

Stat 565**Assignment 6****Fall 2005**

Reading Assignment: Diggle, Heagerty, Liang, and Zeger (DHLZ), Chapters 1, 3, 4, 5
The last three lectures in the course will cover material in chapters 7, 8 and 9 of DHLZ. There will be one more assignment that will be posted by December 1, along with solutions.

Written Assignment: Due in class, Thursday, December 1

Final Exam: Monday, December 12, 7:30-9:30 am

1. The data in the file dental.dat posted on the assignment section of the course web page are distances in millimeters from the center of the pituitary gland to the pteryomaxillary fissure for a sample of 11 girls and 16 boys. The measurements were taken every two years on each child at ages 8, 10, 12, and 14 years. Gender is the single covariate. (Potthoff and Roy, 1964).

There is one line of data for each subject. Each line of data has information in the following order:

id	Subject identification number
gender	0=girl 1=boy
Y_1	measurement at age 8 (in mm)
Y_2	measurement at age 10 (in mm)
Y_3	measurement at age 12 (in mm)
Y_4	measurement at age 14 (in mm)

A. Explore the data with plots and displays:

- Create a profile plot of the data. Use different types of lines for boys and girls. Describe any patterns that you see in the plot. What does this plot reveal about differences between boys and girls?
- Construct separate scatterplot matrices of the data for boys and girls. Compute the sample correlation coefficient for each scatterplot. Report any additional findings.

B. Fit linear models with linear, quadratic and cubic trends across time (age of the child) and interactions with gender. The most complex model would be

$$Y_{ijk} = \beta_{0j} + \beta_{1j}X_{ijk} + \beta_{2j}X_{ijk}^2 + \beta_{3j}X_{ijk}^3 + \varepsilon_{ijk}$$

where Y_{ijk} is the i -th measurement on the k -th child in the j -th gender group, and X_{ijk} is the age of the child when the measurement was taken. Assume an arbitrary (unstructured) covariance matrix for $(\varepsilon_{1jk} \ \varepsilon_{2jk} \ \varepsilon_{3jk} \ \varepsilon_{4jk})$, the random errors for the four measurements taken on a single subject. Is it appropriate to use the same covariance matrix for boys and girls? What is the simplest regression model that fits these data well.

- C. To search for an appropriate covariance model, fit the most complex regression model in part B,

$$Y_{ijk} = \beta_{0j} + \beta_{1j}X_{ijk} + \beta_{2j}X_{ijk}^2 + \beta_{3j}X_{ijk}^3 + \varepsilon_{ijk}$$

where Y_{ijk} is the i -th measurement on the k -th child in the j -th gender group, and

X_{ijk} is the age of the child when the measurement was taken. This model fits a different mean response for each of the eight combinations of gender and age. Consider the following covariance models for $(\varepsilon_{1jk} \ \varepsilon_{2jk} \ \varepsilon_{3jk} \ \varepsilon_{4jk})$, the random errors for the four measurements taken on a single subject:

- (1) Independence with homogeneous variances
- (2) Compound symmetry
- (3) Antedependence
- (4) Autoregressive(1)
- (5) Toeplitz model with homogeneous variances
- (6) Independence with heterogeneous variances
- (7) Equal correlations with heterogeneous variances
- (8) Antedependence with heterogeneous variances
- (9) AR(1) with heterogeneous variances
- (10) Toeplitz with heterogeneous variances
- (11) Unstructured covariance matrix

Which of these covariance models appears to be most appropriate? Give some justification for your selection.

- D. Considering the plots in part A and the range of models considered in parts B and C, what is the simplest model that adequately describes the data? Give some justification for your answer.

2. The data posted in the file cd4data.txt are from a randomized, double blind, study of AIDS patients with advanced immune suppression (CD4 counts less than 50 cells/mm³) (Henry, et al 1998). The patients in this study were randomized to either dual or triple combinations of HIV-1 reverse transcriptase inhibitors. The four treatment groups correspond to daily regimens containing 600 mg of zidovudine: zidovudine alternating monthly with 400 mg didanosine, zidovudine plus 2.25 mg of zalcitabine, zidovudine plus 400 mg of didanosine, or zidovudine plus 400 mg of didanosine plus 400 mg of nevirapine (triple therapy). There are 1313 lines on the data file. On each line, the variables appear in the following order:

id patient identification number
 group treatment group: coded 1 for triple therapy and 0 for dual therapy
 age baseline patient age in years
 sex coded 0 for female and 1 for male.
 week number of weeks since baseline
 logcd4 log(cd4 count + 1)

In this analysis we will combine the first three treatment groups into a single group (dual therapy) and make comparisons with the fourth treatment group (triple therapy). SAS code to read the data and combine the first three groups into a single treatment group is posted in the file CD4.sas.

