

PubH 8452 Spring 2008

Homework #1

Due Feb 18, 2008

1. In a randomized clinical trial, subjects receive either the active treatment or placebo after recording a baseline measurement, Y_{i0} . At the end of trial, the outcome is again recorded for each participant, Y_{i1} .

The goal of the study is to assess whether there is an impact on Y due to treatment. There are a number of potential analyses that can be proposed for this study design. Let TX denote the treatment assignment and let post be an indicator for the follow-up time. Assume that we have m subjects in each group. Assume that $\sigma^2 = \text{Var}(Y_{i0}) = \text{Var}(Y_{i1})$.

- (a) Consider the regression model:

$$E(Y_{ij}) = \beta_0 + \beta_1 \text{TX}_i + \beta_2 \text{post}_{ij} + \gamma \text{TX}_i \text{post}_{ij}.$$

Interpret the parameter γ .

- (b) If we use a repeated measures model with $\mathbf{Y}_i = (Y_{i0}, Y_{i1})^T$, assume an exchangeable covariance matrix, and assume normality, show that the MLE for γ is given by the difference of the differences: $\hat{\gamma}^{(1)}$ equals the means of $Y_{i1} - Y_{i0}$ for the treatment group minus the mean of $Y_{i1} - Y_{i0}$ for the placebo group.
- (c) Calculate the variance of $\hat{\gamma}^{(1)}$.
- (d) Another model uses the fact that groups were randomized to constrain the means at baseline, $E(Y_{i0} | \text{TX}_i = 0) = E(Y_{i0} | \text{TX}_i = 1)$. This model can be written as:

$$E(Y_{ij}) = \beta_0 + \beta_2 \text{post}_{ij} + \gamma \text{TX}_i \text{post}_{ij}.$$

Derive the MLE for γ in the constrained model. Denote this estimator as $\hat{\gamma}^{(2)}$. *Hint:* factor the likelihood $f(\mathbf{Y}_i) = f(Y_{i0})f(Y_{i1} | Y_{i0})$.

- (e) Calculate the variance of $\hat{\gamma}^{(2)}$.
- (f) The estimators $\hat{\gamma}^{(k)}$ are special cases of the general estimator:

$$\hat{\gamma}(\alpha) = \overline{Y_{i1}(\text{TX} = 1) - \alpha Y_{i0}(\text{TX} = 1)} - \overline{Y_{i1}(\text{TX} = 0) - \alpha Y_{i0}(\text{TX} = 0)}.$$

Another common proposed estimator is $\hat{\gamma}^{(0)}$ which simply compares the means at the follow-up time and ignores the baseline. Calculate the expectation of $\hat{\gamma}(\alpha)$ for any fixed α assuming the model in d holds. What is the variance of $\hat{\gamma}(\alpha)$? When is $\hat{\gamma}^{(1)}$ more precise than $\hat{\gamma}^{(0)}$? What is the optimal choice of α ?

- (g) Given the information above, suggest an appropriate analysis of “changes” when there are multiple follow-up measures.
2. Generate correlated data with different forms for the correlation (covariance) structure. Let the number of observations per subject being $n_i = 10$, evaluate at times $t_{ij} = j$ for $j = 1, 2, \dots, 10$. Use the mean model:

$$E(Y_{ij}|X_i) = \beta_0 + \beta_1 t_{ij}. \quad (1)$$

For each of the scenarios below generate vectors, $\mathbf{Y}_i = (Y_{i1}, Y_{i2}, \dots, Y_{i10})$ with the specified covariance structure. Generate data for $m = 25$ subjects for each scenario, using parameter value $\boldsymbol{\beta} = (10.0, 1.0)$.

- (a) **Random Intercepts:** To introduce correlation we assume that each subject has its own intercept. The complete model is given by:

$$\begin{aligned} Y_{ij} &= \beta_0 + \beta_1 t_{ij} + b_{0,i} + \epsilon_{ij}, \\ b_{0,j} &\sim \mathcal{N}(0, \tau^2), \\ \epsilon_{ij} &\sim \mathcal{N}(0, \sigma^2), \end{aligned}$$

where $b_{0,i}$ and ϵ_{ij} are mutually independent.

- i. Give the general form for the covariance matrix $\Sigma = \text{Cov}(\mathbf{Y}_i)$.
 - ii. Generate \mathbf{Y}_i and plot (lines) versus \mathbf{t}_i for $m = 25$ using $\sigma = 1.0$, $\tau = 1.0$.
 - iii. Generate \mathbf{Y}_i and plot (lines) versus \mathbf{t}_i for $m = 25$ using $\sigma = 1.0$, $\tau = 2.0$.
 - iv. Generate \mathbf{Y}_i and plot (lines) versus \mathbf{t}_i for $m = 25$ using $\sigma = 1.0$, $\tau = 5.0$.
- (b) **Random Intercepts and Slopes:** To introduce correlation we assume that each subject has its own intercept *and slope*. The complete model is given by:

$$\begin{aligned} Y_{ij} &= \beta_0 + \beta_1 t_{ij} + b_{0,i} + b_{1,i} t_{ij} + \epsilon_{ij}, \\ \mathbf{b}_j &\sim \mathcal{N}(\mathbf{0}, D), \\ \epsilon_{ij} &\sim \mathcal{N}(0, \sigma^2), \end{aligned}$$

where $\mathbf{b} = (b_{0,i}, b_{1,i})$ and ϵ_{ij} are mutually independent.

