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 of estimators for $\beta$ in a population-averaged model and of 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 $i$th individual, $i = 1, \ldots, m$, observations $Y_{ij}$ are to be taken at times $t_j$, $(t_1, \ldots, t_4) = (8, 10, 12, 14)$. Let $g_i = 0$ for $i = 1, \ldots, 11$ and $g_i = 1$ for $i = 12, \ldots, m$. Then take $x_i$ to contain the among-individual covariate $g_i$ and the four time points. As in (5.14), take

$$E(Y_{ij}|x_i) = \beta_{0,0}g_i + (1 - g_i)\beta_{0,0} + \{\beta_{1,1}g_i + \beta_{1,1}(1 - g_i)\}t_j,$$

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

$$\text{var}(Y_i|x_i) = \Gamma_i(\alpha_i),$$

where $\Gamma_i(\alpha_i)$ is the $(4 \times 4)$ AR(1) correlation matrix as in (2.28) with $n_i = 4$. Let the true distribution of $Y_i$ given $x_i$, (so given $g_i$) be multivariate normal with these moments, with true values of $\beta$, $\sigma^2$ and $\alpha$ be

$$\beta_0 = (17.3, 0.48, 16.3, 0.78)^T, \sigma^2_0 = 5, \alpha_0 = 0.75.$$

Thus, the true population mean vector is given by (1) evaluated at $\beta_0$, and the true overall covariance matrix is given by (2) evaluated at $\sigma^2_0$ and $\alpha_0$ and has a AR(1) correlation structure.

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

1. Estimate $\beta$ 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 $\text{var}(Y_i|x_i) = \sigma^2 I_4$. 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).

2. Estimate $\beta$ in (1) using maximum likelihood under normality as in Section 5.3, taking $\text{var}(Y_i|x_i)$ to be of the form (2); that is, assuming the correct overall covariance structure. Note that you will have to estimate $\alpha$ and $\sigma^2$ jointly with $\beta$. 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 $\beta$ in (1) using maximum likelihood under normality as in Section 5.3, taking $\text{var}(Y_i|x_i)$ to be $\sigma^2$ 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 $\alpha$ and $\sigma^2$ jointly with $\beta$. 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).
4. Estimate $\beta$ in (1) using *maximum likelihood* under normality as in Section 5.3, taking $\text{var}(Y_i | x_i)$ to be a *completely unstructured* covariance matrix, so that the variances at each age are also taken to differ. Again, you will have to estimate the parameters in this covariance model jointly with $\beta$. 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).

(b) For each of 1–4, save the $S$ estimates of $\beta$ and the associated standard errors, and then do the following:

- Calculate the Monte Carlo mean of the $S$ estimates of $\beta$ for each of 1–4.
- Calculate the Monte Carlo standard deviation of the $S$ estimates of $\beta$ for each of 1-4.
- Calculate the Monte Carlo average of the estimated standard errors of both types for each of 1–4.
- Calculate the Monte Carlo mean square error (MSE) based on the $S$ estimates of $\beta$ for each of 1–4, and for 1, 3, and 4, calculate the ratio of the MSE for these estimators relative to that for 2. That is, the ratio of MSE for the estimators based on an incorrect specification of the overall covariance structure relative to the MSE for the estimator based on the true structure, so $\text{MSE (incorrect model)}/\text{MSE (correct model)}$.

(c) Comment on the implications of your results in (b) regarding consistency of the estimators for $\beta$ in 1–4.

(d) Comment on the implications of your results in (b) regarding the relative efficiency of the estimators for $\beta$ in 1–4. Does this agree with what you expect from the theory?

(e) Comment on the implications of your results in (b) regarding the validity of the model-based standard errors and the ability of robust standard errors to correct for misspecification of the covariance structure.

2. Demonstrate that a summand in the REML estimating equation

$$
\sum_{i=1}^{m} \left( (Y_i - X_i\hat{\beta})^T V_i^{-1}(\xi, x_i) \{\partial / \partial \xi_k V_i(\xi, x_i)\} V_i^{-1}(\xi, x_i)(Y_i - X_i\hat{\beta}) 
- \text{tr} \left[ V_i^{-1}(\xi, x_i) \{\partial / \partial \xi_k V_i(\xi, x_i)\} \right] \right) 
+ \text{tr} \left[ \left( \sum_{i=1}^{m} X_i^T V_i^{-1}(\xi, x_i) \right) \right]^{-1} \sum_{i=1}^{m} X_i^T V_i^{-1}(\xi, x_i) \{\partial / \partial \xi_k V_i(\xi, x_i)\} V_i^{-1}(\xi, x_i) X_i \right] = 0
$$

in (5.56) has expectation zero given $\tilde{x}$, thus showing that the REML estimating equations for $\xi$ are unbiased.

3. Protein diets and growth of chicks, continued. Recall the study of the effects of protein diets on the growth of newly hatched chicks from Homework 1. The data set in that homework was actually an abridged version of the actual data set. In this problem we will work with another version of this data set with a different set of time points and where some chicks are missing some observations. The data set is available in the file *allchicks.dat*, and has the following columns:
1 weight (g) 
2 days since birth 
3 chick ID number 
4 diet (coded as 1, 2, 3, 4) 

Note that some chicks are missing data on some of the days. Thus, be careful to account for this and to state any assumptions necessary to justify your analysis.

The investigators are interested in the following questions: (i) Is the pattern of change of chick weight over the period of the study different among the four diets? (ii) Is there evidence that the rate of change of weight is not constant over the study period for any of the diets? (iii) What is the rate of change of weight over the study period for each diet? (iv) What is the mean weight at the end of the study (20 weeks) for each diet? 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. Be sure to describe how you interpreted and formalized the questions of interest from the PA perspective. Comment 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 and embed it in your report if you like).

4. Age-related macular degeneration clinical trial. In the file armd.dat on the class webpage, you will find the data from the clinical trial in patients with age-related macular degeneration (AMD) discussed in Section 5.6 of the notes. There were 240 patients, each of whom was randomized to receive either a placebo or active treatment (interferon-α). The outcome we will consider is the visual acuity outcome discussed in the notes, which was intended to be ascertained for each subject at weeks 0 (baseline), 4, 12, 24, and 52. As discussed in the notes, some subjects dropped out or did not attend all visits, so there are missing outcomes for some of them. The file armd.dat contains the observed data. The data include a categorical variable, baseline lesion grade, which is an assessment of the severity of the macular lesion(s), where 1 is least severe and 4 is most severe.

The data set has the following columns:

1 patient ID 
2 lesion grade (1, 2, 3, 4 representing increasing severity) 
3 treatment group (0 = placebo, 1 = interferon-α) 
4 time (weeks) 
5 visual acuity 

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

The investigators are interested in the following questions: (i) Do either of the treatments improve visual acuity or at least arrest its decline? (ii) Is the pattern of change in visual acuity over the study period different between placebo and interferon-α? How? (iii) What is the rate of change in visual acuity over the study period? (iv) Is baseline visual acuity associated with lesion severity? (v) Is the pattern of change in visual acuity over the study period different depending on lesion severity for one or both treatments?

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 summa-
rizing what you did and the results. Be sure to describe how you interpreted and formalized the questions of interest from the PA perspective. Comment 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 and embed it in your report if you like).