LDA 140.655
MIDTERM - 02/09/2005
DUE DATE - 02/21/2005

1. Suppose you are designing a randomized intervention trial of a social network method for retaining drug users in treatment programs, thereby reducing drug consumption. There will be two groups: the first will get the usual drug treatment; the second will get the usual treatment plus weekly meetings with their social network, who will support their retention in the treatment program.

   Identify an outcome measure that leads to each of a cross-sectional and longitudinal data analysis

2. Suppose we observed longitudinal data $(y_{ij}, x_{ij}),\ j = 1, \ldots, n,\ i = 1, \ldots, m$ where y_{ij} is a response and x_{ij} is a scalar predictor for observation j on person i. Suppose that data can reasonably be described by the linear regression

   $$y_{ij} = x_{i1}\beta_c + (x_{ij} - x_{i1})\beta_L + \epsilon_{ij}$$

   What simple one (predictor) linear regression can be performed to estimate β_c? β_L?

3. Choose one of the course data sets (or one of your own, if you have burning desire to do so) and create a graphical or tabular display of the evidence relevant to a scientific question specified by you. In particular:

   - State the scientific question of interest
   - Design and produce the display
   - Summarize the evidence apparent in your display relevant to your question
   - Make a scatterplot matrix and assess whether the assumption of stationarity is reasonable. If so, estimate an autocorrelation function.

4. In the dental study,

   `http://www.biostat.jhsph.edu/~fdominic/teaching/LDA/dental.dat`

   27 children, 16 boys and 11 girls, were observed at each ages 8, 10, 12, and 14 years. The data set has the following five columns:

   ```
   col1 = observation number
   col2 = child id number
   col3 = age
   col4 = response (distance)
   col5 = gender indicator (0=girl, 1=boy)
   ```

   At each time, the response, a measurement of the distance from the center of the pituitary to the pteryomaxillary fissure was made. Objectives were to learn whether there is a difference between boys and girls with respect to this measure and its change over time.

   Suppose you have **not been told** that the data represent observations on different subjects; rather, you are just told that columns 3 and 4 of this data set represent $n = 4 \times 27 = 108$ pairs of measurements on a covariate x (age) and response y (distance) respectively. You are also told

that the indicator in column 5 (0 or 1) means that the pairs of measurement are of two different types. Finally, you are told that interest focuses on fitting straight line regression models to these data, one for each type

$$\begin{aligned} y_j &= \beta_{00} + \beta_{01} x_j + \epsilon_j \; j = 1, \ldots, 108, \quad \text{for Type } 0 \\ y_j &= \beta_{10} + \beta_{11} x_j + \epsilon_j \; j = 1, \ldots, 108, \quad \text{for Type } 1 \end{aligned}$$

Estimate the parameters in each model, assuming that the all responses $y_j, \; j = 1, \ldots, n$ are independent with var$(y_j) = \sigma_0^2$ for Type 0 and var$(y_j) = \sigma_1^2$ for Type 1.

Now suppose instead that **you have been told** that the data do indeed represent repeated observations on each of 27 children, 11 girls and 16 boys, and that interest focuses on modeling the response (y_{ij}) as a linear function of age (t_{ij}) for boys and girls separately

$$\begin{aligned} y_{ij} &= \beta_{0B} + \beta_{1B} t_{ij} + \epsilon_{ij} \; i = 1, \ldots, 16 \; j = 1, 2, 3, 4 \quad \text{Boys} \\ y_{ij} &= \beta_{0G} + \beta_{1G} t_{ij} + \epsilon_{ij} \; i = 1, \ldots, 11 \; j = 1, 2, 3, 4 \quad \text{Girls} \end{aligned}$$

- Plot the data
- Fit the model with ordinary least squares (OLS)
- Fit the model with Generalized Least Square assuming several different models for the covariance matrix
- Which correlation model is the most appropriate?
- Fit the model using weighted least squares (WLS) and using maximum likelihood with this correlation model
- Compare the estimates and inferences from OLS and WLS
- Summarize your substantive findings as if for a journal.
  (*Two pages maximum – except figures – DO NOT include computer output*)