- i. Give the general form for the covariance matrix $\Sigma = \text{Cov}(\mathbf{Y}_i)$.
- ii. Generate \mathbf{Y}_i and plot (lines) versus \mathbf{t}_i for $m = 25$ using $\sigma = 1.0$, $D = \begin{pmatrix} 2.0 & 0 \\ 0 & 2.0 \end{pmatrix}$.
- iii. Generate \mathbf{Y}_i and plot (lines) versus \mathbf{t}_i for $m = 25$ using $\sigma = 1.0$, $D = \begin{pmatrix} 2.0 & -0.2 \\ -0.2 & 2.0 \end{pmatrix}$.
- iv. Generate \mathbf{Y}_i and plot (lines) versus \mathbf{t}_i for $m = 25$ using $\sigma = 1.0$, $D = \begin{pmatrix} 0.5 & 0 \\ 0 & 0.4 \end{pmatrix}$.

- (c) **Serial Correlation:** To introduce correlation we assume that each subject has his own “process” that is serially correlated. The complete model is given by:

$$\begin{aligned}
 Y_{ij} &= \beta_0 + \beta_1 t_{ij} + W_i(t_{ij}) + \epsilon_{ij}, \\
 W_i &\sim \mathcal{N}(0, D), \\
 \text{Var}[W_i(t_{ij})] &= \tau^2, \\
 \text{Cov}[W_i(t_{ij}), W_i(t_{ik})] &= \tau^2 \rho^{|t_{ij} - t_{ik}|}, \\
 \epsilon_{ij} &\sim \mathcal{N}(0, \sigma^2)
 \end{aligned}$$

where W_i and ϵ_{ij} are mutually independent.

- i. Give the general form for the covariance matrix $\Sigma = \text{Cov}(\mathbf{Y}_i)$.
 - ii. Generate \mathbf{Y}_i and plot (lines) versus t_i for $m = 25$ using $\sigma = 1.0$, $\tau = 2.0$ and $\rho = 0.7$.
 - iii. Generate \mathbf{Y}_i and plot (lines) versus t_i for $m = 25$ using $\sigma = 1.0$, $\tau = 2.0$ and $\rho = 0.9$.
 - iv. Generate \mathbf{Y}_i and plot (lines) versus t_i for $m = 25$ using $\sigma = 2.0$, $\tau = 2.0$ and $\rho = 0.9$.
3. The Six Cities Study of Air Pollution and Health was a longitudinal study designed to characterize lung growth as measured by changes in pulmonary function in children and adolescents, and the factors that influence lung function growth. A cohort of 13,379 children born on or after 1967 was enrolled in six communities across the U.S.: Watertown (Massachusetts), Kingston and Harriman (Tennessee), a section of St. Louis (Missouri), Steubenville (Ohio), Portage (Wisconsin), and Topeka (Kansas). Most children were enrolled in the first or second grade (between the ages of six and seven) and measurements of study participants were obtained annually until graduation from high school or loss to follow-up. At each annual examination, spirometry, the measurement of pulmonary function, was performed and a respiratory health questionnaire was completed by a parent or guardian.

The data set `fev1.dat` (on the class website) contains a subset of the pulmonary function data collected in the Six Cities Study. The data consist of all measurements of FEV1 (a measurement of lung function, total volume of air exhaled in the first second), height and age obtained from a randomly selected subset of the female participants living in Topeka, Kansas. The random sample consists of 300 girls, with a minimum of one and a maximum of twelve observations over time.

The list of variable is: Subject ID, Height, Age, Initial Height, Initial Age, Log(FEV1).

We are interested in characterizing long function growth in these children, in particular, how the age and initial height of the child can influence the growth.

- (a) Summarize the data, using exploratory data analysis techniques that were introduced in class. Explore the correlation structure, plot the variogram. Explain what you have learned.
- (b) For the next couple of questions, we only consider age as a covariate. Consider this model that has both age at entry (x_{i0}) and age since entry ($x_{ij} - x_{i0}$):

$$E(Y_{ij}) = \beta_0 + \beta_C x_{i0} + \beta_L (x_{ij} - x_{i0}).$$

Obtain GLS estimates of $\boldsymbol{\beta}$ using an independence model (the OLS estimator), in addition to estimates under both an exchangeable and AR(1) correlation model. Comment on the differences that are obtained. Given interpretation of $\hat{\beta}_C$ and $\hat{\beta}_L$ that could be understood by a clinician.

- (c) Now fit the GLS model $E(Y_{ij}) = \beta_0 + \beta_1 x_{ij}$ using OLS (independent), exchangeable and AR(1) correlation model. Comment on the differences between the $\hat{\boldsymbol{\beta}}$'s obtained with different correlation models and their relationship to the estimates above.
- (d) Which model(s) appear to be adequate for these data?