STAT/BIOSTAT 571: Takehome Exam

To be handed in on Friday 27th February, 2009. No collaboration!

On the class website you will find data on illiteracy and race collected during the US 1930 census. “Illiterate” is defined as being unable to read and over 10 years of age. Race is categorized as “Native Born White”, “Foreign-Born White”, and “Black”. For each of the \( i = 1, \ldots, 49 \) states that existed in 1930 the data consist of the number of illiterate individuals, \( Y_{ij} \), and the total population aged 10 years and older, \( N_{ij} \), by each of the \( j = 1, 2, 3 \) races, coded as Native Born White, Foreign Born White, Black. Let \( p_{ij} \) be the probability of being illiterate for an individual in state \( i \), and of race \( j \). An additional binary state-level variable, \( x_i = 0/1 \), describes whether Jim Crow laws were absent/present in state \( i = 1, \ldots, 49 \). These laws enforced racial segregation in all public facilities.

In this exam we will model the association between illiteracy and race, state and Jim Crow laws using logistic regression models. In particular we are interested in whether illiteracy in 1930 varied by race, varied across states, and was associated with the absence/presence of presence of Jim Crow laws.

1. **7 marks** Calculate the empirical logits of the \( p_{ij} \)'s, and provide informative plots that graphically displays the association between illiteracy and state, race and Jim Crow laws.

2. **24 marks** First consider the Native Born White data only, \( \{Y_{i1}, N_{i1}, i = 1, \ldots, 49\} \) with the following estimation methods:

   - **Binomial**: \( Y_{i1} \sim \text{Binomial}(N_{i1}, p_{i1}) \), with the logistic model
     \[
     \log \left( \frac{p_{i1}}{1 - p_{i1}} \right) = \beta_1
     \]
     for \( i = 1, \ldots, 49 \).
   - **Quasi-Likelihood**: \( E[Y_{i1}] = N_{i1} p_{i1} \), \( \text{var}(Y_{i1}) = \kappa \times N_{i1} p_{i1} (1 - p_{i1}) \), \( i = 1, \ldots, 49 \), with model (1).
   - **GEE**: \( E[Y_{i1}] = N_{i1} p_{i1} \), \( i = 1, \ldots, 49 \), with working independence, and model (1).
   - **Generalized Linear Mixed Model (GLMM)**:
     \[
     \log \left( \frac{p_{i1}}{1 - p_{i1}} \right) = \beta_1^* + b_{i1}
     \]
     with \( b_{i1} \sim_{iid} N(0, \sigma_1^2) \), \( i = 1, \ldots, 49 \).
(a) 5 marks Give careful definitions of the parameters \( \exp(\beta_1) \) in the GEE model, and \( \exp(\beta'_1) \) in the GLMM.

(b) 4 marks Fit the binomial model to the native born white data, and give a 95% confidence interval for the odds of native born white illiteracy. Is the fit good?

(c) 4 marks Fit the quasi-likelihood model to the data, and give a 95% confidence interval for the odds of native born white illiteracy.

(d) 4 marks Fit the GEE model to the data, and give a 95% confidence interval for the odds of native born white illiteracy. How does the GEE approach differ from quasi-likelihood here? Which do you prefer?

(e) 7 marks Fit the GLMM model to the data using a likelihood approach, and give a 95% confidence interval for the odds of native born white illiteracy, and an estimate of the between-state variability in logits. Are the results consistent with the GEE analysis?

3. 6 marks We will now consider data on all three races. Using GEE fit separate models to each race. Give a 95% confidence interval for the odds ratios comparing illiteracy between foreign-born whites and native born whites, and comparing blacks with native born whites. Is there any problem with this analysis?

4. 8 marks Use GEE to fit a model to all three races simultaneously and compare your answer with the previous part. Which analysis is the most appropriate, and why?

5. 10 marks Fit the GLMM

\[
\log \left( \frac{p_{ij}}{1 - p_{ij}} \right) = \beta_j^* + b_{ij} \tag{3}
\]

with \( b_{ij} \sim_{\text{ind}} N(0, \sigma^2_j) \), \( j = 1, 2, 3 \), using likelihood-based methods. Give 95% confidence intervals for the odds ratios comparing illiteracy between foreign born whites and native born whites, and comparing blacks with native born whites. Are your conclusions the same as with the GEE analysis? Does this model require refinement?

We will now add the state-level Jim Crow law indicator to the model.

6. 10 marks Consider the model

\[
\log \left( \frac{p_{ij}}{1 - p_{ij}} \right) = \beta_j + \gamma_j x_i \tag{4}
\]

Give interpretations of each of \( \exp(\beta_1) \), \( \exp(\beta_2) \), \( \exp(\beta_3) \), \( \exp(\gamma_1) \), \( \exp(\gamma_2) \), \( \exp(\gamma_3) \). Fit this model to the three races using GEE and interpret and summarize the results in a clear fashion.
7. **20 marks** Consider a Bayesian fitting of the GLMM:

\[
\log \left( \frac{p_{ij}}{1 - p_{ij}} \right) = \beta_j^* + \gamma_j^* x_i + b_{ij} \tag{5}
\]

where \( b_i \sim_{iid} N_3(0, D) \) with \( b_i = [b_{i1}, b_{i2}, b_{i3}]^T \) and

\[
D = \begin{bmatrix}
\sigma_1^2 & \rho_{12}\sigma_1\sigma_2 & \rho_{13}\sigma_1\sigma_3 \\
\rho_{21}\sigma_2\sigma_1 & \sigma_2^2 & \rho_{23}\sigma_2\sigma_3 \\
\rho_{31}\sigma_3\sigma_1 & \rho_{32}\sigma_3\sigma_2 & \sigma_3^2
\end{bmatrix}
\]

is a 3 \times 3 variance-covariance matrix for the random effects \( b_i \). Assume improper flat priors for \( \beta_j^*, \gamma_j^*, j = 1, 2, 3 \), and the Wishart prior \( W = D^{-1} \sim \text{Wishart}(r, S) \), parameterized so that \( \mathbb{E}[W] = rS \), with \( r = 3 \) and

\[
S = \begin{bmatrix}
30.45 & 0 & 0 \\
0 & 30.45 & 0 \\
0 & 0 & 30.45
\end{bmatrix}.
\]

Carry out a Bayesian analysis using this model and interpret and summarize the results in a clear fashion.

8. **15 marks** Write a short summary of what you have found, concentrating on the particular substantive questions of interests stated in the introduction.

Please include all code as a separate appendix.