ST 790, Homework 3
Spring 2017

1. Consider the data from the multicenter clinical trial in patients with age-related macular degeneration (AMD) in Problem 2 of Homework 3, which compared an experimental (active) treatment, interferon-α, with a placebo for the treatment of patients with AMD. Here, we will carry out analyses of the visual acuity outcomes on the N = 240 patients considered in that problem using proper multiple imputation. Recall that visual acuity was to be assessed at baseline (week 0) and then at clinic visits at 4, 12, 24, and 52 week, but that some participants have missing data due both to dropout and to intermittent missed visits. All have the baseline measure.

   The data are in the "wide" format in the files `armd.dat`, with missing values indicated using the SAS "." convention, and `armd.R.dat`, with missing values indicated by "`NA`." The columns are (1) patient ID number; (2) baseline lines of vision; (3)-(6) change from baseline lines of vision at 4, 12 24, and 52 weeks; (7)-(11) visual acuity at baseline, 4, 12, 24, and 52 weeks; (12) lesion grade; and (13) treatment, coded as 1 (placebo) and 4 (active treatment). We are again interested in an analysis of the visual acuity outcomes in columns (7)-(11).

   The full data are $Z = (A, Y_1, Y_2, Y_3, Y_4, Y_5)$, where Y_1 is visual acuity at baseline, and $Y_2, \dots, Y_5$ are visual acuity at weeks 4, 12, 24, and 52, and A is the treatment indicator such that $A = 0$ if a patient was assigned to placebo an $A = 1$ if assigned to active treatment. Letting $Y = (Y_1, \dots, Y_5)^T$ and treating the visual acuity measures as continuous, it is not unreasonable to assume that Y has an approximate multivariate normal distribution with possibly different mean vectors for each treatment.

   Assume the same full data model we considered in Problem 2 of Homework 2, namely

   $$Y_{ij} = \mu_{0j} + \beta_j A_i + \epsilon_{ij}, \quad \epsilon_i = (\epsilon_{i1}, \dots, \epsilon_{i5})^T \sim \mathcal{N}(0, \Sigma), \quad j = 1, \dots, 5, \ \ i = 1, \dots, N, \tag{1}$$

   where $\beta_j = \mu_{1j} - \mu_{0j}$, and μ_{0j} and μ_{1j} are the means at times $j = 1, \dots, 5$ for placebo and active treatment, respectively. Assume that Σ, which is common to both treatments, is a symmetric matrix with 15 distinct variance and covariance parameters (i.e., an "unstructured" covariance specification). Let $\theta = (\mu_{01}, \dots, \mu_{05}, \beta_1, \dots, \beta_5)^T$.

   In (a) and (b), you will use proper multiple imputation under the assumption of multivariate normality of $Y = (Y_1, \dots, Y_5)^T$ to obtain inference for the model (1) using both SAS and R. Thus, the "imputer's model" and the "analyst's model" are the same.

   Proper multiple imputation via MCMC is simulation-based, so, in contrast to optimization techniques like the EM algorithm, there is inherent variation in the results. We will examine the extent of this variation between the two different implementations.

   (a) Using SAS `proc mi`, obtain M = 10 imputed data sets using proper imputation. Fit the model (1) to each of the imputed data sets using `proc mixed`. Obtain the multiple imputation estimate $\widehat{\theta}$ of θ and associated standard errors using Rubin's variance formula. Also obtain the within and between imputation covariance matrices and calculate the full covariance matrix for $\widehat{\theta}$ based on Rubin's formula.

   (b) Using the R `norm` package, obtain M = 10 imputed data sets using proper imputation. Fit the model to each of the imputed data sets using `gls`. Obtain the multiple imputation estimate of θ and associated standard errors using Rubin's variance formula. Also obtain the within and between imputation covariance matrices and calculate the full covariance matrix for $\widehat{\theta}$ based on Rubin's formula.

(c) Compare the results in (a) and (b). How similar are they?

(d) From each analysis, construct an appropriate test statistic and obtain an associated p-value for testing the null hypothesis that $\beta_5 = 0$ versus the alternative $\beta_5 \neq 0$. Are the inferences qualitatively similar?

(e) Re-run the analyses in (a) and (b) with $M = 100$. Do you think $M = 10$ imputations are sufficient to achieve stable inferences using multiple imputation? Do you think choice of software implementation makes a difference in the inferences?

2. On the course webpage, you will find data from a study involving $N = 395$ patients suffering from chronic, disease-related pain. At baseline, all patients were started on analgesic treatment for pain and were to continue taking the treatment for 12 months. At baseline and then at 3, 6, 9, and 12 months, participants were to return to the clinic and provide an assessment of the efficacy of the treatment for controlling their pain using a five-point, ordinal"Global Satisfaction Assessment" (GSA) scale, where

$$\text{GSA} = \begin{cases} 1 & \text{very good} \\ 2 & \text{good} \\ 3 & \text{indifferent} \\ 4 & \text{bad} \\ 5 & \text{very bad.} \end{cases}$$

In addition to GSA measures, a number of baseline covariates were recorded.

