ST 790, Homework 1
Spring 2018

1. *Effectiveness of weight loss programs.* In the file `weightloss.dat` on the class website, you will find data from a study to evaluate two weight loss programs. 100 male subjects were recruited to participate and were 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)

   In this problem you will investigate the suitability of different statistical model specifications for these data.

   Let Y_{ij} denote the weight measured on subject i at time t_{ij}, where $i = 1, \dots, m = 100$, $j = 1, \dots, n_i = 5$ for all i, and $t_{ij} = 0, 3, 6, 9, 12$ for all i. Analogous to the individual-specific model (2.13) for the dental study data given on page 38 of the notes, consider the subject-specific model

   $$Y_{ij} = \beta_{0i} + \beta_{1i} t_{ij} + e_{ij}. \qquad (1)$$

   (a) Suppose we assume that, in (1),

   $$\begin{array}{lll} \beta_{0i} = \beta_{0,1} + b_{0i}, & \beta_{1i} = \beta_{1,1} & \text{if } i \text{ followed the control program} \\ \beta_{0i} = \beta_{0,2} + b_{0i}, & \beta_{1i} = \beta_{1,2} & \text{if } i \text{ followed the diet+exercise program} \\ \beta_{0i} = \beta_{0,3} + b_{0i}, & \beta_{1i} = \beta_{1,3} & \text{if } i \text{ followed the diet alone program} \end{array}$$

   where $E(b_{0i}|p_i) = 0$, $\text{var}(b_{0i}|p_i) = D$, p_i is the program followed by subject i (=1 for control, =2 for diet+exercise, and =3 for diet alone), and b_{0i} are independent across i. Suppose further that e_{ij} are independent of each other and of b_{0i} for all (i, j) with $E(e_{ij}|p_i) = 0$ and $\text{var}(e_{ij}|p_i) = \sigma^2$.

   Give expressions for $\text{var}(Y_{ij}|p_i)$, $\text{cov}(Y_{ij}, Y_{ij'}|p_i)$, and $\text{corr}(Y_{ij}, Y_{ij'}|p_i)$ implied by the model (1) with the above specifications for β_{0i} and β_{1i}.

   (b) Comment on the similarities and differences between the model in (a) and the model (3.1) underlying the univariate repeated measures analysis of variance methods in Chapter 3.

   (c) Construct graphical and numerical summaries to explore the nature of population mean response and among- and within-subject variance covariance and correlation in the weight loss data. In light of these analyses, comment on the suitability of the various models in Section 2.5 for representing/approximating the overall pattern of correlation.

   (d) Based on your analyses in (c), are the models in (a) and/or (b) reasonable representations of these data? Explain your answer.

   Note: These data are in the "wide" format of one record per individual. To reconfigure them in the "long" format of one record per observation in R, you can use the `reshape()` function, `melt()` from the `reshape2` package, or a variety of other tools. For example, using `reshape()`,

```
long.data <- reshape(wide.data,varying=c("month0","month3","month6",
   "month9","month12),v.names="weight",idvar="id",times=c(0,3,6,9,12),
   timevar="month",direction="long")
long.data <- long.data[order(long.data$id,]    #  reorder
```

Examples of doing this in SAS are on the course website.

2. *Nonlinear models.* On page 43 of the notes, we indicate that the distinction between subject-specific and population-averaged models is critical when the models involved are *nonlinear*. In this problem, you will examine this in a simple situation with no within- or among-individual covariates.

Consider a study of growth where the response is a continuous measure of growth. Suppose that the inherent trajectory for individual i is represented by the logistic growth model in (1.1) of the notes, which we write here in a different parameterization as

$$\mu_i(t) = \frac{\beta_{1i}}{1 + \exp\{-(\beta_{3i} + \beta_{2i} t)\}}, \quad t \geq 0, \tag{2}$$

where

$$\beta_{1i} = \beta_1, \quad \beta_{2i} = \beta_2 + b_i, \quad \beta_{3i} = \beta_3; \tag{3}$$

$\boldsymbol{\beta} = (\beta_1, \beta_2, \beta_3)^T$ is a fixed parameter vector; and

$$b_i \sim \mathcal{N}(0, D), \quad D > 0. \tag{4}$$

In (2), $\beta_{1i} > 0$ is the asymptotic growth value as time $t \to \infty$, which together with $\beta_{3i} > 0$ characterizes the growth value at time $t = 0$; and the growth rate constant β_{2i} describes the change of growth value over time. In (2)-(4), assume $\beta_2 \gg 0$ so that pr($\beta_{2i} \leq 0$) is negligible.

