ST 790, Homework 4
Spring 2017

1. Recall the situation discussed in Section 5.1, in which the full data are $Z = (Y, V)$, Y can be missing, and V is always observed; and the goal is to estimate $\mu = E(Y)$. As remarked in Section 5.1, this is a nonparametric (so semiparametric) model. Define the iid observed data $(C_i, C_i Y_i, V_i)$, $i = 1, \ldots, N$, as in Section 5.1, and assume as in that section that the missingness mechanism is MAR, so that (5.2) holds and $C \perp\!\!\!\perp Y | V$. Let μ_0 be the true value of μ.

   In class, I remarked that it can be shown via semiparametric theory that the class of all consistent and asymptotically normal (regular, asymptotically linear, RAL) estimators for μ based on the observed data under these conditions has elements of the form

$$\widehat{\mu} = N^{-1} \sum_{i=1}^{N} \left\{ \frac{C_i Y_i}{\pi(V_i)} - \frac{C_i - \pi(V_i)}{\pi(V_i)} h(V_i) \right\}, \tag{1}$$

   where $h(V)$ is an arbitrary function of V. In Section 5.1, it is stated that the optimal such estimator, that will the smallest variance among all estimators in this class, is such that $h(V) = E(Y|V)$. In this problem, you will demonstrate this result.

   As we discussed on page 117 in Section 4.9, generically, RAL estimators $\widehat{\theta}$ for a parameter θ in a statistical model are characterized by their influence functions $\varphi(\cdot)$ and are such that if $N^{1/2}(\widehat{\theta} - \theta_0) \xrightarrow{\mathcal{L}} \mathcal{N}(0, \Sigma)$, Σ is the covariance matrix of the influence function.

   (a) Find the influence function of $\widehat{\mu}$ in (1), and show that the influence function has mean zero.

   (b) Find the variance of the influence function you found in (a) and show that it is minimized when $h(V) = E(Y|V)$.

   Hint: Write the variance of the influence function as the expectation of the influence function, squared. By cleverly adding and subtracting a term involving $E(Y|V)$ and using the MAR assumption, show that this expectation can be written as the sum of two uncorrelated terms, from which the result will follow.

2. Consider again the data from the multicenter clinical trial in patients with age-related macular degeneration (AMD) in Problem 3 of Homework 2 and Problem 1 of Homework 3, which compared an experimental (active) treatment, interferon-α, with a placebo for the treatment of patients with AMD. The study involved $N = 240$ participants.

   We will carry out analyses of the visual acuity outcomes using ***weighted generalized estimating equations*** as implemented in the experimental SAS procedure `proc gee`, assuming that missingness is MAR. Recall that visual acuity was to be assessed at baseline (week 0) and then at clinic visits at 4, 12, 24, and 52 weeks, 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 files `armd.hwk4.dat`, with missing values indicated using the SAS "." convention. 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).

The full data are $Z = (Y_1, Y_2, Y_3, Y_4, Y_5, A, V)$, where Y_1 is visual acuity at baseline, and $Y_2, \ldots, Y_5$ are visual acuity at weeks 4, 12, 24, and 52; A is the treatment indicator such that $A = 0$ if a patient was assigned to placebo an $A = 1$ if assigned to active treatment; and V is lesion grade, recorded on an ordinal scale of 1, 2, 3, 4. In the data set for this problem on the class webpage, A and V are available for all N individuals. (***Note:*** Be sure to download this data set again; one individual was missing lesion grade in the original data set, and a lesion value for this individual has been imputed so that this variable can be used in development of models for the cause-specific dropout hazards in (c) and (d) below.)

Letting $Y = (Y_1, \ldots, Y_5)^T$ and treating the visual acuity measures as continuous, consider a semiparametric model for Y_i of the form

$$E(Y_{ij}|A_i = a_i) = \mu_{0j} + \beta_j a_i, \quad j = 1, \ldots, 5, \ i = 1, \ldots, N, \tag{2}$$

where $\beta_j = \mu_{1j} - \mu_{0j}$; μ_{0j} and μ_{1j} are the means at times $j = 1, \ldots, 5$ corresponding to baseline and weeks 4, 12, 24, and 52 for placebo and active treatment, respectively; and $A_i = a_i$ is the treatment received by subject i. Thus, model (2) characterizes expected outcome given treatment assignment, making no further assumptions on the joint distribution of the visual acuity measures, in terms of the baseline covariate $X = A$ (treatment assignment) and vector of parameters $\beta = (\mu_{01}, \ldots, \mu_{05}, \beta_1, \ldots, \beta_5)^T$. This is, of course, the same model for expected outcome given treatment assumed in the previous homeworks, but without the assumption of multivariate normality.

If full data were available on all N individuals, we could fit (2) via a GEE as in (5.28) of the notes, adopting an assumed working covariance matrix. From previous analyses and on the basis of parsimony, we will take the working correlation structure to be ***compound symmetric*** (***exchangeable***) ***common*** to both treatment groups. This correlation matrix involves a single correlation parameter α, say, which is estimated via moment methods in GEE analyses. Of course, this assumption may be incorrect, so we will use the default robust (empirical) sandwich standard errors.

As you know from previous analyses, only 188 of the 240 individuals have full data; of the remaining 52 individuals, all but 8 exhibit ***monotone*** patterns of missingness (dropout). In (a)–(d), you will use GEE and WGEE methods to obtain inference on β for the model (2) using `proc gee` and, in (d), combining this with ***multiple imputation*** to account for the 8 individuals with ***nonmonotone*** missingness patterns.