In the files `analgesic.dat` and `analgesic.R.dat`, you will find the data in the "wide" format, with missing values indicated by the SAS "." and R "`NA`" conventions, respectively. In each file, the columns are (1) patient ID number; (2) age (years) at baseline; (3) weight (kg) at baseline; (4) genetic health index (scale of 0-100) at baseline; (5) physical functioning status (scale of 0-100) at baseline; (6)-(10) GSA at baseline and 3, 6, 9, and 12 months; and (11)-(15) dichotomized GSA at baseline (0 months) and 3, 6, 9, and 12 months, where this = 1 if GSA $\leq$ 3 and 0 otherwise. There are missing values for all of these variables except age, and there are both intermittent and monotone patterns of missingness of GSA values over time across patients.

In parts (a)-(e) of this problem, you will work with the dichotomized GSA outcomes in columns (11)-(15). In part (f), you will work with the ordered categorical GSA outcomes in columns (6)-(10).

The full intended data are $Z = (X_1, X_2, X_3, X_4, Y_1, Y_2, Y_3, Y_4, Y_5)$, where $X = (X_1, X_2, X_3, X_4)^T =$ (age, weight, genetic health index, physical functioning status)^T, and $Y = (Y_1, \dots, Y_5)^T$ is the vector of dichotomized GSA values at 0, 3, 6, 9, 12 months. Interest focuses on a model for $E(Y|x = x)$ of the form

$$\mu(x;\beta) = \{\mu_1(x;\beta), \dots, \mu_T(x;\beta)\},$$

where $\mu_j(x;\beta)$ depends on time t_j and x, that can be used to investigate whether or not the probability of experiencing pain in this population of pain sufferers changes over time when they are treated with the analgesic treatment. Because genetic health and physical functioning status at baseline are known to be associated with pain control, consider a model that "adjusts" for these variables, consider the model for pr$(Y_j = 1|X = x)$ of the form

$$\mu_j(x;\beta) = \text{expit}(\beta_0 + \beta_1 t_j + \beta_2 x_3 + \beta_3 x_4), \quad \beta = (\beta_0, \dots, \beta_3)^T, \tag{2}$$

where as usual $\text{expit}(u) = e^u/(1 + e^u)$. In (2), because the times are equally spaced, we take $t_j = (0, 1, 2, 3, 4)$ for $j = 1, \ldots, 5$, so that β_1 corresponds to the change in linear predictor over a period of 3 months.

If full data were available, it would be natural to fit this model by solving a GEE with some choice of working covariance matrix. Here, the variance of each Y_j is dictated by the Bernoulli distribution, so specifying a working covariance matrix boils down to specifying a working correlation structure. In this problem, we will take the working correlation structure to be completely unstructured, so that there are 10 distinct correlation parameters.

In (a) and (b), you will use multivariate imputation by chained equations (MICE)/fully conditional specification (FCS) to obtain the desired inferences on β in model (2) using both SAS and R. Note that, although the variables age and weight are not incorporated in model (2), they may still be useful for imputation. As in Problem 1, we will examine the extent of variation between the two implementation in SAS `proc mi` and the R `mice` package.

(a) Using SAS `proc mi`, obtain $M = 10$ imputed data sets using the MICE algorithm and full conditional specifications for each variable. Git (2) with unstructured working correlation structure to each using `proc genmod` (be sure to specify the `descending` option in the `model` statement so that SAS models $\text{pr}(Y_j = 1|X = x)$.) With `pid` and `timecls` as classification variables, the basic syntax is

```
proc genmod descending;
   class timecls pid;
   model gsa = time genhlth physfct / dist=bin link=logit;
   repeated subject=pid / type=un withinsubject=timecls;
```

Obtain the multiple imputation estimate $\widehat{\beta}$ of β and associated standard errors using Rubin's variance formula. Also obtain the within and between imputation covariance matrices and calculate the full covariance matrix for $\widehat{\beta}$ based on Rubin's formula.

(b) Using R `mice`, obtain $M = 10$ imputed data set using the MICE algorithm and full conditional specifications for each variable. Fit (2) with unstructured working correlation structure to each using the `gee` function. The basic syntax is

```
   gee(gsabin ~ time + genhlth + physfct,id=pid,family=binomial,
   corstr="unstructured",scale.fix=TRUE)
```

Obtain the multiple imputation estimate $\widehat{\beta}$ of β and associated standard errors using Rubin's variance formula. Also obtain the within and between imputation covariance matrices and calculate the full covariance matrix for $\widehat{\beta}$ based on Rubin's formula.

(c) Compare the results in (a) and (b).

(d) From each analysis, construct an appropriate test statistic and obtain an associated p-value for testing the null hypothesis that $\beta_j = 0$ versus the alternative $\beta_j \neq 0$, $j = 1, 2, 3$. Are the inferences qualitatively similar?

(e) Re-run the analyses in (a) and (b) with $M = 100$. Do you think $M = 10$ imputations are sufficient to achieve stable inferences using multiple imputation? Do you think choice of software implementation makes a difference in the inferences?

(f) An alternative imputation strategy is to impute the original ordinal GSA outcomes on the 5 point scale and ***then*** dichotomize the imputed ordinal values. In either SAS or R (your choice), carry out this alternative imputation approach to obtain $M = 10$ imputed data sets. Obtain the multiple imputation estimate $\widehat{\beta}$ of β and associated standard errors using Rubin's variance formula, the within and between imputation covariance matrices, the full covariance matrix for $\widehat{\beta}$ based on Rubin's formula, and the test statistics for $\beta_j = 0$ versus $\beta_j \neq 0$, $j = 1, 2, 3$.

Compare the results to those you obtained using the same implementation (SAS or R) in (a) or (b) and (d). Does imputation strategy make a qualitative difference in the results?