

# PubH 8452 Spring 2008

## Homework #3

Due 14 Apr 2008

1. For multivariate binary data there are several possible parameterizations of the model that have been proposed for regression analysis of the correlated outcomes. *All* models for multivariate categorical data are special cases of the canonical log-linear model. The difference between models is the way that the joint probabilities are structured in terms of the parameters of interest.
  - (a) *Log-linear model* (DHLZ 8.2.1) Consider a data set with four binary measurements taken on each of the subject. Define the saturated log-linear model. Give an interpretation for the parameters  $\theta_{i1,j}$  and  $\theta_{i2,jk}$  for a given  $j = 1, 2, 3, 4$  and  $k \neq j$  (where  $i$  indexes the subject).
  - (b) What is  $E(Y_{ij})$  in terms of the log-linear model parameters?
  - (c) *Bahadur model* DHLZ 8.2.2 presents a “fully marginal” model for binary data known as the Bahadur model. For the data above write down the likelihood function based on the Bahadur model with three-way and four-way correlations assumed to be zero. Given an interpretation for the parameters  $\mu_{ij}$  and  $\rho_{ijk}$  for a given  $j$  and  $k \neq j$ .
  - (d) Consider a single pair of binary response variables  $(Y_{ij}, Y_{ik})$ . Derive the correlation  $\rho_{ijk} = \text{Cor}(Y_{ij}, Y_{ik})$  such that it is a function of  $E(Y_{ij} | Y_{ik} = 1) - E(Y_{ij} | Y_{ik} = 0)$ . What's  $\rho_{ijk}$  when  $E(Y_{ij}) = E(Y_{ik})$  (hint:  $\text{Var}(Y_{ij}) = E(Y_{ij})(1 - E(Y_{ij}))$ )? Give a simple interpretation for  $\rho_{ijk}$  in this case.
  - (e) Consider an exchangeable correlation matrix for binary  $Y_{ij}, j = 1, \dots, n_i$  with a single scalar predictor  $X_{ij} = X_i$ . Please derive the score function of GEE and discuss its relationship to quasi-likelihood using a beta-binomial variance model for  $(\sum_j Y_{ij}, n_i)$ . (Note: see DHLZ pages 60-61 for  $R^{-1}$  [the notation for  $R$  in DHLZ is  $V_0$ ]).
  - (f) *Partially marginal model*: Fitzmaurice and Laird (1993) considered the QEF model

$$\Pr(\mathbf{Y}_i) = \exp \left\{ \theta_{i0} + \sum_{j=1}^{n_i} \theta_{ij} Y_{ij} + \sum_{j < k} \omega_{ijk} Y_{ij} Y_{jk} \right\},$$

and the transformation  $(\boldsymbol{\theta}, \boldsymbol{\omega}) \rightarrow (\boldsymbol{\mu}, \boldsymbol{\omega})$ . Derive the score equations (of  $(\boldsymbol{\beta}, \boldsymbol{\alpha})$ ) for

the regressions

$$\begin{aligned}g(\mu_{ij}) &= \mathbf{X}_{ij}^T \boldsymbol{\beta}, \\h(\omega_{ijk}) &= \mathbf{Z}_{ijk}^T \boldsymbol{\alpha}.\end{aligned}$$

Hint: See Fitzmaurice and Laird (1993).

- (g) Comment on the relationship between the score function for  $\boldsymbol{\beta}$  under the partly marginalized model in (f) and that from GEE. Comment on the consistency of  $\hat{\boldsymbol{\beta}}$  if  $\boldsymbol{\alpha}$  is modeled wrong. (No proof is needed.)
2. In this exercise we will consider a mechanism for overdispersed Poisson data and evaluate the resulted impact on standard errors. Consider the following hierarchical, or mixture model:

$$\begin{aligned}Y_i | z_i &\sim \text{Poisson}(\mu_i z_i) \\ \gamma z_i = \nu_i &\sim \text{Gamma}(1, \gamma)\end{aligned}$$

where  $\gamma$  is the shape parameter. Let  $\alpha = 1/\gamma$ , we have

$$\begin{aligned}\mathbb{E}(\nu_i) &= \gamma & \text{Var}(\nu_i) &= \gamma, \\ \mathbb{E}(Y_i) &= \mu_i & \text{Var}(Y_i) &= \mu_i + \alpha \mu_i^2.\end{aligned}$$

