BIST P8157 : Analysis of Longitudinal Data
Homework #1
Due Oct.04, 2023

Question 1:

Suppose interest lies in characterizing the efficacy of treatment A versus treatment B with respect to some continuous outcome Y. Let Y_{ki} denote the response of the k^{th} study participant at the i^{th} time, where k=1, ... K, and i=1, 2. Furthermore, suppose that the variance of the response is σ^2 for i=1, 2, and that the correlation between the repeated measurements (within a study participant) is ρ.

(a) For each of the following designs derive the variance for the given estimate of the treatment effect:

(i) **Cross-sectional design:** A total of $K/2$ study participants are randomized to treatment A and $K/2$ study participants randomized to treatment B. All study participants are measured after they received treatment, and the treatment effect is estimated with $\hat{\gamma}_a = \overline{Y}_1^A - \overline{Y}_1^B$. Note, with this study design we have K total study participants who are each measured once.

(ii) **Longitudinal comparison of change from baseline:** A total of $K/2$ study participants are randomized to treatment A and $K/2$ study participants randomized to treatment B. All study participants are measured at baseline (time=1) prior to receiving treatment and after receiving treatment (time=2), and the treatment effect is estimated with $\hat{\gamma}_b = (\overline{Y}_1^A - \overline{Y}_0^A) - (\overline{Y}_1^B - \overline{Y}_0^B)$. Note, with this study design we have K total study participants who are each measured twice.

(iii) **Longitudinal comparison of treatment A and B (Crossover study)** All K study participants are observed on treatment A (time=1) AND treatment B (time=2), and the treatment effect is estimated with $\hat{\gamma}_c = \overline{Y}_1^A - \overline{Y}_2^B$. Note, with this study design we have K total study participants who are each measured twice.

(iv) **Longitudinal comparison of averages:** A total of $K/2$ study participants are randomized to treatment A and $K/2$ study participants randomized to treatment B. All study participants are measured twice on the randomized treatment assignment, and the treatment effect is estimated with $\hat{\gamma}_d = \overline{\overline{Y}}^A - \overline{\overline{Y}}^B$ where $\overline{\overline{Y}}^{tx}$ is the average of the $K/2$ study participant-specific averages under treatment tx. Note, with this study design we have K total study participants who are each measured twice.

(b) Assume we have a budget of $\$300,000$, and it costs \$500 each time the response is measured. How many people can be enrolled under each design? Calculate and compare the variances of the estimators, and discuss which you would choose for each of $\rho = \{0.2, 0.5, 0.8\}$ in order to minimize uncertainty (i.e. variance).

(c) Assume we have a budget of $\$300,000$, it costs \$250 to enroll someone into the study and then \$250 each time the response is measured. How many people can be enrolled under each design? Calculate and compare the variances of the estimators, and discuss which you would choose for each of $\rho = \{0.2, 0.5, 0.8\}$ in order to minimize uncertainty (i.e. variance).

Question 2:

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.

On the course website you'll find a dataset that contains a subset of the pulmonary function data collected in the Six Cities Study. The data consist of all measurements of FEV1, 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.

(a) Conduct an initial exploratory data analysis (EDA) for the Topeka data. In particular, consider the extent to which there are any unusual observations/outliers, as well as an initial exploration of the mean and dependence structure. For each component of your EDA, comment on how it would inform how you move forward. Report your results in a concise manner, using tables and/or figures. Note, what you submit for this may not be all of the EDA you conduct.

(b) Similar to what we did in class, consider the types of questions that one might be able to address with the Topeka data.

(c) Suppose that, instead of repeated measurements on each of the 300 girls, only a single measurement was obtained (say, at the start of the study). For any question that you considered in part (b), discuss the extent to which the question could be addressed using cross-sectional data albeit possibly with additional assumptions.

Question 3:

Consider the CD4+ cell count data we have been looking at in the notes. Specifically, consider the K^*=266 participants with at least one pre- and one post-seroconversion measurement (see slide 41 of the notes). As in the notes, restrict attention to those patients for whom the pre-seroconversion measurement was within 6 months of seroconversion. For the purposes of this analysis, take that measurement to be the measurement at time 0 (i.e baseline).

(a) Construct a 'Table 1' summarizing the sample on the basis of their covariates at baseline.

(b) Conduct a two-stage least squares analysis of the CD4+ cell count progression post-seroconversion. Towards this, at the first stage model each patients trajectory as a function of time since seroconversion. For these models you may consider the relationship to be linear or some other, more flexible, form. At the second stage, model the coefficients you obtained at the first stage as a function of baseline covariates. Report your results succinctly in the form of tables and/or figures. In addition, provide a brief summary of the results using language that would be suitable for a non-biostatistician collaborator.