BIO 226
Final Take Home Exam: Spring 2015

Instructions:

1. You must work alone on this exam, though you may use any books, SAS software documentation, course notes etc.

2. You may provide brief extracts from your SAS program or SAS output as an Appendix.

3. The exam is due at 1:30pm at class on Thursday May 14, 2015. If you won't be at that class, please turn it in to Brent at his office (HSPH Building II, Room 413) prior to the class. Exams not turned in by the due date/time will not be graded unless this has been agreed upon with the Instructor in advance.

4. The TAs will hold their usual office hours but they won't be able to answer questions or emails about the content of the exam. They will be able to help with questions about the course material.

5. If you have important questions of clarification about the exam, please email me at bcoull@hsph.harvard.edu.

Question 1:

The question concerns the analysis of a subset of data from a clinical trial which randomized patients with arthritis to one of two treatments: auranofin (A) or placebo (P). The data provided are from a subset of 51 patients. The outcome measure of interest is a binary self-assessment of arthritis status (0=poor, 1=good). The self-assessments were scheduled at weeks -1 and 0, and then at weeks 4, 8 and 12. Randomization occurred immediately after the week 0 assessment. The primary objective of the analysis is to compare the effects of auranofin and placebo on self-assessed arthritis status.

The data from this study are stored in the file arthritis.txt, which can be found on the course web site. Each row of the data set contains the following variables: subject ID number, sex (M=male, F=female), subject's age in years at entry to the study, the randomized treatment group (A or P), and the self-assessment results at weeks -1, 0, 4, 8 and 12, respectively. Although the subjects' ages are provided, these data are not used in any question.

a. Briefly summarize the outcome data from the study in a single table. Specifically, show for each treatment group at each scheduled assessment: how many provided an assessment and what number and percentage rated their arthritis status as "good". Based on the descriptive statistics in the table, briefly characterize the pattern of change in the proportions with a "good status" over time including a comparison of the two treatments.

b. In the following questions, use PROC GENMOD in SAS to fit marginal logistic regression models using the approach of generalized estimating equations (GEE) to obtain parameter estimates. Define the time variable, T, to be a categorical variable with 4 levels: taking the value 99 for measurements at weeks -1 and 0 (i.e. for both "baseline" measurements prior to randomization) and values 4, 8 and 12 for the measurements at weeks 4, 8 and 12. The treatment variable, X, is an indicator variable taking the value 1 if a subject received auranofin (A) and the value 0 if placebo (P).

i. Consider a marginal logistic regression model for the repeated arthritis self-assessment measurements as the outcome variable and with just the T variable and a T*X interaction variable as categorical explanatory variables, using an independence working correlation matrix to describe the within-subject correlation structure (i.e. using the TYPE=IND on the REPEATED statement). Define appropriate notation and write down the algebraic form for the model that will be fitted using the GEE method, including any assumptions.

ii. The model does not include the variable X as a main effect. What assumption is therefore being made? Briefly describe why this assumption might be reasonable in this study.

iii. Fit the model described in (i). Provide the PROC GENMOD code used and the estimates of the parameters in the mean part of the model obtained as well as empirical standard errors for these estimates. [HINT: Add "/ param=ref " after your list of variables in your class statement to ensure that SAS doesn't try to outsmart you and add main effects of X back into the model].

iv. How is the model fit in (iii) different from an ordinary logistic regression analysis that assumes that all observations are independent? Note here the question is not "how do the results from the two analyses differ for these data", but how the models themselves differ – that is, why would they give different results. Which approach do you prefer and why?

v. For the model in (i), conduct a Wald test of the joint hypothesis that there is no effect of auranofin compared with placebo at all three of weeks 4, 8 and 12. [HINT: Use the TYPE3 and WALD options together on the MODEL statement]. What do you conclude concerning the effect of auranofin versus placebo? [Note: do not just provide a p-value from the test of the joint hypothesis; instead provide a broader interpretation of the results concerning any treatment differences].

vi. Now using PROC GENMOD, fit the same model but using an exchangeable log odds ratio structure for the working association structure. Write out how your model formulation given in (i) changes under this alternative assumption. Describe whether the conclusions of your GEE analyses in (vi) are sensitive to the change in association structure.

c. The investigator is interested in knowing whether the results comparing the effect of auranofin versus placebo on arthritis status over time is affected by any imbalance in sex between the treatment groups. Conduct analysis to address this specific issue. Briefly describe how you approached this (include your PROC GENMOD code) and what your key findings are.

Question 2:

This question is based on the study that was featured in Homework 8. Briefly, recall that the Skin Cancer Prevention Study was a randomized, double-blind, placebo-controlled multi-center clinical trial designed to test the effectiveness of beta-carotene in the prevention of non-melanoma skin cancer in high-risk subjects. Subjects were randomized to either placebo or 50 mg of beta-carotene per day for 5 years. Subjects were examined once per year and the outcome variable, Y, is a count of the number of new skin cancers per year.

