*
- Prob(>chi-squared), (5) degrees of freedom: 0.000000*
- Prob(>chi-squared), (2) degrees of freedom: 0.451221
Check # 6  

Is the Adjusted R-Squared value strong enough?

<table>
<thead>
<tr>
<th>Input Features</th>
<th>Hot Spot Study Area</th>
<th>Dependent Variable</th>
<th>TOTAL_COSTS_2010</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of Observations:</td>
<td>110</td>
<td>Akaike's Information Criterion (AICc) [d]:</td>
<td>1672.966945</td>
</tr>
<tr>
<td>Multiple R-Squared [d]:</td>
<td>0.870841</td>
<td>Adjusted R-Squared [d]:</td>
<td><strong>0.864631</strong></td>
</tr>
<tr>
<td>Joint F-Statistic [e]:</td>
<td>140.241872</td>
<td>Prob(&gt;F), (5,104) degrees of freedom:</td>
<td>0.000000*</td>
</tr>
<tr>
<td>Joint Wald Statistic [e]:</td>
<td>556.069919</td>
<td>Prob(&gt;chi-squared), (5) degrees of freedom:</td>
<td>0.000000*</td>
</tr>
<tr>
<td>Koenker (BP) Statistic [f]:</td>
<td>27.470483</td>
<td>Prob(&gt;chi-squared), (5) degrees of freedom:</td>
<td>0.000046*</td>
</tr>
<tr>
<td>Jarque-Bera Statistic [g]:</td>
<td>1.591597</td>
<td>Prob(&gt;chi-squared), (2) degrees of freedom:</td>
<td>0.451221</td>
</tr>
</tbody>
</table>
Online help is ... helpful!

Six checks

- Coefficients have expected sign
- Coefficients are statistically significant
- No redundancy among explanatory variables
- Residuals are normally distributed
- Residuals are not spatially autocorrelated
- Strong Adjusted R2 (good model performance)
Exploratory Regression

<table>
<thead>
<tr>
<th>Tradeoffs</th>
</tr>
</thead>
<tbody>
<tr>
<td>Learn a lot about your data and relationships among variables</td>
</tr>
<tr>
<td></td>
</tr>
</tbody>
</table>

Variables

- Cultural
- Environmental
- Socio Economic
- Lifestyle
- Spatial

Tests all variable combinations for:

1. Redundancy
2. Completeness
3. Significance
4. Bias
5. Performance

Creates Output Diagnostic Report
Demo

exploratory regression
Exploring Spatial Variation

Geographically Weighted Regression
## Global vs. local regression models

<table>
<thead>
<tr>
<th>OLS</th>
<th>GWR</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Global regression model</td>
<td>• Local regression model</td>
</tr>
</tbody>
</table>
| • One equation  
  • Calibrated using data from all features | • One equation for each feature  
  • Calibrated using data from nearby features |
| • Relationships are fixed | • Relationships can vary across the study area |
each feature gets a separate equation
\[ y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n + \varepsilon \]
Demo modeling geographic variation

Geographically Weighted Regression
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
Defining *local*

- 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
questions?
Please fill out a course survey!!
Online – www.esri.com/ucsessionssurveys
Paper – pick up and put in drop box
Offering ID: 1450

Want to learn more???

Wednesday:
10:15a Applying Spatial Statistics: The Analysis Process in Action
1:30p Spatial Statistics: Simple Ways to Do More with Your Data
3:15p Spatial Data Mining: A Deep Dive into Cluster Analysis

Thursday:
10:15a Beyond Where: Modeling Spatial Relationships Using Regression Analysis
1:30p Applying Spatial Statistics: The Analysis Process in Action

lbennett@esri.com
fvale@esri.com
jdacosta@esri.com
esriurl.com/spatialstats