Thus, in (2) and (3), growth value at $t = 0$ and asymptotic growth value are taken to be the same for all individuals, while β_{2i}, and thus the nature of change in growth value over time, is individual-specific. In (2), $\mu_i(t)$ depends on β_{2i}, and thus on b_i, in a nonlinear fashion.

(a) In (2.7) of the notes, we represented $\mu_i(t)$ as

$$\mu_i(t) = \mu(t) + \mathcal{B}_i(t),$$

where $\mu(t)$ is the overall population mean response at time t and $E\{\mathcal{B}_i(t)\} = 0$, so that $E\{\mu_i(t)\} = \mu(t)$, where for $\mu_i(t)$ in (2) this expectation is thus with respect to the distribution of b_i. Under the model in (2)–(4), analytical calculation of this expectation is intractable; however, it can be approximated by exploiting the fact that

$$\frac{1}{(1 + e^{-u})} \approx \Phi(u/c), \quad c = \frac{15\pi}{16\sqrt{3}} \approx 1.7, \tag{5}$$

where $\Phi(\cdot)$ is the cumulative distribution function of the standard normal distribution. (5) is a common so-called probit approximation to the logistic cumulative distribution function. The approximation (5) is excellent for a wide range of u values.

Using (5), derive an approximate expression for the overall population mean $\mu(t)$.

(b) As we will see in later chapters, when the model for $\mu_i(t)$ is nonlinear in individual-specific parameters like β_{2i} (and thus depends in a nonlinear way on random effects like b_i), it is common to approximate $\mu_i(t)$ by taking a linear Taylor series about $b_i = 0$ (its mean). Do this and identify the resulting approximation to $\mu(t)$.

(c) Comparing the (cruder) approximation to $\mu(t)$ in (b) to the (more direct) approximation in (a), comment on the implications of using that in (b) for valid inference on $\boldsymbol{\beta}$.

3. Consider the situation in Chapter 3, where $Y_{h\ell j}$ is the response on individual h in group ℓ at time j, $h = 1, \dots, r_\ell$; $\ell = 1, \dots, g$; and $j = 1, \dots, n$. The mean response for group ℓ at time j is $E(Y_h \ell j) = \mu_{\ell j}$. Suppose that $g = 3$ and $n = 5$ and that the time points are equally spaced. Define the matrix $\mathcal{M}$ as in (3.17).

   (a) Write down matrices $\boldsymbol{C}$ and $\boldsymbol{U}$ in terms of which the null hypothesis that the pattern of change is the same in each group can be written as

   $$H_0 : \boldsymbol{C}\mathcal{M}\boldsymbol{U} = \mathbf{0}.$$

   If this is not possible, explain why not.

   (b) Write down a matrix $\boldsymbol{U}$ in terms of which the null hypothesis that the rate of change of mean response is constant for all groups can be written as

   $$\mathcal{M}\boldsymbol{U} = \mathbf{0}.$$

   If this is not possible, explain why not.

4. *Effectiveness of weight loss programs, continued.* Consider again the data from the weight loss study described in Problem 1.

   The investigators were interested in the following questions:

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

   (ii) Is the pattern of change in weight different among the three conditions (control and the two weight loss programs)?

   Using models and methods in Chapter 2 and 3 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. (You may wish to incorporate analyses your carried out in Problem 1 in your report.) 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).

5. *Cholesterol Study.* In the file `cholesterol.dat` on the class website, you will find data from the National Cooperative Gallstone Study (NCGS), where one of the major interests was to study the safety of the drug chenodiol for the treatment of cholesterol gallstones. In the study, subjects were randomly assigned to high-dose (750 mg per day), low-dose (375 mg per day), or placebo. Our data set includes only data from subjects assigned to the high-dose (coded as 1) and placebo (coded as 2) groups. The NCGS investigators conjectured that chenodiol would dissolve gallstones but in doing so might increase levels of serum cholesterol. To

evaluate this conjecture, serum cholesterol (mg/dL) was measured at baseline (month 0) and at 6, 12, 20, and 24 months of follow-up .

The data set has the following columns:

1	treatment, coded as above (1=high-dose chenodiol, 2=placebo)
2	subject ID
3-7	serum cholesterol (mg/dL) at months 0, 6, 12, 20, and 24

The investigators were interested in the following questions:

(i) Does high-dose chenodiol lead to an increase in serum cholesterol levels relative to placebo?

(ii) In particular, is the pattern of change of cholesterol different under high-dose chenodiol and placebo?

Using models and methods in Chapter 2 and 3 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).