Selected data from the study are in the dataset called "skin.txt" on the course web site. Each row of the dataset contains the following 9 variables: ID, Center, Age, Skin, Gender, Exposure, Y, Treatment, Year. These variables take values as follows:

ID: Subject identifier number
Center: Identifier number for center of enrollment.
Age: Subject's age in years at randomization
Skin: Skin type (1=burns; 0 otherwise) [evaluated at randomization and doesn't change with time]
Gender: 1=male; 0=female
Exposure: Count of number of previous skin cancers [prior to randomization]
Y: Count of number of new skin cancers in the Year of follow-up
Treatment: 1=beta-carotene; 0=placebo
Year: Year of follow-up after starting randomized treatment

As in HW8, you may assume that the counts of new skin cancers, Y, are from exact one-year periods (so that no offset term is needed).

The investigator at Center=1 is interested in characterizing the distribution of risk among subjects at her center. **In the following, only include the subset of subjects with Center=1 in the analysis.**

a) Provide an algebraic definition for a generalized linear marginal model in which the only effects are for the intercept and Year (as a continuous variable). Fit this model and provide a table from your SAS output which includes the estimates of the parameters in your model.

b) Provide an algebraic definition for a generalized linear mixed model (GLMM) in which the only fixed effects are for the intercept and Year (as a continuous variable), and the only random effect is the intercept. **Assume there is no overdispersion.** What is being assumed about how the distribution of risk among subjects changes with time?

c) Fit your chosen GLMM from (b) in SAS and provide the SAS code for the fitting the model (just provide the code from the PROC that you use). Provide a table from

your SAS output which includes the estimates for the parameters in your GLMM, and provide careful interpretation of the Year term.

d) Are the estimates for the fixed intercept terms the same or different in the GLMM compared with the marginal model fitted in question 2a? Why are they the same or different?

e) Use the parameter estimates from your GLMM and your model definition to characterize the distribution of expected counts of new skin cancers among subjects at center 1 during their first year of follow-up.

<u>Optional</u> Extra Credit Problem* (This problem is meant to offer another chance to demonstrate understanding of some of the material on the mid-term. If you choose to do this problem and your score is higher than your mid-term grade, then your mid-term grade will be reweighted to be New Midterm Grade = .7*Old Midterm Grade + .3*Extra Credit Problem).

Onychomycosis, popularly known as toenail fungus, is a fairly common condition that not only can disfigure and sometimes destroy the nail but that also can lead to social and self-image issues for sufferers. Tight-fitting shoes or hosiery, the sharing of common facilities such as showers and locker rooms, and toenail polish are all thought to be implicated in the development of onychomycosis. This question relates to data from a study conducted by researchers that recruited sufferers of a particular type of onychomycosis, dermatophyte onychomycosis. The study conducted by the researchers was focused on comparison of two oral medications, terbinafine (given as 250 mg/day, denoted as treatment 1 below) and itraconazole (given as 200 mg/day, denoted as treatment 2 below). The trial was conducted as follows. 200 sufferers of advanced toenail dermatophyte onychomycosis in the big toe were recruited, and each saw a physician, who removed the afflicted nail. Each subject was then randomly assigned to treatment with either terbinafine (treatment 1) or itraconazole (treatment 2). Immediately prior to beginning treatment, the length of the unafflicted part of the toenail (which was hence not removed) was recorded (in millimeters). Then at 1 month, 2 months, 3 months, 6 months, and 12 months, each subject returned, and the length of the unafflicted part of the nail was measured again. A longer unafflicted nail length is a better outcome. Also recorded on each subject was gender and an indicator of the frequency with which the subject visited a gym or health club (and hence might use shared locker rooms and/or showers).

The data are available in the file toenail.txt on the class web page. The data are presented in the form of one data record per observation; the columns of the data set are as follows:

1 Subject id
2 Health club frequency indicator (= 0 if once a week or less, = 1 if more than once a week)
3 Gender indicator (= 0 if female, = 1 if male)
4 Month
5 Unafflicted nail length (the response, mm)
6 Treatment indicator (= 1 if terbinafine, = 2 if itraconazole)

The researchers had several questions, which they stated to you as follows:

i) Is there a difference in the pattern of change of lengths of the unafflicted part of the nail between subjects receiving terbinafine and itraconazole over a 12 month

period? Does one treatment show results more quickly?

ii) Is there an association between the pattern of change of nail lengths and gender and/or health club frequency in subjects taking terbinafine? This might indicate that this drug brings about relief more swiftly in some kinds of subject versus others.

In answering these scientific questions of interest, clearly write out the analytic models you consider for answering these questions. Clearly outline your decision making process for how you selected your final models. Fit your chosen final models and report to the project investigators on the stated scientific questions of interest.

* We acknowledge Marie Davidian for the development of this problem.