Homework 5 Statistics 8320 (Wikle) March 17, 2016

DUE: March 24, 2016

1. Assume we have count data, y_i, $i = 1, \ldots, m$ that are assumed to come from an independent Poisson distribution, conditioned on the rate parameter:

$$Y_i|\lambda \sim Poisson(\lambda),$$

where

$$f(y_i|\lambda) = \frac{\lambda^{y_i}}{y_i!}\exp(-\lambda).$$

In addition, we assume that λ is a random effect that follows a Gamma(a,b) distribution:

$$f(\lambda) = \frac{1}{\Gamma(a)b^a}\lambda^{a-1}\exp(-\lambda/b).$$

   (a) Show that $f(y_i) = \int f(y_i|\lambda)f(\lambda)d\lambda$ is a negative binomial distribution.

   (b) Discuss the relevance of this derivation of the negative binomial to the notion of over-dispersed count data.

2. Consider growth data on 6 plants measured at 10 times. It is hypothesized that the growth curves for these plants follows the following nonlinear model:

$$Y_{ij} = \frac{\beta_1 + u_i}{1 + \exp(-(t_{ij} - \beta_2)/\beta_3)} + \epsilon_{ij},$$

where $i = 1, \ldots, 6$ is the number of plants and $j = 1, \ldots, 10$ is the number of times, $u_i \sim N(0, \sigma_u^2)$ and $\epsilon_{ij} \sim N(0, \sigma_\epsilon^2)$ are both independent random effects. Consider the data on the Blackboard website `growthdata.dat`, which gives the 10 times in the first column, and each plant's growth measurement for those times follows in the remaining 6 columns. Fit this model in PROC NLMIXED and test whether it is necessary to have the random effect in this model. In addition, test the null hypothesis that $\beta_3 = 350$. Provide a plot of the fitted model for each plant, along with its measurements.

3. Tornados can be devastating storms in terms of loss of life and property in the US midwest. There have been some studies to suggest a connection between the sea surface temperature (SST) in the tropical Pacific ocean and the occurrence of tornados in certain parts of the US. In this analysis, we will see if there is a possible dependence of yearly tornado counts in the central portion of Missouri to the tropical Pacific SST. In particular, the dataset `ssttornado532001.dat` contains a 21 x 49 matrix. The first row of this matrix contains the SST values for 49 years from 1953-2001. Each of the other 20 rows (rows 2-21) contain the 49 year time series of counts of strong tornados at a grid box associated with the latitude/longitude given in the file `MOtornlatlon.dat`. Consider a Poisson regression model with the covariate of interest being the SST, which can influence

the tornado counts at each spatial location differently (in principle). That is, consider the model:

$$y_{it}|\lambda_{it} \sim Poi(\lambda_{it})$$
$$\log(\lambda_{it}) = \beta_{0,i} + \beta_{1,i}x_t,$$

for location $i = 1, \ldots, n$ and time $t = 1, \ldots, T$. The key question is whether the $\beta_{1,i}$ are significant. In particular, we are interested in whether these are significant and whether a more appropriate model considers spatially-dependent random effects in β_0 and/or β_1. Support your analysis and include supporting plots and output **summaries**.

4. Let $\mathbf{X}$ be a normally distributed random vector with

$$\bar{\boldsymbol{\mu}} = \begin{bmatrix} -3 \\ 1 \\ 2 \end{bmatrix} \quad \boldsymbol{\Sigma} = \begin{bmatrix} 4 & 1 & 0 \\ 1 & 2 & 1 \\ 0 & 1 & 3 \end{bmatrix}$$

   (a) What is the distribution of $\mathbf{Y} = (X_1, X_3)'$?

   (b) What is the conditional distribution of $\mathbf{Y} = (X_1, X_3)'$ given $X_2 = x_2$?

   (c) What is the conditional distribution of X_2 given $X_1 = x_1$ and $X_3 = x_3$?

   (d) What is the distribution of $Z = X_1 + 3X_2$?

5. Twenty-five subjects were examined in a clinical study of the side effects of the use of a particular drug (drug A). The concentrations of a certain substance in liver tissue were measured immediately before the drug was first administered (Y_1), at 30 days (Y_2) and 60 days (Y_3) of continuous treatment with the drug. The vector of sample means and the sample covariance matrix are:

$$\bar{\mathbf{y}} = \begin{bmatrix} \bar{y}_1 \\ \bar{y}_2 \\ \bar{y}_3 \end{bmatrix} = \begin{bmatrix} 14 \\ 24 \\ 22 \end{bmatrix} \quad \mathbf{S} = \begin{bmatrix} 30 & 10 & 14 \\ 10 & 15 & 4 \\ 14 & 4 & 37 \end{bmatrix}$$

   (a) Are the mean concentrations the same at all three time points? (Do a T^2 test and report your results and conclusion).

   (b) The measurements were also taken on a sample of 17 subjects who were given a different drug (drug B). The mean vector and sample covariance matrix are:

$$\bar{\mathbf{z}} = \begin{bmatrix} 15 \\ 20 \\ 24 \end{bmatrix} \quad \mathbf{W} = \begin{bmatrix} 26 & 12 & 10 \\ 12 & 17 & 8 \\ 10 & 8 & 43 \end{bmatrix}$$

      i. Assume the population covariances for the two drugs (A and B) are the same; obtain the pooled estimate of the common covariance matrix and give the degrees of freedom.

      ii. Evaluate Bartlett's test of the null hypothesis $H_0 : \boldsymbol{\Sigma}_A = \boldsymbol{\Sigma}_B$ against the alternative $H_A : \boldsymbol{\Sigma}_A \neq \boldsymbol{\Sigma}_B$. Report M, C^{-1}, MC^{-1},df and your conclusion.

iii. Using the pooled covariance matrix, compute the estimate of the squared Mahalanobis distance between $\mathbf{y}$ and $\mathbf{z}$.

iv. Test the hypothesis of equal means for drug A and B (use T^2 from part (iii)). Report your conclusion.

v. Test the null hypothesis of no drug by time interaction (i.e., parallel profiles) using a T^2 test. Assume equal covariance matrices for drugs A and B.