STAT/BIOSTAT 571: Homework 10 — Final Exam

Please note this is an exam so **no conferring or discussion with other students is allowed**. You may ask the instructor and/or TAs for clarifications.

To be handed in on Weds March 16th, by noon, in Ken's F-wing mailbox. Please see 'Chapter 0' of the slides for a summary of how to answer questions appropriately, and the guidelines from 570. Where solutions require use of `R`, summarize your findings in a written answer, and append your **annotated** code, to show what you did. For each question, write up your solution on your own, using **full sentences**.

1. [**Conditional likelihoods, 3 points**] Consider the conditional likelihood approach to the progabide example. As an alternative to the model-based standard errors seen in class (slide 3.134) consider use of robust standard errors, applied to the conditional likelihood's score equations. Two different ways to do this are coded below;

```
> library("sandwich")
> glm1 <- glm(cbind(Yi1, Ti-Yi1) ~ trtbas , family=binomial)
> round( cbind( est=coef(glm1), rob.se=sqrt(diag(vcovHC(glm1, "HC0"))) ) , 2)
              est rob.se
(Intercept) -0.11   0.12
trtbas       0.10   0.21
>
> Ylong   <- rep( rep(c(1,0),each=59), c(Yi1, Ti-Yi1) )
> trtbaslong <- rep( c(trtbas,trtbas),  c(Yi1, Ti-Yi1) )
> glm2 <- glm(Ylong ~ trtbaslong, family=binomial)
> round( cbind( est=coef(glm2), rob.se=sqrt(diag(vcovHC(glm2, "HC0"))) ) , 2)
              est rob.se
(Intercept) -0.11   0.05
trtbaslong   0.10   0.07
```

   The point estimates are identical but these analyses have different standard error estimates. By thinking about the robustness that each provides, state why they are different, and which (if either) is appropriate for the use in the conditional likelihood method presented in class.

2. [**Bayesian analysis of mixed models, 5 points**] The class site contains data from a non-randomized longitudinal study, in which subjects taking one of two treatments have their cholesterol measured at 5 time points. Some missing values are present, which you should assume are Missing At Random.

   (a) Fit the following model to the complete observations, and report posteriors for each of the parameters.

$$\begin{aligned}
\tau_b, \tau_Y &\overset{i.i.d.}{\sim} \Gamma(\text{shape} = 0.1, \text{rate} = 0.1) \\
\beta_0, \beta_1, \beta_2, \beta_3 &\overset{i.i.d}{\sim} N(0, 10000) \\
b_i|\tau_b &\overset{i.i.d}{\sim} N(0, 1/\tau_b) \\
Y_{it}|b_i, \text{trt}_i &\overset{indept}{\sim} N(\beta_0 + b_i + \beta_1 \text{trt}_i + \beta_2 t + \beta_3 \text{trt}_i t, 1/\tau_Y)
\end{aligned}$$

   where $t = 1, 2, 3, 4, 5$ and trt_i is an indicator for receiving treatment 1 (versus treatment 0). Note: you are free to use any numerical method, and several options are available, but say what you did

(b) Give diagnostics plots for the model fitted in a), briefly describing what they show. Note that you are not being asked for MCMC diagnostics

(c) The data contains some missing values. Suppose you had used MCMC in part a) but instead of using complete-observation data, had used Bayesian methods to impute the missing outcomes. Briefly, state how you could use a small number of steps from this chain in a multiple-imputation analysis. Note you are **not** required to implement such an analysis

3. **[Marginal and conditional: 6 points]** Read the paper by Ritz and Spiegelman (Statistical Methods in Medical Research, 2004) that is available on the course website. Using full sentences and paragraphs throughout, write a brief summary of all the main ideas and results (1–2 double spaced typed pages, with 12 point font and 1 inch margins). You may omit §2.5, which deals with survival analysis. While it may be helpful for your summary's structure to be similar to the original paper, do not plagiarize the paper or other sources; write using your own words and do not copy complete sentences or phrases from elsewhere.

