Logistic Regression for Distribution Modeling

GIS5306 GIS Applications in Environmental Systems

Presented by:

Andrea Palmiotto
John Perry
Theory
Familiar Territory

• Linear Regression

\[ Y_i = \beta_0 + B_1 X_i + \epsilon_i \]

\[ \epsilon \sim N(0, \sigma^2) \]

• Relevant Assumptions
  – The expectation is a linear function of inputs
  – Y is continuous
  – Errors are normally distributed
y = 1.5761x + 1.0227
Familiar Territory
Motivation

• Dichotomous (Binomial, Binary) response

\[ Y_i \in \{0,1\} \quad \text{pdf}(Y_i) = \pi_i^{Y_i}(1 - \pi_i)^{1-Y_i} \]

• Residuals tend to have a logistic distribution

\[ F(x) = \frac{1}{1 + e^{-(x-\mu)/s}} \]

• In expectation the binary variable is the probability of occurrence (Bernoulli)

\[ E[Y_i] = \pi_i \]
Motivation

Linear Regression Gone Wrong
• Constrain probability using the logistic function

\[ \pi_i = \frac{1}{1 + e^{-z_i}} \]

• Make it dependent on a linear combination of independent variables

\[ z_i = \beta_0 + \beta_1 X_{i,1} + \cdots + \beta_m X_{i,(m-1)} \]
Model

- The logit

$$\text{logit}(\pi_i) = \ln\left(\frac{\pi_i}{1 - \pi_i}\right) = \beta_0 + \beta_1 X_{i,1} + \cdots + \beta_m X_{i,(m-1)}$$

- Clear relationship to the odds ratio

$$\frac{\pi_i}{1 - \pi_i}$$

- The “log odds” is a linear function of the input variables
Model

“Probability of... $Y$”

“Given that... $X$”
Solution

• From basic probability

\[ P(A \text{ and } B) = P(A)P(B) \]

• Probability of \( n \) trials with \( m \) different factors

\[ f(Y_1, ..., Y_n | \beta_0, ..., \beta_m, X_{1,1}, ..., X_{n,m-1}) = \prod_i \pi_i^{Y_i} (1 - \pi_i)^{1-Y_i} \]

• The maximum likelihood principle yields

\[ \mathcal{L}(\beta_0, ..., \beta_m | X_{1,1}, ..., X_{n,m-1}, Y_1, ..., Y_n) = \prod_i \pi_i^{Y_i} (1 - \pi_i)^{1-Y_i} \]
We calculate the maximum likelihood estimate of the logistic regression parameters using

$$\hat{\beta}_{ML} = \max_\beta \left( \prod_i \pi_i^{y_i} (1 - \pi_i)^{1-y_i} \right)$$

This is a relatively complex calculation (compared to linear regression), particularly for multivariate regression.

It is a numeric optimization problem, so there are a number of computational techniques that can be used.
Rudimentary Test Statistic

- Used to test the significance of the effect of individual parameters on the outcome

\[ H_0: \beta_{\text{true}} = \hat{\beta}_{\text{null}} \]
\[ H_1: \beta_{\text{true}} = \hat{\beta}_{\text{alt}} \]

\[
D = -2 \ln \left( \frac{L_{\text{null}}(\beta|X,Y)}{L_{\text{alt}}(\beta|X,Y)} \right)
\]

\[ D \sim \chi^2_{\text{df}_{\text{alt}}-\text{df}_{\text{null}}} \]
The assumptions inherent in linear regressions do not work in the real world (i.e., Y cannot continue forever – there are physical limitations that restrict continuous growth)

By introducing simple binomial Y/N queries, the regression formula is altered to better account for real world variation. These binomial variables include:

- Absence vs. presence; survival vs. death; failure vs. success, etc.

Binary variables in a logistic regression also guarantee that your probability rate never exceeds 100% (which can happen when using linear regressions)

- Results are limited between 0 and 1 (as percentages, not binary values)
- No longer a simple linear line, but results now in a logistic curve
- Accounts for errors better than linear regressions

The Maximum Likelihood Principle helps determine what are the most likely factors to affect the subject, and are used to draw the logistic regression curve.

