ST 790, Homework 2
Spring 2018

1. *Finite-sample properties of estimators in a population-averaged model.* In this problem, you will design and carry out a simulation study to evaluate the performance of several estimators for β in a population-averaged model and the relevance of the large-sample theory approximations to the sample distributions of the estimators. A review of the basic steps in a typical simulation study is in Appendix D of the notes.

   Your simulation study will be based on the following situation, modeled after that of the dental data. The true model from which you will generate data is as follows. Let $m = 27$, and, for the ith individual, $i = 1, \ldots, m$, observations Y_{ij} are to be taken at the same times t_j, $(t_1, \ldots, t_4) = (8, 10, 12, 14)$ for all i. Let $g_i = 0$ for $i = 1, \ldots, 11$ and $g_i = 1$ for $i = 12, \ldots, m$. Then take $\boldsymbol{x}_i$ to contain the among-individual covariate g_i and the four time points. As in (5.14), take

$$E(Y_{ij}|\boldsymbol{x}_i) = \beta_{0,B} g_i + (1 - g_i)\beta_{0,G} + \{\beta_{1,B} g_i + \beta_{1,G}(1 - g_i)\} t_j, \qquad j = 1, \ldots, 4, \tag{1}$$

   where $\beta = (\beta_{0,G}, \beta_{1,G}, \beta_{0,B}, \beta_{1,B})^T$. Further, take

$$\text{var}(\boldsymbol{Y}_i|\boldsymbol{x}_i) = \boldsymbol{V}_i(\boldsymbol{x}_i) = \sigma^2 \boldsymbol{\Gamma}_i(\alpha), \tag{2}$$

   where $\boldsymbol{\Gamma}_i(\alpha)$ is a (4×4) unstructured covariance matrix as in (2.24) with $n_i = 4$, so that α is a $n_i(n_i - 1)/2 \times 1$ vector of distinct correlation parameters.

   Let the true distribution of $\boldsymbol{Y}_i$ given $\boldsymbol{x}_i$ (so given g_i) be multivariate normal with these moments, with true values of β, σ^2 and α be

$$\beta_0 = (17.3, 0.48, 16.3, 0.78)^T, \sigma_0^2 = 5,$$

   and

$$\alpha_0 = (0.4, 0.2, 0.0, 0.7, 0.2, 0.8)^T,$$

   so that the true overall correlation matrix is

$$\boldsymbol{\Gamma}_i(\alpha_0) = \begin{pmatrix} 1 & 0.4 & 0.2 & 0 \\ 0.4 & 1 & 0.7 & 0.2 \\ 0.2 & 0.7 & 1 & 0.8 \\ 0 & 0.2 & 0.8 & 1 \end{pmatrix} \tag{3}$$

   for any individual i, and the true overall covariance matrix is then

$$\sigma_0^2 \boldsymbol{\Gamma}_i(\alpha_0).$$

   Thus, the true population mean vector is given by (1) evaluated at β_0, and the true overall covariance matrix is given by (2) evaluated at σ_0^2 and α_0.

   (a) For each of $S = 1000$ generated data sets from this scenario, do the following:

   (1) Estimate β in mean model (1) using *maximum likelihood* under normality as in Section 5.3, taking $\text{var}(\boldsymbol{Y}_i|\boldsymbol{x}_i)$ to be of the form (2), so assuming the *correct* overall covariance structure, where in fact the correlation matrix is *known* and equal to (3), so that $\alpha = \alpha_0$

is *known*, but σ^2 is not known. Note that you will have to estimate σ^2 (but not $\boldsymbol{\alpha}$) jointly with $\boldsymbol{\beta}$. Obtain both the usual, *model-based standard errors* based on the approximate sampling distribution in (5.68) and (5.69) and the *robust standard errors* based on that in (5.81) and (5.82).

Note: Fitting the model with this covariance structure with $\boldsymbol{\alpha} = \boldsymbol{\alpha}_0$ known using `gls()` in R can be accomplished using the specification

```
correlation=corSymm(value=c(0.4,0.2,0.0,0.7,0.2,0.8),fixed=TRUE,
                    form ~ 1 | as.factor(child))
```

(2) Estimate $\boldsymbol{\beta}$ in (1) using *maximum likelihood* under normality as in Section 5.3, taking var($\boldsymbol{Y}_i|\boldsymbol{x}_i$) to be of the form (2), so, as in 1 above, assuming the *correct* overall covariance structure, where now both σ^2 and $\boldsymbol{\alpha}$ are not known. Thus, you will have to estimate $\boldsymbol{\alpha}$ and σ^2 jointly with $\boldsymbol{\beta}$. Obtain both the *model-based standard errors* based on the approximate sampling distribution in (5.68) and (5.69) and the *robust standard errors* based on that in (5.81) and (5.82).

