ST 790, Homework 1
Spring 2017

1. In ***EXAMPLE 1*** of Chapter 1 of the notes, it is shown at the bottom of page 22 that the complete case estimator for the mean μ of an outcome Y given in (1.18) under MNAR satisfies
$$\widehat{\mu}^c \xrightarrow{p} \frac{E\{Y\pi(Y)\}}{E\{\pi(Y)\}},$$
where $\pi(y) = \text{pr}(R = 1|Y = y)$. At the top of page 23, it is noted that, if $\pi(y)$ is an increasing function of y, so that the probability of observing Y increases with the value of Y then
$$\frac{E\{Y\pi(Y)\}}{E\{\pi(Y)\}} > \mu.$$
Provide an argument justifying this claim.

2. *Missingness in regression analysis.* In this exercise, you will write a program to carry out a simulation study to investigate the analytical results presented in ***EXAMPLE 2*** on pages 28 - 30 of the notes. Here, the full data are $Z = (Y, X)$, where Y is a scalar outcome and X is a vector of covariates, and we are interested in estimation of β in a model $\mu(x;\beta)$ for $E(Y|X = x)$. Each of Y and X is either observed or is missing, so that $R = (R_1, R_2)$, where $R_1 = 1$ if Y is observed and $R_1 = 0$ if Y is missing, and similarly for R_2 and X. As in the notes, let $\pi\{r, (Y, X)\} = \text{pr}(R = r|Y, X\}$, so that the probability of observing a complete case ($r = (1, 1)$) is $\pi\{(1, 1), (Y, X)\}$. We will consider the three cases in the notes and the claims regarding consistency of the least squares estimator $\widehat{\beta}^c$ for β based only on the complete cases, obtained by solving (1.27):

   (i) When $\pi\{(1, 1), (Y, X)\} = \pi\{(1, 1), X\}$, so depends only on X, the claim is that $\widehat{\beta}^c$ is consistent under MNAR and MAR.

   (ii) When $\pi\{(1, 1), (Y, X)\}$ depends on Y, the claim is that $\widehat{\beta}^c$ is inconsistent in general under MNAR and MAR.

   (iii) When $\pi\{(1, 1), (Y, X)\}$ does not depend on (Y, X), which is the case if the missingness mechanism is MCAR, $\widehat{\beta}^c$ is consistent.

   The objective of a simulation is to approximate the properties of an estimator by generating some large number S independent data sets from a known situation and computing the estimator for each data set. The sample mean of the estimates over all S data sets is an estimate of the mean of the sampling distribution of the estimator; similarly, the standard deviation of the estimates over the S data sets is an estimate of the standard deviation of the sampling distribution (how good these quantities are at capturing the true features of the sampling distribution obviously depends on the size of S).

   If an estimator is consistent, we expect, for reasonably large sample sizes where we might expect large sample theory to be a good approximation, the sample mean of the S estimates to be very close to the true value of the parameter being estimated. For a regression parameter β with p elements and estimator $\widehat{\beta}$, we can assess this by computing for each element β_k, $k = 1, \dots, p$, the *Monte Carlo mean* $S^{-1}\sum_{s=1}^{S}\widehat{\beta}_{s,k}$ and *Monte Carlo bias*
$$S^{-1}\sum_{s=1}^{S}\widehat{\beta}_{s,k} - \beta_{0,k},$$

where $\widehat{\beta}_s$ is the estimate calculated for the sth data set and β_0 is the true value of β used to generate the data sets. We would expect the Monte Carlo mean/bias to be close to the true value/zero for a consistent estimator. Likewise, the *Monte Carlo standard deviation*, the usual standard deviation of the S estimates, reflects the variation in the sampling distribution of $\widehat{\beta}$.

(a) In your favorite programing language, write a program to carry out a simulation of the performance of the complete case least squares estimator $\widehat{\beta}^c$ under a known data generation scenario. Your program should have the following features:

- Each of S data sets should contain N observations, where $S = 1000$ and $N = 200$. For each data set, for each $i = 1, \dots, N$, generate independently $X_{i1} \sim \mathcal{N}(10, 3^2)$ (standard deviation 3) and X_{i2} as Bernoulli with $\text{pr}(X_2 = 1) = 0.4$. Then generate

$$Y_i = \beta_1 + \beta_2 X_{i1} + \beta_3 X_{i2} + \epsilon_i,$$

  where ϵ_i are independent $\mathcal{N}(0, 8^2)$ (standard deviation 8), and $\beta = (20, 5, -5)^T$. Thus, the true value $\beta_0 = (20, 5, -5)^T$, and the model is $\mu(x; \beta) = \beta_1 + \beta_2 x_1 + \beta_3 x_2$.
- You will run your program three times, once under each of scenarios (i)–(iii). In each case, for each of the S data sets and for each $i = 1, \dots, N$, generate an indicator of observing a complete case, i.e., observing (Y_i, X_{i1}, X_{i2}), by calculating