For a set of predictors  $\mathbf{X}_i$ , the marginal mean of  $Y_i$  is modeled via

$$\log(\mu_i) = \mathbf{X}_i^T \boldsymbol{\beta}.$$

- (a) If we assume  $\text{Var}(Y_i) = \mu_i$  and use Poisson regression with log link, write down the score equations for Poisson regression. What is the variance of the estimator,  $\hat{\boldsymbol{\beta}}$ , that solves the score equations above? Hint: See the *Generalized Estimating Equations* handout on *Impact of Model Misspecification*. The variance should have a sandwich form.
- (b) What is the variance of the Quasi-Likelihood estimator if the variance of  $Y_i$  is given as  $\text{Var}(Y_i) = \phi \mu_i$ ? Note: you will need estimate  $\phi$  – use simple moment estimate.
- (c) What is the variance of the Quasi-Likelihood estimator if the variance of  $Y_i$  is given as  $\text{Var}(Y_i) = \mu_i + \alpha \mu_i^2$ ? Note: you will need estimate  $\alpha$  – use simple moment estimate.
- (d) Simulate over-dispersed data using two covariates: a continuous  $X_1 \sim \mathcal{N}(0, 1)$ , the binary  $X_2 = 0$  for half the subjects and  $X_2 = 1$  for the second half. The sample size is  $n = 200$ . The regression coefficients (including intercept) are  $\boldsymbol{\beta} = (2, 1.0, 0.5)$ , and  $\alpha = 0.5$ . Sample R code is given below:

```
n <- 200
x1 <- rnorm (n, mean = 0, sd = 1)
```

```

x2 <- rep (0:1, each = n/2)
mu <- exp (2 + 1 * x1 + 0.5 * x2)

alpha <- 0.5
gamma <- 1 / alpha
nu <- rgamma (n, shape = gamma, scale = 1)

```

```
y <- rpois (n, lambda = mu * nu)
```

Please use seed=1 before you simulate the data, for all parts except for (g).

```

seed <- 1
set.seed(seed, kind=NULL)

```

Fit an ordinary Poisson regression model and construct residual plot(s) that can be used to guide whether the variance from in (a) or (b) is suggested by the residuals.

- (e) Write your own program to get the variance estimates given in (a), (b), and (c). Compare the resulted standard error estimates to those obtained from the ordinary Poisson regression. Hint: you can check if your program for (b) is correct by comparing the result with

```
glm(y~x1+x2, family=quasi(link=log, variance="mu"))
```

- (f) Use R function, `gee` to get the robust standard error estimates (i.e. `link=log`, `family=poisson`). Compare the robust standard error estimates to those obtained in (e).
- (g) Simulate 100 data sets and obtain  $\hat{\beta}$  and the 4 standard error estimates (from (a), (b), (c), and (f)). Compare the standard errors estimates and evaluate whether they yield nominal coverage for 95% confidence intervals.

Please use seed=number of simulation in your program, say

```

for (i in 1:100)
{
  seed <- i
  set.seed(seed, kind=NULL)
  ...
}

```

3. In a clinical trial of patients with respiratory illness, 111 patients from two different clinics were randomized to receive either placebo or an active treatment. Patients were examined at baseline and at four visits during treatment. At each examination, respiratory status (categorized as 1 = good, 0 = poor) was determined. The main objective of the analysis is to understand the joint effects of treatment and time on the probability that respiratory status is classified as good. It is also of interest to determine whether the effect of treatment is the same for patients from two clinics.

- (a) Ignoring the clinic variable, consider a model for the log odds that respiratory status is classified as good, including the main effects of treatment and time (where time is regarded as a categorical variable with 5 levels), the their interaction.

Using generalized estimating equations (GEE), assuming separate pairwise log-odds ratios (or separate pairwise correlations, if available software does not permit the within-subject association to be parameterized in terms of log odds ratios) among the five binary responses. Construct a test of the null hypothesis of no effect of treatment on *changes* in the log odds that respiratory status is classified as good based on the empirical standard errors.

- (b) What conclusions do you draw about the effect of treatment on changes in the log odds? Provide results that support your conclusions.
- (c) Patients in this trial were drawn from two separate clinics. Repeat the analysis to allow the effects of treatment (and possibly, time) to depend upon clinic.
  1. Is the effect of treatment the same in the two clinics? Present results to support your conclusion.
  2. Find a parsimonious model that describes the effects of clinic, treatment and time, on the log odds that respiratory status is classified as good. For the selected final model, give a clear interpretation of the estimated regression parameters .
- (d) For the final model selected above, construct a table of the estimated probabilities that respiratory status is classified as good as a function of both time and treatment group (and possibly, clinic). What do you conclude from this table?