Biostat/Stat 571
Winter 2010
P. Heagerty

Exercise #5
Due: February 10, 2010

Reading:
- Verbeke & Molenberghs, Chapters 4,5, and 6
  - Diggle, Heagerty, Liang & Zeger, Chapter 6 (ANOVA)

1. In Exercise #4 we generated correlated data assuming: random intercepts; random intercepts and slopes; and autocorrelated random errors. The goal of this first question is to assess the impact of correlation on the standard error estimates obtained from ordinary least squares. Recall that OLS estimators of regression parameters remain unbiased when data are correlated. However, the "naive" standard errors based on $\sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1}$ will generally be invalid. Consider the following two regression model designs that are represented by a linear model:

$$E(Y_{ij} \mid \boldsymbol{X}_i) = \beta_0 + \beta_1 \cdot t_{ij} + \beta_2 \cdot x_{ij}$$

Mean Model (**A**): A linear model with time, $t_{ij} = j$, and a group indicator variable, $x_{ij} = 0$ or $x_{ij} = 1$, where the group indicator variable is "cluster-specific" meaning that x_{ij} is constant for all j within unit i. Assume that 100 clusters have $x_{ij} = 0$ and 100 clusters have $x_{ij} = 1$, and that for all clusters $n_i = 12$.

Mean Model (**B**): A linear model with time, $t_{ij} = j$, and a group indicator variable, $x_{ij} = 0$ or $x_{ij} = 1$, where the group indicator variable is "subject-specific" meaning that x_{ij} is not constant for all j within unit i. Assume that 100 clusters have $x_{ij} = 0$ for $j = 1:6$, and $x_{ij} = 1$ for $j = 7:12$, while 100 clusters have $x_{ij} = 1$ for $j = 1:6$, and $x_{ij} = 0$ for $j = 7:12$. Assume that for all clusters $n_i = 12$.

(a) If data actually have a random intercepts correlation structure as described in Question 2(a) of Exercise #4 then what is the standard error that is reported by OLS (the "naive estimate") and what is the correct standard error? Consider both $(\sigma = 20, \tau = 5)$ and $(\sigma = 5, \tau = 20)$ for both Model A, and Model B above.

(b) If data actually have a random intercepts and random slopes covariance structure as described in Question 2(b) of Exercise #4 then what is the standard error reported by OLS (the "naive estimate") and what is the correct standard error? Let $\sigma = 1$ and consider both $\boldsymbol{D} = \text{diagonal}(100, 2)$ and $\boldsymbol{D} = \text{diagonal}(50, 4)$ for both mean models (A and B above). (Note: In this case the variance of Y_{ij} is not constant over time, so to compute the OLS standard errors you will need to obtain the average variance.)

(c) If data actually have autocorrelated errors as described in Question 2(c) of Exercise #4 then what is the standard error reported by OLS (the "naive estimate") and what is the correct standard error? Let $\sigma = 10$, $\tau = 20$ and consider both $\rho = 0.9$ and $\rho = 0.5$ for both Model A, and Model B above.

(d) Summarize the patterns that you see in (a)-(c). In particular is the impact of correlation similar for β_1 in Model A and β_1 in Model B? Is the impact of correlation similar for β_2 in Model A and β_2 in Model B?

2. Here we investigate ordinary least squares, weighted least squares, relative efficiency, and the impact of missing data.

Generate a single data set with $m = 200$ subjects and $n_i = 10$ using the random intercepts and slopes model given by

$$Y_{ij} = \beta_0 + \beta_1 \cdot t_{ij} + b_{i0} + b_{i1} \cdot t_{ij} + \epsilon_{ij}$$

with $\boldsymbol{\beta} = (100.0, -5)$, $\text{var}(\epsilon) = \sigma^2 = 1.0$ and $\text{cov}(\boldsymbol{b}_i) = \boldsymbol{D} = \begin{bmatrix} 100 & 0 \\ 0 & 2 \end{bmatrix}$ for $\boldsymbol{b}_i = (b_{i0}, b_{i1})$. Assume that $t_{ij} = j$ for $j = 1, 2, \ldots, 10$. The design matrices are $\boldsymbol{X}_i = [\boldsymbol{1}, \boldsymbol{t}_i]$ for $\boldsymbol{t}_i = (1, 2, \ldots, 10)^T$
Recall that under this model we have $E(\boldsymbol{Y}_i \mid \boldsymbol{X}_i) = \boldsymbol{X}_i\boldsymbol{\beta}$, and $\text{cov}(\boldsymbol{Y}_i \mid \boldsymbol{X}_i) = \boldsymbol{X}_i\boldsymbol{D}\boldsymbol{X}_i^T + \sigma^2\boldsymbol{I}$. Structure (store) the data in the following "stacked" fashion:

```
id  y      time
 1  110.6  1
 1  112.3  2
     .
     .
 1  121.6  10
 2  109.2  1
 2  110.3  2
     .
     .
```

Save a copy of the complete data before proceeding with the following questions.