$$\pi_i = \exp(U_i)/\{1 + \exp(U_i)\}, \quad U_i = \psi_1 + \psi_2 X_{i1} + \psi_3 X_{i2} + \psi_4 Y_i + \psi_5 X_{i1} Y_i + \psi_6 X_{i2} Y_i,$$

  and then generating C_i, the indicator of whether or not the ith observation is a complete case ($R_i = (1, 1)$), as Bernoulli with $\text{pr}(C_i = 1 | X_{i1}, X_{i2}, Y_i) = \pi_i$. For each case, take ψ as follows:

$$\begin{array}{ll} \text{Case (i)} & \psi = (2, -0.025, 0.5, 0, 0, 0)^T \\ \text{Case (ii)} & \psi = (6, 0, 0, -0.075, -0.003, 0.05)^T \\ \text{Case (iii)} & \psi = (0.5, 0, 0, 0, 0, 0)^T \end{array}$$

  Thus, you should write your program so that it can be run for each case by simply changing the value of ψ.
- For each case, for each of the S data sets of size N, calculate two least squares estimates of β in the above linear model: (1) the *ideal* full data estimate $\widehat{\beta}^f$, say, that could be calculated if the full data were available, which will serve as a "gold standard;" and (2) the complete case estimate $\widehat{\beta}^c$ based only on the observations with $C_i = 1$. For each data set $s = 1, \dots, S$, save the values of $\widehat{\beta}^f_s$ and $\widehat{\beta}^c_s$ and the proportion p^c_s out of the N intended observations that were complete cases.
- For each case, calculate the Monte Carlo mean, bias, and standard deviation of the S estimates and the average of the proportions p^c.
- Note: Most languages, including SAS and R, allow generation of random deviates from normal and binomial distributions. For each run of your program, set the "starting seed" for the random number generation so that you can reproduce the results rather than let the random number generator start where it likes. The random number generator will update the seed automatically and internally each time a random number generation function is called, so you need only to set the seed initially at the beginning of your program. Random number generators work by generating a sequence of random deviates that should be approximately independent. Thus, you do not want to change the seed yourself over the course of the simulation once it starts.

(b) Discuss whether or not the results are in line with the analytical results presented in the notes. Do you think that mindlessly proceeding with the default (in SAS and other software) complete case analysis when there are missing data is to be recommended?

3. *Missing at random and dropout (monotone missingness).* As discussed in the notes, *missing at random* (MAR) corresponds to the situation where the probability of missingness patterns depends only on the data that are observed. In the notation in the notes, then, MAR corresponds to assuming that $\text{pr}(R = r|Z) = \text{pr}(R = r|Z_{(r)})$. When missingness is *monotone*, as in the case of *dropout* in a longitudinal study, on page 16 of the notes we defined for full data $Z = (Z_1, \dots, Z_T)$ collected over T times t_j, $j = 1, \dots, T$, with corresponding missingness indicators $(R_1, \dots, R_T)$, $D = 1 + \sum_{j=1}^{T} R_j$, where $D = j$ indicates that $Z_1, \dots, Z_{j-1}$ are observed and $Z_j, \dots, Z_T$ are missing, so $D = T + 1$ means that the full data are observed. As on page 16, write $Z_{(j)} = (Z_1, \dots, Z_{j-1})$ when using this notation.

   In this situation, it is convenient to represent the probability of missingness through the *hazard function*
$$\lambda_j(Z) = \text{pr}(D = j|D \geq j, Z), \quad j = 1, \dots, T + 1.$$
   (Note that $\lambda_j(Z) = 1$ for $j = T + 1$.)

   (a) Show that there is a one-to-one relationship between
$$\text{pr}(D = j|Z) \quad \text{for } j = 1, \dots, T + 1$$
   and $\lambda_j(Z)$ for $j = 1, \dots, T$. That is, for any $j = 1 \dots, T$, $\lambda_j(Z)$ can be expressed in terms of $\text{pr}(D = j'|Z)$, $j' = 1, \dots, j$, and, for any $j = 1, \dots, T + 1$, $\text{pr}(D = j|Z)$ can be expressed in terms of $\lambda_{j'}(Z)$, $j' = 1, \dots, j$.

   (b) Show that MAR holds, i.e.,
$$\text{pr}(D = j|Z) = \text{pr}(D = j|Z_{(j)})$$
   for all $j = 1, \dots, T + 1$ if and only if
$$\lambda_j(Z) = \lambda(Z_{(j)})$$
   for all j.

4. *More on longitudinal data and dropout.* In ***EXAMPLE 3*** on pages 30 - 32 of the notes, we discussed the situation where $Z = (Y_1, \dots, Y_T)$, where Y_j is a scalar outcome recorded at time t_j, $j = 1, \dots, T$. In this problem, we will consider a more general situation where, along with Y_j, a scalar covariate X is also recorded at baseline ($j = 1$) and does not change with j. Letting $Y = (Y_1, \dots, Y_T)^T$, suppose that 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 only on time t_j and x. As on page 18 of the notes, this is a *semiparametric* model, as only this conditional expectation, and not the entire distribution of Z, is modeled parametrically, with the rest of the joint density of Y and X left unspecified.