(3) Estimate $\boldsymbol{\beta}$ in mean model (1) using *ordinary least squares* (OLS), so pretending that all Y_{ij} given g_i are mutually independent for all i and j; that is, wrongly taking var($\boldsymbol{Y}_i|\boldsymbol{x}_i$) = $\sigma^2\boldsymbol{I}_4$. Obtain both the usual, *model-based standard errors* based on the approximate sampling distribution in (5.68) and (5.69) and the *robust standard errors* based on that in (5.81) and (5.82).

Note: This model can be fitted using `gls()` by using the specification

```
correlation=corCompSymm(value=0,fixed=TRUE,form ~ 1 | as.factor(child))
```

(4) Estimate $\boldsymbol{\beta}$ in (1) using *maximum likelihood* under normality as in Section 5.3, taking var($\boldsymbol{Y}_i|\boldsymbol{x}_i$) to be σ^2 times a *compound symmetric* correlation matrix as in (2.27) of the notes; that is, assuming this *incorrect* overall covariance structure. Again, you will have to estimate $\boldsymbol{\alpha}$ and σ^2 jointly with $\boldsymbol{\beta}$. Obtain the *model-based standard errors* based on the approximate sampling distribution in (5.68) and (5.69) and the *robust standard errors* based on that in (5.81) and (5.82).

(5) Estimate $\boldsymbol{\beta}$ in (1) using *maximum likelihood* under normality as in Section 5.3, taking var($\boldsymbol{Y}_i|\boldsymbol{x}_i$) to be σ^2 times a *AR(1)* correlation matrix as in (2.28) of the notes; that is, assuming this *incorrect* overall covariance structure. Again, you will have to estimate $\boldsymbol{\alpha}$ and σ^2 jointly with $\boldsymbol{\beta}$. Obtain the *model-based standard errors* based on the approximate sampling distribution in (5.68) and (5.69) and the *robust standard errors* based on that in (5.81) and (5.82).

Thus, (1) is an "ideal" situation in which the overall covariance matrix is *known* up to the scale parameter σ^2, which would of course be unlikely in practice, so serves as a "benchmark" against which realistic specifications that could be made in practice can be judged. (2) is the situation in which the form of the overall covariance matrix is *correctly specified*, and (3)–(5) are cases in which the form of the overall covariance matrix has been *incorrectly specified*.

(b) For each of (1)–(5), save the S estimates of $\boldsymbol{\beta}$ and the associated standard errors, and then do the following:

- Calculate the Monte Carlo mean of the S estimates of each of the components of $\boldsymbol{\beta}$ for each of (1)–(5).

- Calculate the Monte Carlo standard deviation (SD) of the S estimates of each of the components of β for each of (1)–(5).
- Calculate the Monte Carlo average of the estimated standard errors (SEs) of both types (model-based and robust) for each of the components of β for each of (1)–(5). Using these and the Monte Carlo SDs, for each component of β, calculate the ratio

$$\text{Monte Carlo SD}/\text{Average of Monte Carlo SEs.}$$

- Calculate the Monte Carlo mean square error (MSE) based on the S estimates of β for each component of β for each of (1)–(5). From these, for each component of β, calculate the Monte Carlo relative efficiency of each of (2)–(5) relative to (1); that is, for each component, calculate

$$\text{MSE}\{\text{estimator from model (1)}\}/\text{MSE}\{\text{estimator from model (a)}\}, \quad (a) = (2) - -(5),$$

so that a ratio < 1 reflects inefficiency of the estimator from model (a) relative to the "benchmark" estimator (1) for which the correlation matrix is known.

(c) Comment on the implications of your results in (b) regarding consistency of the estimators for β in (1)–(5).

(d) Comment on the implications of your results in (b) regarding the validity of the model-based standard errors in terms of reflecting the true sampling variation of the estimators for the components of β and the ability of robust standard errors to correct for misspecification of the covariance structure.

(e) Comment on the implications of your results in (b) regarding the relative efficiency of the estimators for β in (1)–(5). Do your results agree with what you expect from the theory?

