Lecture 27: Introduction to Correlated Binary Data

Dipankar Bandyopadhyay, Ph.D.

BMTRY 711: Analysis of Categorical Data Spring 2011
Division of Biostatistics and Epidemiology
Medical University of South Carolina
Introduction

- Previously, we discussed correlated data in the context of a matched pair design.
- Correlated data are common in:
  1. Longitudinal studies: same subject followed over time.
  2. Cluster randomized trials: Treatment is assigned to groups, the members within a group tend to be correlated responses.
  3. Family studies: individuals within a family are more similar (genetically and environmentally).
- While we could have spent a semester on this topic alone, we will briefly move through a methods-driven examination.
Motivating Data

- We will look at the “Six Cities” study of the health effects of air pollution (Ware et al. 1984).
- The data analyzed are the 16 selected cases in Lipsitz, Fitzmaurice, et al. (1994).
- The binary response is the wheezing status of 16 children at ages 9, 10, 11, and 12 years.
- The probability of wheezing at each age is to be modeled as a logistic regression model using the explanatory variables city of residence, age, and maternal smoking status at the particular age.
General summary points to consider

- Data between subjects are assumed to be independent
- Data within a subject are assumed to be dependent
- The dependence is modeled as a covariance (or correlation) pattern
Covariance Patterns

You will learn more about covariance patterns models in Multivariate Analysis, for now, consider the following structures:

- **Compound Symmetry (or exchangeable):** Correlation is the same for all outcomes within a subject (i.e., the \( \text{corr}(y_{ij}, y_{ik}) = \rho, \forall j \neq k \) and the \( \text{corr}(y_{ij}, y_{i'k}) = 0, \forall i \neq i' \) )

- **Unstructured** (i.e., \( \text{corr}(y_{ij}, y_{ik}) = \rho_{jk} \))

The compound symmetry model is a good place to begin.

- You estimate the fewest number of correlation parameters
- Compound symmetry is often used in sample size calculations
- **re:example** - The binary responses (there are 4 of these, one for each age) for individual children are assumed to be equally correlated, implying an exchangeable correlation structure.
Generalized Estimating Equations (GEE)

- GEE is a generalized form of a GLM
- GEE differs from a GLM in that the distribution of the outcome is not completely specified
- GEE is known as a marginal model. A marginal model is appropriate when inference on group effects (population effects) is of interest. Group effects may include a "treatment" effect.
- Solutions are obtained by the estimating equations (AKA as score equations), which for exponential class variables, can be written as

\[
S(\beta) = \sum_{i=1}^{n} \frac{\partial E(Y_i|X_i)}{\partial \beta} \left[ \frac{Y_i - E(Y_i|X_i)}{Var(Y_i|X_i)} \right]
\]

The estimation of the variance-covariance matrix is more complicated, and the solutions are obtained iteratively. For the purpose of this class, lets rely on GENMOD for the calculations.
Before we develop our model, let's formalize some more notation.

Notation

- $i = 1, \ldots, N$ subjects
- $j = 1, \ldots, t_i$ observations (I like $t$ to represent time, others may use $n_i$)
- $Y_i = [y_{i1}, y_{i2}, \ldots, y_{it_i}]'$ is the $t_i \times 1$ vector of responses for subject $i$
  - $y_{i1}$ is the response for subject $i$ at time 1
  - $y_{i2}$ is the response for subject $i$ at time 2
  - etc. (recall $t_i$ is the maximum observed time for subject $i$ – this does not have to be the same for all subjects)

For our example, $t_i = 4 \quad \forall i$
Covariate notation

- \( x_{ij} = [x_{ij1}, x_{ij2}, \ldots, x_{ijp}]' \) is the \( p \times 1 \) covariate vector for the subject \( i \) at time \( j \)
- \( X_i = [x_{i1}, x_{i2}, \ldots, x_{it_i}]' \) is the \( t_i \times p \) matrix of covariates for subject \( i \)
- \( \beta \) is the \( p \times 1 \) vector of true population parameters

NOTES:
Covariates are typically static (same at all time measurements, e.g., gender, race, etc) or time-dependent (change over time, e.g., drug dose, smoking status, etc.)
This notation accounts for the characteristics.
Our data

data six;
  input case city$ @@;  \(<---\) Time indepdendent covariates
  do i=1 to 4;  \(<---\) "time"
    input age smoke wheeze @@;  \(<---\) time dependent
  output;
end;
datalines;
  1 portage  9 0 1 10 0 1 11 0 1 12 0 0
  2 kingston 9 1 1 10 2 1 11 2 0 12 2 0
  3 kingston 9 0 1 10 0 0 11 1 0 12 1 0
  4 portage  9 0 0 10 0 1 11 0 1 12 1 0
  5 kingston 9 0 0 10 1 0 11 1 0 12 1 0
  6 portage  9 0 0 10 1 0 11 1 0 12 1 0
  7 kingston 9 1 0 10 1 0 11 0 0 12 0 0
  8 portage  9 1 0 10 1 0 11 1 0 12 2 0
  9 portage  9 2 1 10 2 0 11 1 0 12 1 0
 10 kingston 9 0 0 10 0 0 11 0 0 12 1 0
 11 kingston 9 1 1 10 0 0 11 0 1 12 0 1
 12 portage  9 1 0 10 0 0 11 0 0 12 0 0
 13 kingston 9 1 0 10 0 1 11 1 1 12 1 1
 14 portage  9 1 0 10 2 0 11 1 0 12 2 1
 15 kingston 9 1 0 10 1 0 11 1 0 12 2 1
 16 portage  9 1 1 10 1 1 11 2 0 12 1 0
run;
Ignoring Clustering

