

PubH 8452 Spring 2008

Homework #4

Due 16 May 2008 (5:00PM)

1. Consider the design used in the Progabide data. For $i = 1, \dots, m$, we observe independent count processes specified as follows

$$Y_{ij} | w_{ij} \sim \text{Poisson}(\lambda_{ij}) \quad (1)$$

$$\log \lambda_{ij} = \mathbf{x}_{ij}^T \boldsymbol{\beta} + w_{ij} \quad (2)$$

where the random effect w_{ij} is a stationary first-order autoregressive process, i.e.,

$$w_{ij} - \mu_w = \psi(w_{ij-1} - \mu_w) + Z_{ij}, \quad (3)$$

$$Z_{ij} \stackrel{\text{iid}}{\sim} \mathcal{N}(0, \sigma_Z^2). \quad (4)$$

For identifiability, let $E(e^{w_{ij}}) = 1$. The predictors in the models are intercept, treatment, time and treatment-time interaction (Example 9.5 of DHLZ).

- (a) Find μ_w and $\text{Var}(w_{ij}) = \sigma_w^2$ in terms of ψ and σ_Z^2 .
- (b) Assume that the standard GLM methods are used (that is, serial dependence is ignored in deriving the likelihood). Determine the formula for the asymptotic covariance matrix for the GLM estimates $\tilde{\beta}_0$ and $\tilde{\beta}_1$ and show that this is no smaller (in the ordering of n.n.d. matrices) than the asymptotic covariance that would be calculated assuming that $\sigma_Z^2 = 0$ (i.e. no latent process w_{ij}).
- (c) Discuss the result of part (b) in relation to the parallel result for ordinary least squares for correlated Gaussian data.
- (d) Use Theorems 1 and 2 in (Liang & Zeger, 1986) and compare the standard errors of the GLMM estimate $\tilde{\beta}_3$ and GEE estimate $\hat{\beta}_3$ (based on the GEE with the correct covariance structure) versus ψ using graphs.
- (e) Is it possible to specify the correct covariance structure using a matrix representation of the form?

$$\tilde{V}_i(\boldsymbol{\beta}, \boldsymbol{\alpha}, \phi) = \phi A^{1/2} R(\boldsymbol{\alpha}) A^{1/2}.$$

2. The Skin Cancer Prevention Study was a randomized, double-blind, placebo-controlled clinical trial of beta carotene to prevent non-melanoma skin cancer in high-risk subjects. A total of 1805 subjects were randomized to either placebo or 50mg of beta-carotene per day

for 5 years. Subjects were examined once a year and biopsied if a cancer was suspected to determine the number of new skin cancers occurring since the last exam. The main objective of the analyses is to compare the effect of beta carotene on skin cancer rates.

The raw data is in file `skin.dat`. The outcome variable (Y) is a count of the number of new skin cancers per year. The categorical variable “Treatment” is coded 1 = beta-carotene, 0 = placebo. The variable “Year” denotes the year of follow-up. The categorical variable “Gender” is coded 1 = male, 0 = female. The categorical variable “Skin” denotes skin type and is coded 1 = burns, 0=otherwise. The variable “Exposure” is a count of the number of previous skin cancers. The variable “Age” is the age (in years) of each subject at randomization. Complete data are available on 1683 subjects comprising a total of 7081 measurements.

- (a) Consider a Poisson-Gaussian generalized linear mixed model, with random intercepts, for the subject-specific log rate of skin cancers:

$$\log E(Y_{ij} | b_i) = (\beta_0 + b_i) + \beta_1 \text{Year}_{ij} + \beta_2 \text{Treatment}_i + \beta_3 \text{Treatment}_i \times \text{Year}_{ij}.$$

Fit this model using both penalized quasi-likelihood and maximum likelihood methods. Try different number of points in Gauss-Hermite quadrature method for integration. Comment on their relative computing time (including PQL) and estimates. part What is the estimated of the variance of the random effects σ_b^2 ? Given an interpretation to the magnitude of the estimated variance.

- (b) What is the interpretation of the estimate of β_1 ?
- (c) What is the interpretation of the estimate of β_3 ?
- (d) From these results, what conclusions do you draw about the effect of beta carotene on skin cancers and why?
- (e) Obtain the predicted (empirical Bayes) random effects for each subject. What does its distribution look like?
- (f) Calculate the sample variance of the predicted random effects. How does it compare with the estimate of σ_b^2 ? Why might they differ?
- (g) Plot the predictions against age and the count of the number of previous skin cancers. What do you conclude?
- (h) Carry out a marginal model analysis using GEE, using the same covariates as above. What do you conclude about the effect of beta carotene on skin cancers and why? Comment on the difference (or lack of) with the random effect model.
- (i) Repeat the random effect model analysis adjusting for skin type, age and the count of the number of previous skin cancers. What conclusions do you draw about effect of beta carotene on skin cancers and why?
3. Here we investigate ordinary least squares, weighted least squares, relative efficiency, and the impact of missing data.

Generate a single data set with $m = 200$ subjects using the random intercepts and slopes model:

$$y_{ij} = \beta_0 + \beta_1 t_{ij} + b_0 + b_1 t_{ij} + \epsilon_{ij} \quad (5)$$

where $t_{ij} = j$, for $j = 1, 2, \dots, 10$, $(b_0, b_1)^T \sim \mathcal{N}(\mathbf{0}, D)$, $\epsilon_{ij} \sim \mathcal{N}(0, \sigma^2)$.

We will use $\boldsymbol{\beta} = (10.0, 1.0)$, $\sigma = 1.0$, and $D = \begin{pmatrix} 0.5 & 0 \\ 0 & 0.3 \end{pmatrix}$. Structure (store) the data in the following “long” form:

| id | y | time |
|----|------|------|
| 1 | 10.6 | 1 |
| 1 | 12.3 | 2 |
| 1 | 12.1 | 3 |
| | . | |
| | . | |

Save a copy of the complete data before proceeding with the following questions:

- Randomly delete 20% of the observations (do this by removing rows *not* subjects). Calculate the OLS estimator $\hat{\beta}(I)$, the WLS estimator $\hat{\beta}(W^1)$, where W^1 assumes random intercepts with standard deviation $\tau = 1.0$, error standard deviation $\sigma = 1.0$; and $\hat{\beta}(W^2)$ uses $W = \Sigma^{-1}$ where Σ^{-1} is the true covariance matrix of \mathbf{Y}_i (and known). Comment on the properties of these estimators in this situation (i.e. MCAR, unbalanced data).
- Calculate the asymptotic relative efficiency of $\hat{\beta}(I)$ and $\hat{\beta}(W^1)$ relative to $\hat{\beta}(W^2)$.
- Start with the complete data. For each subject delete all observations that come after the first measurement which is below 6.0. That is, set to missing Y_{ij} for $j > j'$ where $Y_{ij'} < 6.0$ (and is the first such value for subject i). Calculate $\hat{\beta}(I)$, $\hat{\beta}(W^1)$, and $\hat{\beta}(W^2)$ and comment on their properties in this situation (i.e., MAR).
- Start with the complete data. For For each subject delete all observations that come after the first measurement which is below 6.0 **and** delete all values below 6.0. That is, remove any other values that are below 6.0 from the data set created in 2(c). Calculate $\hat{\beta}(I)$, $\hat{\beta}(W^1)$, and $\hat{\beta}(W^2)$ and comment on their properties in this situation.