2. *Another perspective on REML.* In this problem, you will see another way to interpret REML. With $\boldsymbol{X}$, $\boldsymbol{V}$, $\boldsymbol{X}_i$, and $\boldsymbol{V}_i$, $i = 1, \ldots, m$, as in the notes, so suppressing dependence on $\tilde{\boldsymbol{x}}$ and $\boldsymbol{\xi}$ for brevity, write

$$\widehat{\beta} = (\boldsymbol{X}^T \boldsymbol{V}^{-1} \boldsymbol{X})^{-1} \boldsymbol{X}^T \boldsymbol{V}^{-1} \boldsymbol{Y}$$

and define

$$\boldsymbol{H}_i = \boldsymbol{V}_i^{-1/2} \boldsymbol{X}_i (\boldsymbol{X}^T \boldsymbol{V}^{-1} \boldsymbol{X})^{-1} \boldsymbol{X}_i^T \boldsymbol{V}_i^{-1/2},$$

and

$$\boldsymbol{V}_{\xi_k i} = \partial / \partial \xi_k \, \boldsymbol{V}_i(\boldsymbol{\xi}, \boldsymbol{x}_i).$$

(a) Derive expressions for

(i) the conditional expectation of

$$(\boldsymbol{Y}_i - \boldsymbol{X}_i \widehat{\beta})^T \boldsymbol{V}_i^{-1} \boldsymbol{V}_{\xi_k i} \boldsymbol{V}_i^{-1} (\boldsymbol{Y}_i - \boldsymbol{X}_i \widehat{\beta})$$

with respect to $\tilde{\boldsymbol{x}}$

(ii) the final term in the kth REML estimating equation,

$$\text{tr} \left[\left\{ \sum_{i=1}^{m} \boldsymbol{X}_i^T \boldsymbol{V}_i^{-1} \boldsymbol{X}_i \right\}^{-1} \sum_{i=1}^{m} \boldsymbol{X}_i^T \boldsymbol{V}_i^{-1} \boldsymbol{V}_{\xi_k i} \boldsymbol{V}_i^{-1} \boldsymbol{X}_i \right],$$

in terms of $\boldsymbol{H}_i$.

(b) Using your results in (a), show that the conditional expectation given $\tilde{\boldsymbol{x}}$ of a summand of the maximum likelihood estimating equations (5.35) for ξ is not equal to zero when $\widehat{\beta}$ is substituted for β but that of a summand of the REML estimating equations (5.56) is equal to zero.

(c) Now suppose that $\boldsymbol{Y}_i$ is a scalar outcome, Y_i, say, for $i = 1, \ldots, m$ ($n_i = 1$), so that $\boldsymbol{X}_i$ is a $(1 \times p)$ vector depending on $\boldsymbol{x}_i$ and thus $\boldsymbol{X}$ is a $(m \times p)$ matrix with ith row $\boldsymbol{X}_i$ of full column rank p; and $\boldsymbol{V}_i(\xi, \boldsymbol{x}_i) = V_i(\xi, \boldsymbol{x}_i) = \sigma^2$, a scalar constant variance, so that $\xi = \sigma^2$. That is, consider the simple special case of the general population-averaged model corresponding to the usual linear model for independent (across i) scalar outcomes with constant variance.

(i) Letting $\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T$ be the so-called "hat matrix," use your results in (a) to show that the usual least squares residual $r_i = Y_i - \boldsymbol{X}_i\widehat{\beta}$ has conditional variance

$$\sigma^2(1 - h_{ii}), \tag{4}$$

where h_{ii} is the ith diagonal element of $\boldsymbol{H}$, often referred to as the *leverage* associated with the ith observation due to the location of $\boldsymbol{X}_i$ in the covariate space, with large h_{ii} indicating that $\boldsymbol{X}_i$ is remote in the covariate space from the bulk of the observations and thus could have large influence in determining the fit $\widehat{\beta}$.

(ii) Use your result in (b) to find an expression for the REML estimator $\widehat{\sigma}^2$ for σ^2.

Interpretation: Thus, in the case of the usual linear model with constant variance, (4) is the motivation for the use of *studentized* residuals

$$\frac{r_i}{\widehat{\sigma}(1 - h_{ii})^{1/2}}$$

in place of raw or standardized residuals in regression diagnostics.

Accordingly, in the general population-averaged model, the REML estimator for ξ can be interpreted correcting for bias in estimation of ξ based on "residual vectors" ($\boldsymbol{Y}_i - \boldsymbol{X}_i\widehat{\beta}$) due to "high leverage" individuals.

3. *Effectiveness of weight loss programs, continued.* Recall the data from the weight loss study described in Problem 1 of Homework 1, involving 100 male subjects randomly assigned to maintain their current eating and exercise habits (the control condition, coded as 1), to follow a rigorous program of modified diet and exercise (coded as 2), or to adopt a modified diet (coded as 3). All participants were weighed at baseline (month 0) and then again at months 3, 6, 9, and 12.

