Logic Regression and its Applications in SNP Association Studies

Ingo Ruczinski

Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health

Email: ingo@jhu.edu. The slides and software used for this presentation are available at http://biostat.jhsph.edu/~iruczins

A Public Health Related Example

- The Women's Health and Aging Study (WHAS) began in 1992 to study the causes and the course of disability in moderately to severely disabled older women living in the community.

- The WHAS is a population-based longitudinal study of women with at least mild disability, 65 years of age or older, living at home in eastern Baltimore city or county.

- 1002 women agreed to participate and provided written informed consent.

- The major chronic diseases at baseline were ascertained by using complex algorithms. Follow-up evaluations were conducted every 6 months for 3 years.

- There is evidence that disability results from chronic diseases, and that interactions between diseases (comorbidities) are of importance in causing disability.

- The chronic diseases recorder included cancer, congestion heart failure, diabetes, degenerative disc disease, hip fracture, myocardial infarction, arthritis, osteoporosis, Parkinson’s disease, pulmonary disease, stroke.
A Public Health Related Example

\[ p = \Pr(\text{death in round } j \mid \text{survival to round } j-1, X, \text{age}) \]

\[ \logit(p) = -9.01 + 0.06 \cdot \text{age} + 1.07 \cdot L(X) \]

Motivation

[Lucek and Ott]

"Current methods for analyzing complex traits include analyzing and localizing disease loci one at a time. However, complex traits can be caused by the interaction of many loci, each with varying effect."

"... patterns of interactions between several loci, for example, disease phenotype caused by locus A and locus B, or A but not B, or A and (B or C), clearly make identification of the involved loci more difficult. While the simultaneous analysis of every single two-way pair of markers can be feasible, it becomes overwhelmingly computationally burdensome to analyze all 3-way, 4-way to N-way 'and' patterns, 'or' patterns, and combinations of loci."
Logic Regression

- $X_1, \ldots, X_k$ are 0/1 (False/True) predictors.
- $Y$ is a response variable.
- Fit a model
  
  $$g(E(Y)) = b_0 + \sum_{j=1}^{t} b_j \cdot L_j,$$

  where $L_j$ is a Boolean combination of the covariates, e.g., $L_j = (X_1 \lor X_2) \land X_4^c$.
- Determine the logic terms $L_j$ and estimate the $b_j$ simultaneously.

- SNPs are usually coded as dominant and recessive:

<table>
<thead>
<tr>
<th>SNP</th>
<th>X</th>
<th>X.R</th>
<th>X.D</th>
</tr>
</thead>
<tbody>
<tr>
<td>AA</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>AT</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>TT</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

Logic Trees

An equivalent representation of $\text{SNP2.D}^c \lor (\text{SNP17.R} \land \text{SNP22.D})$:

This is a Logic Tree!
Decision Trees and Logic Trees

A Decision Tree is different from a Logic Tree!

An Example
Example: GAW 12

\[
\logit(\text{affected}) = \beta_0 + \beta_1 \times X_{\text{ENV}_1} + \beta_2 \times X_{\text{ENV}_2} + \beta_3 \times X_{\text{Gender}} + \sum_{i=1}^{K} \beta_{i+3} \times L_i
\]

The Move Set for Logic Regression
Simulated Annealing for Logic Regression

We try to fit the model \( g(E(Y)) = b_0 + \sum_{j=1}^{t} b_j \cdot L_j \).

- Select a scoring function (RSS, log-likelihood, \ldots).
- Pick the maximum number of Logic Trees.
- Pick the maximum number of leaves in a tree.
- Initialize the model with \( L_j = 0 \) for all \( j \).
- Carry out the Simulated Annealing Algorithm:
  - Propose a move.
  - Accept or reject the move, depending on the scores and the temperature.

Growing Logic Models

```
X.7
X.7 X.2
or
X.7 X.2
and X.4
and
X.4 X.2 X.7 X.1
or and
or
X.1 X.7
and so on...
```
A Public Health Related Example

Model Selection 1: Cross Validation
Model Selection 2: Permutation Tests

```
<p>| | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
```

Model Selection 2: Permutation Tests

```
X
```

```
Y
```

```
\pi(Y)
```

permutation
Model Selection 2: Permutation Tests

\[
\begin{array}{c|cc|c}
X & T & Y & \pi(Y) \\
\hline
0 & . & . & . \\
1 & . & . & . \\
2 & . & . & . \\
3 & . & . & . \\
4 & . & . & . \\
5 & . & . & . \\
6 & . & . & . \\
\end{array}
\]

permutation
permutation
Model Selection 2: Permutation Tests

Multiple Models
Multiple Models

Let $\gamma_S$ be the score of a certain state $S$.

- We use the acceptance function
  $$\alpha(\gamma_{\text{old}}, \gamma_{\text{new}}, t) = \min\{1, \exp([\gamma_{\text{old}} - \gamma_{\text{new}}]/t)\}$$

- If we keep the temperature constant, this defines a homogeneous Markov chain.

- We constructed the move set to be irreducible and aperiodic, therefore each homogeneous Markov chain has a limiting distribution $\pi_t(S)$.

- If we know the model size where the signal ends and the noise starts, we can read off the corresponding temperature from the diagnostic plot!

Multiple Models

Example: Simulate 10 binary predictors $X_1, \ldots, X_{10}$.

Let $Y = 5 + 1 \times L(X_1, X_2, X_3, X_4) + \epsilon$, $\epsilon \sim N(0,1)$.

Run a homogeneous Markov chain during “crunch time” for two separate cases:

Case 1: All $X$ are independent.

Case 2: All $X$ are independent, except $X_4$ (in the signal) and $X_5$ (not in the signal), which are heavily correlated.
Multiple Models

SNPs

predictors