In all analyses below with `proc gee`, you can specify the model (2) in the `model` statement in an identical fashion to that used with `proc mixed`, taking `week` to be a classification variable and using the `noint` option. In the `model` statement, the options `dist=normal` and `link=identity` yield (2) and the appropriate estimating equations; these are the defaults so can be omitted if you like.

(a) ***Naïve analysis 1.*** Using `proc gee`, fit (2) with the working compound symmetric (exchangeable) correlation structure to the ***available data***; that is, using all observed data on all $N = 240$ individuals.

(b) ***Naïve analysis 2.*** Identify and delete the 8 individuals with nonmonotone missingness patterns. Using `proc gee`, fit the same model as in (a) to the data from the 232 individuals who have full data or exhibit monotone (dropout) patterns. Thus, this is another ***available data*** analysis, restricted to individuals with monotone (or no) missingness.

(c) ***Weighted GEE analyses.*** Using the data on the 232 individuals with full data or exhibiting monotone patterns of missingness, use `proc gee` to carry out a WGEE analysis to

fit (2) assuming compound symmetric working correlation structure, using (i) ***subject level weighting*** and (ii) ***occasion level weighting***.

For (i) and (ii), models for the dropout hazards are required. Letting $H_j = (Y_1, \ldots, Y_j, A, V)$ for $j = 1, 2, 3, 4$, adopt the following models for $j = 2, \ldots, 5$:

$$\text{logit}\{\, \lambda_j(H_{j-1}; \psi_j)\,\} = \psi_{0j} + \psi_{1j} Y_{j-1} + \psi_2 I(V > 2); \qquad (3)$$

in (3), note that the parameter ψ_2, the coefficient of the dichotomized lesion grade, $I(V > 2)$, is ***common*** to all j. Model (3) thus allows the dropout hazards to depend on the most recent previous visual acuity measure and lesion grade at baseline.

(d) ***Combining WGEE with multiple imputation.*** The analyses in (a) and (b) are ***available data*** analyses that do not take account of the missingness. WGEE methods provide a principled approach under the assumption of MAR, but, as we have discussed, are feasible only in the case of ***monotone missingness*** (dropout). Thus, the analyses in (c) are restricted to the 232 individuals who have full data or exhibit such patterns.

Of course, excluding individuals from an analysis is always suspect. In this particular application, only 8 of the $N = 240$ (3.3%) individuals exhibited nonmonotone patterns and are excluded, so as a practical solution this may not be unreasonable. However, in most applications, as we have seen in other data sets, the proportion of individuals exhibiting nonmonotone patterns can be much ***higher***, rendering methods such as WGEEs infeasible.

As suggested by Molenberghs and Kenward (2007, Section 14.6.4), an *ad hoc* approach in this situation is to use ***multiple imputation*** to "fill in" only the intermittently missing values, thereby creating M imputed data sets that exhibit only ***monotone*** missingness patterns. The WGEE analysis of choice can then be carried out on each of the M imputed data sets, and the results combined in the usual way, with standard errors, etc, obtained using Rubin's variance formula.

(Of course, WGEE methods are not likelihood based methods, so the theory for multiple imputation ***does not*** strictly apply. However, the use of non-likelihood based analyses with multiple imputation is widespread and seems to work well in practice, despite the absence of theoretical justification. The following procedure based on ***monotone imputation*** likewise has no theoretical justification.)

SAS `proc mi` implements such ***monotone imputation*** in the case where all variables are assumed to be ***multivariate normal*** for imputation purposes, filling in only intermittently missing values and yielding imputed data sets that exhibit only monotone missingness. This is accomplished by invoking the `impute=monotone` option in the `mcmc` statement.

Carry out such an analysis as follows:

(i) Using `proc mi` and taking $Y_1, \ldots, Y_5$ to be multivariate normal, create $M = 10$ imputed data sets in which the intermittent missing values for the 8 individuals are filled in using the `monotone` option as above.

(ii) For each of the M data sets, carry out the WGEE analyses as in (c) using `proc gee` using both subject level and occasion level weighting and assuming the compound symmetric working correlation structure, with dropout hazards modeled as in (3). For each of these analyses, output the estimate of β and its robust asymptotic covariance for use by `proc mianalyze`; this can be accomplished using the following `ods` statement in the call to `proc gee`:

```
ods output GEEEmpPEst=xxxparms ParmInfo=xxxinfo GEERCov=xxxcovb;
```

Here, "xxx" is whatever you choose to distinguish the output data sets for each of the subject level and occasion level analyses. To obtain the robust asymptotic covariance matrices, you will need to include the `ecovb` option in the `repeated` statement. See the `proc gee` documentation for details.

(iii) For each type of WGEE analysis (subject/occasion level), combine the results for the *M* data sets using `proc mianalyze` to obtain an estimate of β and standard errors for each component of β via Rubin's formula. ***Note:*** If you specified `week` as a `class` variable in `proc gee`, you will need to use the `classvar` option in the `parms` option:

```
proc mianalyze parms(classvar=level)=xxxparms parminfo=xxxinfo
   covb=xxxcovb wcov bcov;
```

(e) For each analysis in (a) –(d), write down the estimate and associated standard error for β_5 (difference in treatment means at 52 weeks) and obtain an appropriate test statistic and associated p-value for testing the null hypothesis that $\beta_5 = 0$ versus the alternative $\beta_5 \neq 0$.

Compare the results across the analyses. Can you think of an explanation for similarities and differences in the inferences across approaches? Which approach, if any, do you feel comfortable recommending?