In the equations, $\beta$ refers to how much X influences Y

- Influential factors are weighted to indicate degree of influence
Example
• Underlying binary variable
  – “Is (-82°42′36″, 27°11′03″) a suitable nesting site for the gopher tortoise?”

• Prediction is not binary, it is a probability
  – “The probability of the site located at (-82°42′36″, 27°11′03″) of being a suitable nesting site is 73%”

• Since the underlying variable is binary, it is “best” modeled by the logistic function of a linear combination of predictive factors

• The logistic model is fit to the data by maximum likelihood estimation
  – logistic regression
Gopher Tortoises

- Located throughout the southeastern United States in areas with well-drained sandy soil and vegetation cover including:
  - Pine flatwoods, mixed hardwood-pine communities, and dry prairies
- Dig burrows several meters long that provide protection from extreme warm or cold temperatures, fire, and predation
- Considered a keystone species because their burrows provide habitats for hundreds of other animals and turning the soil returns leached nutrients to the surface.
- Endangered due to habitat loss from urbanization, and predation

http://www.flickriver.com/photos/michaelrupert/tags/reptile/
Factors

• Example Factors
  – Prefer Dry Ground
  – Prefer Near Water

• “Habitat Modeling Within a Regional Context: An Example Using Gopher Tortoise”
  Baskaran, et al.

[Cooperative Conservation America
http://www.cooperative conservation.org/]
Big Picture

- Start with mapped factors (layers)
- Use Logistic Regression to fit a model using a training set of observations
- Predict the probability of occurrence over the rest of the map

\[
\text{Probability of a gopher tortoise burrow} = \frac{\exp(A)}{1 + \exp(A)}, \quad (1)
\]

\[
A = \left(\begin{array}{c}
(\text{Dist2strms} \times 0.004) - (\text{Dist2rds} \times 0.003) - (\% \text{ clay} \times 0.152) \\
+ (\text{Transportation} \times 1.751) + (\text{Utilityswaths} \times 2.327) \\
+ (\text{Clearcut} \times 2.684) + (\text{Decid} \times 1.913) + (\text{Evergreen} \times 1.004) \\
+ (\text{Mixed} \times 1.8) + (\text{Pasture} \times 3.987) + (\text{Rowcrop} \times 2.435) - 0.757
\end{array}\right). \quad (2)
\]

[Baskaran, et al.]
Types of Factors - Ordinal

- Prefer Being Near Water?
  - “Transition Zone”
  - “Diminishing Returns”
### “Categorical” vs “Ordered”

<table>
<thead>
<tr>
<th>Ordered</th>
<th>Categorical</th>
</tr>
</thead>
<tbody>
<tr>
<td>6</td>
<td>Excessively Drained</td>
</tr>
<tr>
<td>5</td>
<td>Well Drained</td>
</tr>
<tr>
<td>4</td>
<td>Somewhat Drained</td>
</tr>
<tr>
<td>3</td>
<td>Somewhat Poorly Drained</td>
</tr>
<tr>
<td>2</td>
<td>Poorly Drained</td>
</tr>
<tr>
<td>1</td>
<td>Very Poorly Drained</td>
</tr>
</tbody>
</table>

#### Graph

- **Ordered**
  - Excessively Drained
  - Well Drained
  - Somewhat Drained
  - Somewhat Poorly Drained
  - Poorly Drained
  - Very Poorly Drained

- **Categorical**

#### Probability

- Logistic Model
- “Actual”
“Categorical” or “Ordered”

“Binary Factors with a Binary Response degenerates to a contingency table”
Data Set

- FGDL.org
- Polk, Highland, and Glades Counties, FL
- Response
  - Gopher Tortoise Observations (238)
    - 80% Training
    - 20% Validation
  - Pseudoabsences (238)
- Factors
  - Elevation
  - Elevation Slope
  - Soil Drainage
  - Distance to Streams
  - Distance to Roads
  - Distance to Permanently Flooded Wetlands
  - Distance to Seasonally Flooded Wetlands
  - Population Density
  - Public/Private Land
  - Land Cover
    - Urban
    - Pasture
    - Scrub
References