(a) Randomly delete 20% of the observations (do this by removing individual values not subjects). Calculate the OLS estimator $\widehat{\boldsymbol{\beta}}(I)$, the WLS estimator $\widehat{\boldsymbol{\beta}}(W^1)$, where W^1 assumes random intercepts with $\sigma = 1.0, \tau = 10.0$, and $\widehat{\boldsymbol{\beta}}(W^2)$, that uses $W = \boldsymbol{\Sigma}^{-1}$ where $\boldsymbol{\Sigma}$ is the true covariance matrix (and known). Comment on the properties of these estimators in this situation (ie. MCAR, unbalanced data).

(b) Calculate the asymptotic relative efficiency of $\widehat{\boldsymbol{\beta}}(I)$ and $\widehat{\boldsymbol{\beta}}(W^1)$ relative to $\widehat{\boldsymbol{\beta}}(W^2)$.

(c) [**optional**] Start with the complete data. For each subject delete all observations that come after the first measurement which is below 65. That is, set to missing Y_{ij} for $j > j'$ where $Y_{ij'} < 65$ (and is the first such value for subject i). Calculate $\widehat{\boldsymbol{\beta}}(I)$, $\widehat{\boldsymbol{\beta}}(W^1)$, and $\widehat{\boldsymbol{\beta}}(W^2)$ and comment on their properties in this situation (ie. MAR).

(d) [**optional**] Start with the complete data. For each subject generate a dropout time based on the CRM model:

$$\text{logit}P(\text{dropout} = j \mid \text{dropout} \geq j,\ b_{i0},\ b_{i1}) = -2.5\ -\ 0.1 \cdot b_{i0}\ -\ (j/2.5) \cdot b_{i1}$$

Delete all observations after the subjects' dropout time. Plot these data and add a smooth curve to indicate the mean as a function of time. Finally, calculate $\widehat{\boldsymbol{\beta}}(I)$, $\widehat{\boldsymbol{\beta}}(W^1)$, and $\widehat{\boldsymbol{\beta}}(W^2)$ and comment on their properties in this situation.

3. The Multi-Center Aids Cohort Study (MACS) recruited HIV negative (uninfected) men starting in 1984 and followed them prospectively with follow-up visits scheduled every six months. The goal of the study was to learn about factors associated with HIV infection and with disease progression among subjects that became infected. The data `MACS-cd4-vload0.data` in the Data section of the course web page contain repeated measures on a subset of the MACS participants. The analysis data contains measurements of the number of CD4 positive cells, a biological marker of one aspect of the health of the participants. The data set contains measurements taken during the first four years after infection with HIV (the approximate date of infection is termed the "seroconversion" date). Any subject that died during the first four years has been excluded from our analysis data.

Scientific interest is in whether "viral load" measured at baseline (ie. within the first year after infection) is associated with the rate of decline in health as characterized by the CD4 count. Viral load refers to the amount of HIV virus that is present in the blood of an infected person. Present graphical and confirmatory summaries that assess whether the rate of decline in CD4 counts is associated with baseline viral load.

(a) Summarize the data set using simple numerical and/or graphical summaries as relevant to the scientific question above.

(b) Use appropriate exploratory methods to characterize the covariance structure of these data. What structured covariance model(s) appear plausible for these data?

(c) One approach to the analysis of "rates of decline" would be to fit a separate linear model to each subject and then compare the rate of decline to the baseline value of viral load. Perform this analysis and interpret your results. Please also comment on the observed magnitude of variation in rates of decline for this patient population – that is, summarize the magnitude of unexplained variation in the rates of decline.

(d) Comment on the limitations of your analysis in part (c) and suggest an alternative regression approach that could be used to address the scientific question.

(e) [**optional**] Execute the regression approach that you suggest in 1(d) and interpret the results.