- A. Use the loess smoothing technique or some other smoother to plot a smooth curve for $\log(\text{CD4 count} + 1)$ against time since baseline for the dual and triple therapy groups.
- B. Consider a piecewise linear model with a knot at 16 weeks. That is, the response for each patient can be described by an intercept and the slopes for two straight lines that are connected at 16 weeks. Letting t_{ij} denote the time since baseline for the j -th measurement on the i -th patient (with $t_{ij} = 0$ at baseline), and including random coefficients for each patient, the model can be expressed as

$$Y_{ij} = \beta_0 + \beta_1 t_{ij} + \beta_2 (t_{ij} - 16)_+ + \beta_4 z_i t_{ij} + \beta_5 z_i (t_{ij} - 16)_+ + b_{0i} + b_{1i} t_{ij} + b_{2i} (t_{ij} - 16)_+ + \varepsilon_{ij}$$

where

Y_{ij} is the j -th measurement of $\log(\text{CD4 count} + 1)$ on the i -th patient

$$(t_{ij} - 16)_+ = \begin{cases} t_{ij} - 16 & \text{if } t_{ij} \geq 16 \\ 0 & \text{if } t_{ij} < 16 \end{cases}$$

$$z_i = \begin{cases} 0 & \text{for dual therapy} \\ 1 & \text{for triple therapy} \end{cases}$$

$$\mathbf{b}_i = \begin{pmatrix} b_{0i} \\ b_{1i} \\ b_{2i} \end{pmatrix} \sim \text{NID} \left(\begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} g_{11} & g_{12} & g_{13} \\ g_{21} & g_{22} & g_{23} \\ g_{31} & g_{32} & g_{33} \end{pmatrix} \right) \quad \text{and} \quad \varepsilon_{ij} \sim \text{NID}(0, \sigma^2) \quad \text{and}$$

any ε_{ij} is independent of any b_i .

- i. Obtain REML estimates of g_{11} , g_{12} , g_{13} , g_{22} , g_{23} , g_{33} and σ^2 and the standard errors of the estimates. What do these estimates indicate?
- ii. Obtain estimates of $\beta_0, \beta_1, \beta_2, \beta_4, \beta_5$, standard errors of the estimates, and corresponding t-values and p-values. State your conclusions.
- iii. In the context of this model, test the null hypothesis that there are no differences between the dual and triple therapy treatment groups. State your conclusion.
- iv. Give interpretations of $\beta_0 + b_{0i}$, $\beta_1 + b_{1i}$, $\beta_1 + \beta_4 + b_{1i}$, $\beta_1 + \beta_2 + b_{1i} + b_{2i}$ and $\beta_1 + \beta_2 + \beta_4 + \beta_5 + b_{1i} + b_{2i}$ for this model.
- v. This model has the same intercept for each treatment group. Do you think this is reasonable? Explain...

C. Now consider a model that accounts for associations with gender and age. In particular, fit the following model:

$$Y_{ij} = \beta_0 + \beta_1 t_{ij} + \beta_2 (t_{ij} - 16)_+ + \beta_4 z_i t_{ij} + \beta_5 z_i (t_{ij} - 16)_+ + \beta_6 \text{Age}_i + \beta_7 \text{Gender}_i + b_{0i} + b_{1i} t_{ij} + b_{2i} (t_{ij} - 16)_+ + \varepsilon_{ij}$$

- i. Are there significant age or gender effects?
- ii. Do inferences about the parameters in model in part B change after adjusting for age and gender terms in the model in part C?

D. The models fit in parts Band C are only two of many possible models that could be fit to these data. If possible, identify and fit a model that is an improvement over the models in parts B and C or provides some insight about the data that is not provided by the models in parts B and C. Provide a short argument or justification for your decision.