   The data set has the following columns:

1	subject ID number
2-6	weight (lbs) at months 0, 3, 6, 9, and 12
7	program (coded as 1, 2, 3 as above)

   The investigators are interested in the same sorts of questions as in Problem 5 of Homework 1, which are stated more fully and precisely as: following questions:

   (i) Are the weight loss programs effective at lowering weight?

   (ii) What is the nature of the pattern of change in weight under the three conditions (control and the two weight loss programs), and is this pattern different among the three conditions?

(iii) What is the rate of change of weight for each program? Is it constant over the study period?

(iv) What is the mean weight at 12 months for each of the programs? Do these means differ?

Taking a population-averaged perspective on these questions, using methods in Chapter 5 of the notes, carry out analyses to address these questions and write a brief report summarizing what you did and the results, following the basic outline for writing a data analysis report in Appendix F of the course notes. As in the guidelines there, be sure to describe how you formalized the questions of interest within the framework of these models and interpret the results in the context of the subject matter. Comment on any limitations or concerns you might have and on how confident you feel about the reliability of the inferences and conclusions.

Please turn in your code and output along with your report (you can edit the output to include only the portions that pertain directly to your report).

4. *Cervical dystonia clinical trial.* Cervical dystonia, also called spasmodic torticollis, is a painful condition in which the neck muscles contract involuntarily, causing the head to twist or turn to one side or to uncontrollably tilt forward or backward. In the file `cdystonia.dat` on the class webpage, you will find the data from a randomized clinical trial conducted to evaluate the effectiveness of 4 weeks of treatment with botulinum toxin type B (BotB) in patients suffering from cervical dystonia. In the study, $m = 109$ subjects were randomized to a four-week regimen of placebo, 5000 units (U) of BotB, or 10000 U of BotB. The outcome, total score on Toronto Western Spasmodic Torticollis Rating Scale (TWSTRS), a measure of severity, pain, and disability of cervical dystonia, where high scores mean greater impairment, was intended to be ascertained for each subject at baseline (week 0), and weeks 2, 4, 8, 12, and 16. As can be seen from inspection of the data file, not all subjects appeared at all intended visits for TWSTRS score to be measured, with several dropping out before completing the study. The file `cdystonia.dat` contains the *observed* data. The data also include the age and gender of each subject.

   The data set has the following columns:

   1 patient ID
   2 week
   3 treatment group (0=placebo, 5000=5000 U, 10000=10000 U)
   4 age (years)
   5 gender (0=female, 1=male)
   6 TWSTRS score (the outcome)

   Because some patients are missing data from some of the intended visits, be careful to account for this and to state any assumptions necessary to justify your analysis.

   As noted above, each subject received his/her assigned treatment for only the first part of the study; treatment was started at baseline and at week 4. The investigators expected that, in the groups where subjects were given 5000 U or 10000 U BotB for the first 4 weeks, TWSTRS scores would decline, reflecting the efficacy of BotB for alleviating impairment. From week 4 onward, they then expected TWSTRS scores to show a rebound . They were especially interested in comparing the steepness of the decline from week 0 to week 4 ("phase 1") and then the steepness of the rebound after treatment was withdrawn ("phase 2," from week 4 to week 16). Their hope was that, the higher the dose of BotB given in first four weeks,

the steeper the decline would be in phase 1, reflecting better ability to alleviate impairment; and that the rebound would be shallower (less steep) the higher the dose, reflecting a better sustained effect the higher the dose even after it was withdrawn.

More precisely, the investigators are interested in the following questions:

(i) Are there differences in the steepness of the decline in phase 1 among the three treatments? What is the rate of change of TWSTRS for each treatment in phase 1?

(ii) Are there differences in the steepness of the rebound in phase 2 among the three treatments? What is the rate of change of TWSTRS for each treatment in phase 2?

(iii) What is the mean TWSTRS score for each group at the end of the study (week 16)? Are their differences in mean score at the end of the study among the treatments?

(iv) Is there evidence that TWSTRS score at baseline is associated with age or gender? Is the steepness of the decline in phase 1 associated with age or gender?

Taking a population-averaged perspective on these questions, using methods in Chapter 5 of the notes, carry out analyses to address these questions and write a brief report summarizing what you did and the results, following the basic outline for writing a data analysis report in Appendix F of the course notes. As in the guidelines there, be sure to describe how you formalized the questions of interest within the framework of these models and interpret the results in the context of the subject matter. Comment on any limitations or concerns you might have and on how confident you feel about the reliability of the inferences and conclusions.

Please turn in your code and output along with your report (you can edit the output to include only the portions that pertain directly to your report).