Beyond Where: Modeling Spatial Relationships and Making Predictions

Lauren Bennett
Jenora D’Acosta
Alberto Nieto
Flora Vale

esriurl.com/spatialstats
Modeling Relationships

when variables are related, information can be learned about one variable by observing the values of the related variable(s)
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/
Generalized Linear Regression
Modeling linear relationships
a statistical process for estimating linear relationships between variables
\[ y = B_0 + B_1 x_1 + B_2 x_2 + \ldots + B_n x_n + e \]
Dependent Variable

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

What are you trying to predict or understand?
Explanatory Variables

\[ y = B_0 + B_1x_1 + B_2x_2 + \ldots + B_nx_n + \epsilon \]

Variables you believe to cause or explain the dependent variable
Coefficients

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

Represent the strength and type of relationship that \( X \) has to \( y \)
Coefficients

\[ y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_n X_n + \varepsilon \]

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

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

Negative relationship - as foreclosure rates rise, home values drop
Coefficients

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

No relationship - the value for X is not correlated with the value for y
Residual

\[ y = B_0 + B_1 x_1 + B_2 x_2 + \ldots + B_n x_n + \epsilon \]

Model over and under predictions
Residual

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

predicted value

observed value
Residual

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

over prediction
Residual

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

\{ difference between the observed value and the predicted value = \epsilon \}
Evaluating our model
Every variable should be 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>
Each variable should tell a different part of the story

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

<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>
Residuals should not be clustered in location or in value
Residuals should not be clustered in location or in value

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

<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>[Akaike's Information Criterion (AICc)] [d]: 1672.966945</td>
</tr>
<tr>
<td>Multiple R-Squared [d]:</td>
<td>0.870841</td>
<td>[Adjusted R-Squared] [d]: 0.864631</td>
</tr>
<tr>
<td>Joint F-Statistic [e]:</td>
<td>140.241872</td>
<td>Prob(&gt;F), (5,104) degrees of freedom: 0.000000*</td>
</tr>
<tr>
<td>Joint Wald Statistic [e]:</td>
<td>556.069919</td>
<td>Prob(&gt;chi-squared), (5) degrees of freedom: 0.000000*</td>
</tr>
<tr>
<td>Koenker (BP) Statistic [f]:</td>
<td>27.470483</td>
<td>Prob(&gt;chi-squared), (5) degrees of freedom: 0.000046*</td>
</tr>
<tr>
<td>Jarque-Bera Statistic [g]:</td>
<td>1.591597</td>
<td>Prob(&gt;chi-squared), (2) degrees of freedom: 0.451221</td>
</tr>
</tbody>
</table>
Model should have a strong R-Squared

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

<table>
<thead>
<tr>
<th>Dependent Variable:</th>
<th>TOTAL_COSTS_2010</th>
</tr>
</thead>
<tbody>
<tr>
<td>Akaike's Information Criterion (AICc) [d]:</td>
<td>1672.966945</td>
</tr>
<tr>
<td>Adjusted R-Squared [d]:</td>
<td>0.864631</td>
</tr>
<tr>
<td>Prob(&gt;F), (5,104) degrees of freedom:</td>
<td>0.000000*</td>
</tr>
<tr>
<td>Prob(&gt;chi-squared), (5) degrees of freedom:</td>
<td>0.000000*</td>
</tr>
<tr>
<td>Prob(&gt;chi-squared), (5) degrees of freedom:</td>
<td>0.000046*</td>
</tr>
<tr>
<td>Prob(&gt;chi-squared), (2) degrees of freedom:</td>
<td>0.451221</td>
</tr>
</tbody>
</table>
When comparing models, lower AICc is better

