MIDTERM EXAM
February 12, 2010

Biostat/Stat 571
Winter 2010
P. Heagerty

This exam is to be completed **independently**. Do not discuss your work with other students. If you have any questions then please contact me via e-mail or by coming to my office. The exam is due in class on Friday February 19, 2010.

The data are available in the **DataSets** section of the course web page.

---

1. [40 points total] INTRODUCTION:

A study was conducted to evaluate various doses of a monoclonal antibody for the treatment and prevention of pollen allergy symptoms. A total of 150 subjects were randomized to either placebo, or to 15 mg/day, 30 mg/day, or 60 mg/day of the active agent. The study was conducted during the spring when "cedar fever" is a common condition that constitutes a reaction to cedar pollen. Patients reported nasal symptoms (sneezing, itching, runny nose, stuffy nose) on an ordered categorical severity scale (1=none, 2=mild, 3=moderate, 4=severe).

Several factors thought to be associated with the presentation of symptoms were also recorded, including the pollen count for the day on which the symptom data were collected, in addition to gender, height, weight and the date (days since **Sept 1**).

(a) [20 pts] Summarize the evidence for a treatment effect for the symptom "*stuffy*" (e.g. stuffy nose). Specifically, provide graphical and numerical summaries that can be used to evaluate the efficacy of treatment. Justify the methods that you use, and provide a written summary of your analysis with a summary of your conclusions (appropriate for a general scientific audience). Provide an interpretation of key parameters that characterize the effect of treatment.

(b) [15 pts] For the symptom *stuffy* is there evidence that the effect of treatment depends on the level of pollen? Provide graphical and numerical summaries that support your conclusions. Justify the methods that you use, and provide a written summary of your analysis with a summary of your conclusions (appropriate for a general scientific audience).

(c) [5 pts] For the symptom *stuffy* your collaborator suggests that the data simply be reduced to the binary indicator 0=no symptoms, 1=mild or greater symptoms. For the question of treatment efficacy (e.g. part (a) above) provide a summary calculation that would communicate the loss of information associated with use of a binary predictor rather than the use of the complete ordered categorical outcome.

2. [40 points total] INTRODUCTION:

In the Vaccine Preparedness Study (VPS) there are 1,588 MSM (men who have sex with men) who reported having more than one partner during the 6 months prior to enrollment. Scientific interest is in whether drug and/or alcohol use correlate with a higher proportion or number of partners of unknown HIV status. For each subject the number of partners that were of unknown HIV status (`n.unk`), the number of partners that were HIV+ (`n.pos`), and the the number of partners that were HIV− (`n.neg`) were recorded. In addition, demographic and behavioral factors were measured.

The covariates of primary interest are:

`poppers` - an indicator of popper use (1=yes,0=no).
`amphetamines` - an indicator of amphetamine use (1=yes,0=no).
`alcohol` - an indicator of excessive alcohol use (1=yes,0=no).

The study was a multi-site study with recruitment from 6 locations. Subjects were recruited either as individuals who had previously participated in an HIV study, or as new participants. The available demographic variables are:

`site` (1,2,3,4,5,6)
`age` (in years)
`cohort` 1 if the subject was a new recruit, and
0 if previously enrolled in a study.

There was prior evidence that popper use varied across the recruitment sites, and varied with the age of the subject.

(a) [20pts] One analysis could focus on the *proportion* of partners that are of unknown HIV status. However, it could be argued that the *number* of **HIV unknown partners** is a more meaningful outcome since this may reflect the actual risk of infection. (That is, perhaps a subject with 2 HIV unknown partners and 6 total partners is at greater risk than a subject with 1 HIV unknown parter and 2 total partners.) Therefore, summarize the relationship between the predictors of interest (`amphetamines`, `poppers`, and `alcohol`) and the mean number of unknown partners (`n.unk`). State the methods that you use, and interpret the results of your analysis. Do these data provide evidence that specific drug use, or alcohol use is associated with an increase in the number of partners that are of unknown HIV status?

(b) [20pts] A new study is being planned that will target a reduction in the mean number of HIV unknown partners. The investigators plan to use Poisson regression to estimate the impact of a new behavior change program (individual counseling) as compared to a control program (only distribute written information). Thus a model for the mean number of partners over 6 months post-intervention would be given as $\log \mu_i = \beta_0 + \beta_1 X_i$ where $X_i = 1$ if subject i is randomized to the new intervention, and 0 if to the control group. The primary test will

be a simple (valid) Wald test of $H_0 : \beta_1 = 0$. If the investigators wish to have 80% power to detect a reduction in the mean number of partners of 25% then what is the total sample size that they would need? Assume that the investigators will recruit from a similar population and therefore you can use your analysis in (a) to inform the sample size calculation. Please outline the assumptions / methods that you use to determine the sample size.

3. [20 points total] INTRODUCTION:

Let $U_n(\theta)$ be an estimating function where $U_n(\theta)$ is a $p \times 1$ vector that is a function of the $p \times 1$ parameter θ and the data (Y_i, X_i). Let $\widehat{\theta}$ be the estimator that solves the estimating equation $U_n(\theta) = 0$. Assume that the estimating equation is formed as the sum of contributions from $i = 1, 2, \ldots, n$ independent observations:

$$U_n(\theta) = \sum_{i=1}^{n} U(Y_i, \theta)$$

where $U(Y_i, \theta)$ may also depend on covariates (as we have used for GLM regression setting).

(a) [10pts] Suppose that the estimating function is not "unbiased" in the sense that $E[U(Y_i, \theta)] = \delta_i$ (**Note**: this clarifies the notation here by clearly labeling $U(\theta)$ as referring to $U(Y_i, \theta)$ as opposed to $U_n(\theta)$) rather than the standard assumption where $E[U(Y_i, \theta)] = 0$. The "bias" for the estimating function may be the result of some form of model mis-specification. **Note**: assume either that $\delta_i \equiv \delta$ or that $\frac{1}{n}\sum_i \delta_i \to \delta$.

Sketch **a proof that shows what** the asymptotic distribution of the estimator $\widehat{\theta}$ is in this situation. Please briefly state the main assumptions that you use (ie. a CLT, or a WLLN) but you need not list the specific technical assumptions that are needed to invoke these limit theorems.

(b) [10pts] Now suppose that the estimating equations are the score equations for a model where the parameter θ can be partitioned into two components: $\theta = (\beta, \alpha)$ where β is a $(p \times 1)$ vector and α is a $(q \times 1)$ vector. Furthermore, we can partition the estimating equations into two corresponding components:

$$\begin{aligned} U_{1,n}(\theta) &= \frac{\partial}{\partial \beta} \sum_{i=1}^{n} \log \mathcal{L}_i \\ U_{2,n}(\theta) &= \frac{\partial}{\partial \alpha} \sum_{i=1}^{n} \log \mathcal{L}_i \end{aligned}$$

where $\mathcal{L}_i = P(Y_i; \theta)$. Finally, suppose that $E[U_{1,n}(\theta)] = n \cdot \delta_1 = 0$, while $E[U_{2,n}(\theta)] = n \cdot \delta_2 \neq 0$. (**Note**: this clarifies the order of the bias). This situation corresponds to, for example, the negative binomial scenario where the mean model is correctly specified but the dependence model may not be. Based on your results in part (a), under what assumptions will the component $\widehat{\beta}$ of the estimator $\widehat{\theta}$ remain consistent even though $U_{2,n}(\theta)$ is biased?