4. **[Missing data in GEE and mixed models, 5 points]** Suppose you are working with a statistician, on linear regression-based analysis of data that contains missing values. To understand the behavior of GEE and LMMs under MAR, your co-author has simulated data very similarly to the setting of HW6, and 3.204–3.206;

$$\begin{aligned} b_i &\overset{i.i.d}{\sim} N(0, \sigma_b^2) \\ X_i &\overset{i.i.d}{\sim} Bern(0.5) \\ Y_{it}|b_i, X_i = x &\overset{indept}{\sim} N(\beta_0 + b_i + \beta_1 t + \beta_2 x, \sigma_Y^2), \text{for } t = 1, 2, 3, \end{aligned}$$

where $n = 1000$, $\sigma_b = 1, \sigma_Y = 0.5, \beta_0 = 1, \beta_1 = -1, \beta_2 = 0.5$, and if any $Y_{it} < -1$ then all subsequent $Y_{it'}$ are missing. (Note the original version used $Y_{it} < 0$; this version is less extreme)

Despite the robustness of GEE, and correct mean-model and variance (i.e. second-moment) assumptions, your coauthor is surprised to see that 95% CIs from GEE with an exchangeable working matrix do not provide as good coverage of β_1 and β_2 as those from LMM analysis. They feel the GEE approach should provide the same robustness to MAR missing outcome as LMM, as it relies on assumptions about the first two moments, just like LMM.

(a) Provide simulation results confirming your co-authors' findings; that GEE linear regression here does not give good coverage, but LMM analysis does. These simulations need not be large-scale, but should be clear enough to confirm this finding unambiguously. As always, you should report what you did

(b) After carefully considering the differences between the two anlayses, write a short explanation for your co-author, of why the GEE approach does not work as well as LMM

5. **[(G)LMMs and assumptions, 8 points]** Suppose we have a data-generating mechanism where

$$\begin{aligned} Z_i &\overset{i.i.d.}{\sim} Bern(p) \\ X_{ij}|Z_i = z &\overset{indept}{\sim} Bin(2, \text{expit}(\theta_0 + \theta_1 z)) \\ b_i|\boldsymbol{X}_i, Z_i &\overset{i.i.d.}{\sim} N(0, \sigma_b^2) \\ Y_{ij}|\boldsymbol{X}_i, Z_i, b_i &\overset{indept}{\sim} Bern(\text{expit}(\beta_0 + b_i + \beta_1 X_{ij} + \beta_2 Z_i)), \end{aligned}$$

and the inference of interest is a test of the null hypothesis, that keeping cluster and confounding variable Z fixed, there is no association between Y and X. Throughout, our simulations will use $\beta_1 = 0$, i.e. generating data under the null hypothesis.

(a) Using a simulation study for $n_i = 3$ and $n = 1000$, with $p = 1/4, \theta_0 = -2, \theta_1 = 1.5, \beta_0 = -2.5, \beta_2 = 1.5, \sigma_b = 0.1$, summarize the distribution of p-values produced by fitting LMM regression of Y on X and Z, with random intercepts for each cluster, and performing a Wald test using the X coefficient. (The choice of how to summarize is up to you, but at minimum you should indicate the nominal and actual Type I error rates of the test at $\alpha = 0.05$, indicating the extent of any Monte Carlo error in your answers)

(b) Contrast your summary in a) with the same summary of Wald test p-values from fitting GLMM logistic regression, of Y on X adjusting for Z with random intercepts for each cluster. Hint: use the same datasets as in part a)

(c) Repeat a) and b) for the same data-generating mechanism, but with $\beta_0 = -0.75$

(d) Using what you know about the importance of assumptions of LMMs and GLMM (correct mean model, correct within-cluster variance, mean-variance relationship, Normality of outcomes etc) carefully explain any differences or lack of difference in the results between the two simulation setups

6. **[Review, 3 points]** Using the most up-to-date version of the class slides, find 6 remaining typos, and state what a correct version would be. Note: do not start this question until noon on Monday, March 14th, or until notified by the instructor.