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

<table>
<thead>
<tr>
<th>Dependent Variable:</th>
<th>TOTAL_COSTS_2010</th>
</tr>
</thead>
<tbody>
<tr>
<td>Akaike's Information Criterion (AICc) [d]:</td>
<td>1672.966945</td>
</tr>
<tr>
<td>Adjusted R-Squared [d]:</td>
<td>0.864631</td>
</tr>
<tr>
<td>Prob(&gt;F), (5,104) degrees of freedom:</td>
<td>0.000000*</td>
</tr>
<tr>
<td>Prob(&gt;chi-squared), (5) degrees of freedom:</td>
<td>0.000000*</td>
</tr>
<tr>
<td>Prob(&gt;chi-squared), (5) degrees of freedom:</td>
<td>0.000046*</td>
</tr>
<tr>
<td>Prob(&gt;chi-squared), (2) degrees of freedom:</td>
<td>0.451221</td>
</tr>
</tbody>
</table>
Exploratory Regression

Tests all variable combinations for:

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

Creates Output Diagnostic Report
Geographically Weighted Regression

Exploring spatial variation
each feature gets a separate equation
\[ y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n + \epsilon \]
Defining local
Defining *local*

**Neighborhood Type**

- Number of neighbors
Defining *local*

**Neighborhood Type**
- Number of neighbors
Defining local

Neighborhood Type

- Number of neighbors
Defining local

Neighborhood Type

- Number of neighbors
- Distance band
Defining local

Neighborhood Type

- Number of neighbors
- Distance band
Defining local

Neighborhood Type

- Number of neighbors
- Distance band
Defining *local*

**Neighborhood Type**
- Number of neighbors
- Distance band

**Neighborhood Selection Method**
- Golden search
- Manual intervals
- User defined
Demo
Three model types

• Gaussian – continuous
• Logistic – binary
• Poisson – count
Gaussian – model a continuous variable

- Sales profits
- Healthcare spending
- Mortality rate
- Temperature
Poisson – model a **count** variable

- Traffic accidents
- Number of people with cancer per 10,000
- Crime counts
- Sales per month

Bar chart with counts for each category from 0 to 5.
Logistic – model a **binary** variable

- Disease presence
- Fire damage
- Insurance fraud
- Pass/fail inspection
Demo
Local Bivariate Relationships

Examining relationships across space
“... the measurement of a relationship depends on where the measurement is taken.”
two variables

determine relationship significance and type

relationships across geography
... what does it mean for two variables to be related to each other?
“good” relationships
“good” relationships

null relationships
“good” relationships

low entropy

null relationships

higher entropy
variable Y

low entropy
variable Y

high entropy
variable Y

mutual information

variable X
minimum spanning trees

low entropy

higher entropy
...how do we know if the relationship is **SIGNIFICANT**???
Permutation-based distribution estimation
feature
each feature has two values

- dependent
- explanatory
relationship between $Y$ and $X$ is evaluated for a neighborhood.
relationship between $Y$ and $X$ is evaluated for a neighborhood.
neighborhood → distribution of variables → minimum spanning tree → entropy value

0.56
what is the probability that the observed entropy exists while $Y$ and $X$ are actually independent from each other?
observed local entropy

0.56
start a permutation
the dependent values $Y$ are shuffled, while explanatory values $X$ are kept the same
minimum spanning trees and entropy are calculated for the permutation
minimum spanning trees and entropy are calculated for the permutation
neighborhood

permutations

0.56

0.61

0.57

0.64

...
observed entropy is converted to a p-value

observed entropy = 0.56
p-value = 0.0012
Entropy helps determine if relationships are significant, but it does not tell us what type of relationship exists.
classifying the local relationships
start with the distribution of the significant feature’s neighborhood
estimate an ordinary linear regression model and calculate AICc
estimate a second linear regression model using the square of the explanatory variable
AICc values are compared, and a model is selected

AICc: -42.3
linear model

AICc: -51.4
quadratic model

square of X
is the $R^2$ value greater than 0.05?

- Yes
- No

which model was chosen?

- linear
- quadratic

sign of coefficient?

- positive
- negative

undefined complex

Convex

Concave
entropy

classify type
Positive Linear
Convex
Demo
"Essentially, all models are wrong, but some are useful."

- George E. P. Box