

LDA 140.655  
HW1 - 02/04/2009  
Solutions given on 02/11/2009

Notes:

1. Homework is not required and does not count towards your final grade.
2. Please keep your write-ups short and do not include reams of analyses output.
3. Please attach an appendix of the (well documented) code you used (.do, .sas, or .R).

**Exploratory Data Analyses:**

Use the Nepal children's growth data to practice describing mean and association structures for an outcome (children's mean weight) over time (age).

1. Download the data from the course web site. Perform and describe standard quality control investigations on weight and age to make sure that the data have no implausible values.
2. Describe the structure of the data set, including distribution of observation times and a list of baseline and time-varying variables. Are the data balanced? Equally spaced?
3. Plot weight against age and obtain predictions from:
  - i. a linear regression
  - ii. a linear spline with knots at the 33rd and 67th percentile
  - iii. a cubic spline with knots at the 33rd and 67th percentile of age.

Plot the predicted curves from these three models and compare them. Interpret the output from the models in i. and ii.

4. Create the following spaghetti plots:
  - i. using all the data
  - ii. using a 20% sample
  - iii. a ZAP spaghetti plot

and include a smooth lowess fit on these plots.

5. Group the time (age) variable appropriately and create both a scatterplot matrix and a plot the ACF of the weight data. Include a confidence interval around the ACF estimate. Describe and interpret. (Remember the ACF is based on residuals; use the predictions from the linear regression above to obtain these residuals.)
6. Create and plot a variogram of the weight data. Use this variogram to create an ACF plot that is not based on the grouped time data. You may add this plot to the grouped data ACF from 5. or plot it separately. (Also use residuals as in 5.)

7. Write down a marginal linear regression model for weight on age using an independence model and a uniform (exchangeable) correlation structure. Adjust for gender, height, breastfeeding, mother's age, and mother's literacy, and use an appropriate function of age (as determined from 3.). Fit these models with ordinary least squares (OLS) and weighted least squares (WLS), respectively. Make a table comparing the intercept and slope estimates (and their standard errors) from the two models and interpret. Which results do you think are more appropriate? Does the average weight of children at a given age differ according to the mother's literacy?
8. Write down and fit the marginal model you would need to determine whether or not the rate of increase in weight differs according to whether or not the child's mother is literate.
9. Refit the model from 7. using a conditional, random effects model. What random effect specification do you need to include in the model to induce a uniform correlation structure?
10. Show that the OLS estimate of  $\beta$  is unbiased using matrix notation.
11. Derive the variance of the  $\beta_{OLS}$  estimate using matrix notation.

Biostat Students (and anyone else interested):

12. Show that  $\beta_{WLS}$  is optimally efficient when using  $W = V^{-1}$ . If  $V = R\sigma^2$ , do we need to know  $\sigma^2$  for this result to hold?