- Here, we have 4 observations per individual
- What happens if we assume we have 64 independent observations? (4 outcomes per 16 people)
- Here is the code:

```plaintext
proc genmod data=six desc;
   class case city ;
   model wheeze = city age smoke / dist=bin;
run;
```
Selected Results

Model Information

Data Set WORK.SIX
Distribution Binomial
Link Function Logit
Dependent Variable wheeze

Number of Observations Read 64
Number of Observations Used 64
Number of Events 19
Number of Trials 64

Analysis Of Parameter Estimates

<table>
<thead>
<tr>
<th>Parameter</th>
<th>DF</th>
<th>Estimate</th>
<th>Error</th>
<th>Pr &gt; ChiSq</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>1</td>
<td>1.2597</td>
<td>2.6104</td>
<td>0.6294</td>
</tr>
<tr>
<td>city kingston</td>
<td>1</td>
<td>0.1391</td>
<td>0.5527</td>
<td>0.8013</td>
</tr>
<tr>
<td>city portage</td>
<td>0</td>
<td>0.0000</td>
<td>0.0000</td>
<td>.</td>
</tr>
<tr>
<td>age</td>
<td>1</td>
<td>-0.2003</td>
<td>0.2508</td>
<td>0.4245</td>
</tr>
<tr>
<td>smoke</td>
<td>1</td>
<td>-0.1284</td>
<td>0.4102</td>
<td>0.7544</td>
</tr>
</tbody>
</table>
Repeated Statement

- We need to tell SAS that we have correlated data (or repeated observations)
- We do this by using the repeated statement

```plaintext
proc genmod data=six desc;
  class case city ;
  model wheeze = city age smoke / dist=bin;
  repeated subject=case / type=exch;
run;
```
Selected Results

You get the same ‘‘model based’’ information
Data Set WORK.SIX
Distribution Binomial
Link Function Logit
Dependent Variable wheeze

Number of Observations Read 64
Number of Observations Used 64
Number of Events 19
Number of Trials 64

So...you have to go to the end of the report for the GEE summary
### Selected Results

#### GEE Model Information

- **Correlation Structure**: Exchangeable
- **Subject Effect**: case (16 levels)
- **Number of Clusters**: 16
- **Correlation Matrix Dimension**: 4
- **Maximum Cluster Size**: 4
- **Minimum Cluster Size**: 4

#### Analysis Of GEE Parameter Estimates

**Empirical Standard Error Estimates**

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Estimate</th>
<th>Standard Error</th>
<th>95% Confidence Limits</th>
<th>Z</th>
<th>Pr &gt;</th>
<th>Z</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>1.2751</td>
<td>3.0561</td>
<td>-4.7148</td>
<td>7.2650</td>
<td>0.42</td>
<td>0.6765</td>
<td></td>
</tr>
<tr>
<td>city kingston</td>
<td>0.1223</td>
<td>0.6882</td>
<td>-1.2266</td>
<td>1.4713</td>
<td>0.18</td>
<td>0.8589</td>
<td></td>
</tr>
<tr>
<td>city portage</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
<td>.</td>
<td>.</td>
<td></td>
</tr>
<tr>
<td>age</td>
<td>-0.2036</td>
<td>0.2789</td>
<td>-0.7502</td>
<td>0.3431</td>
<td>-0.73</td>
<td>0.4655</td>
<td></td>
</tr>
<tr>
<td>smoke</td>
<td>-0.0935</td>
<td>0.3613</td>
<td>-0.8016</td>
<td>0.6145</td>
<td>-0.26</td>
<td>0.7957</td>
<td></td>
</tr>
</tbody>
</table>
### Comparison of Estimates

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Regular MLE (64 indep obs)</th>
<th>GEE (16 clusters of 4)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Estimate</td>
<td>Error</td>
</tr>
<tr>
<td>Intercept</td>
<td>1.2597</td>
<td>2.6104</td>
</tr>
<tr>
<td>city kingston</td>
<td>0.1391</td>
<td>0.5527</td>
</tr>
<tr>
<td>city portage</td>
<td>0.0000</td>
<td>0.0000</td>
</tr>
<tr>
<td>age</td>
<td>-0.2003</td>
<td>0.2508</td>
</tr>
<tr>
<td>smoke</td>
<td>-0.1284</td>
<td>0.4102</td>
</tr>
</tbody>
</table>

Conclusion: Parameter estimates approximately equal; standard errors wrong under regular MLE

However, for this example, the effect isn’t that dramatic (See Kleinbaum & Klein Ch 11 for more dramatic example)
Properties of GEE

- GEE estimates have desirable asymptotic properties
- For correctly specified models and
- As the number of clusters gets large, the estimates are
  1. **Consistent**: $\hat{\beta} \to \beta$ as $K \to \infty$
  2. **Asymptotically normal**: $\hat{\beta} \sim \text{normal}$ as $K \to \infty$
- Correctly specified means the correct link and correlation have been specified
- However, GEE is robust to misspecification of the correlation pattern
- The closer the correlation is to the true correlation, the more efficient (smaller standard errors)
Model Testing

- Gone are the likelihood based methods
- We have not specified our likelihood and are using quasi-likelihood
- Recall from the GLM slides, we formulated the score equations
- With GEE, we are using “score like” equations since we have not fully specified the likelihood (we’ve only specified the variance (based on the binomial) and the correlation of the outcomes
- We can compute “score like” and Wald tests
Other forms of correlated data

- Correlated data also arises in survey research
- This follows a cluster sampling approach
- Randomly select clusters (e.g., families) and survey family members
- Need to take this clustering into account in the analysis
- You can use SUDAAN, GEE or new SAS procedures (SURVEYFREQ, SURVEYLOGISTIC, etc)