   In this situation, analogous to (1.32) of the notes, a standard approach to estimation of β is to solve a GEE of the form
$$\sum_{i=1}^{N} \mathcal{D}^T(X_i; \beta)\mathcal{V}^{-1}(X_i; \beta) \begin{pmatrix} Y_{i1} - \mu_1(X_i; \beta) \\ \vdots \\ Y_{iT} - \mu_T(X_i; \beta) \end{pmatrix} = 0, \tag{1}$$

where $\mathcal{D}^T(x;\beta)$ is a $(p \times T)$ matrix of partial derivatives of $\mu(x;\beta)$ multiplied by a $(T \times T)$ *working covariance matrix* $\mathcal{V}(x;\beta)$, both of which depend on the observation times and possibly on x. A GEE like (1) can be solved, and an estimate $\widehat{\beta}$ obtained, using, for example, SAS `proc genmod` or the R packages `gee` or `geepack`.

In the case of dropout at time $j+1$, we observe $(Y_1, \dots, Y_j)^T$ and X only, and, analogous to (1.33), using the dropout notation, it is common to base estimation of $\widehat{\beta}$ on these available data, in which case one would solve

$$\sum_{i=1}^{N} \left\{ \sum_{j=1}^{T} I(D_i = j+1)\, \mathcal{D}_j^T(X_i;\beta)\mathcal{V}_j^{-1}(X_i;\beta) \begin{pmatrix} Y_{i1} - \mu_1(X_i;\beta) \\ \vdots \\ Y_{ij} - \mu_j(X_i;\beta) \end{pmatrix} \right\} = 0, \qquad (2)$$

where $\mathcal{D}_j^T(x;\beta)$ $(p \times j)$ and $\mathcal{V}_j(x;\beta)$ $(j \times j)$ are the corresponding submatrices of $\mathcal{D}^T(x;\beta)$ and $\mathcal{V}(x;\beta)$, In SAS, `proc genmod` will automatically disregard missing values (indicated using "." in the data set) and solve (2), and a function like R `gee` can be instructed to do the same (where missing values are indicated using "`NA`" in the data set).

The working covariance model is a "guess" at the covariance matrix var$(Y|X)$ and is meant to capture possible correlation over the t_j among observations on the same individual. It may or may not be correct; accordingly, it is important to use so-called "sandwich" or "robust/empirical" standard errors that take account of this possibility rather than "model-based/naive" standard errors that assume it is correct. If you take ST 732 and ST 793, you will learn all about this.

In this problem, we will use a world-famous data set to examine the consequences of solving (2) when there is dropout/monotone missingness that is MAR, and we will learn how to implement (1) and (2) in SAS and/or R. The so-called *orthodontic* or *dental data* were presented by Pothoff and Roy (1964), and are from a study conducted at UNC- Chapel Hill involving 27 children, 16 boys and 11 girls. On each child, the distance (mm) from the center of the pituitary to the pterygomaxillary fissure ("dental distance" Y) was made at ages 8, 10, 12, and 14 years of age. There were no missing data; i.e., full data were observed. Interest focused on comparing the growth patterns between boys and girls. As you will evaluate in (a) below, letting $(t_1, \dots, t_4) = (8, 10, 12, 14)$ and $X = 0$ for girls and $X = 1$ for boys, a possible model for $E(Y|X = x)$ takes

$$\mu_j(x;\beta) = \beta_1 + \beta_2 x + \beta_3 t_j + \beta_4 x t_j, \qquad (3)$$

so that the mean is represented as a straight line for each gender with intercept and slope β_1 and β_3 for girls and $(\beta_1 + \beta_2)$ and $(\beta_3 + \beta_4)$ for boys. The difference in growth patterns is thus reflected in β_4, which in (3) represents the difference in slope between boys and girls.

In the file `dental.dat` on the class website, you will find the original, full data. The file has 5 columns: (1) observation number, (2) child number (1-27), (3) age, (4) distance measurement, and (5) indicator of gender (0=girl, 1=boy).

(a) In your favorite programming language, calculate the mean dental distance for girls and boys separately at each age, and plot and connect the means at each t_j for girls and boys on the same axes. Is (3) a reasonable model based on the visual evidence?

(b) Using SAS or R (your choice), fit the model (3) using the full data by solving (1) assuming that the working covariance matrix is a constant variance σ^2 times an *exchangeable* or *compound symmetric* correlation matrix with 1s on the diagonal and the same parameter ρ in all off-diagonal positions. Here, ρ thus represents a pattern of correlation that is the same regardless of how close or far apart two observations are in time (age).

This may be accomplished as follows. In SAS using `proc genmod`:

```
data dent1; infile 'dental.dat';
  input obsno child age distance gender;
run;

proc genmod data=dent1;
    class child;
    model distance = gender age gender*age /  dist=normal link=identity;
    repeated subject=child / type=exch corrw modelse;
run;
```

In interpreting the results, you should look at the table `Analysis Of GEE Parameter Estimates, Empirical Standard Error Estimates`; the estimate of σ is the `Scale` in the `Model-Based Standard Error Estimates` table.

In R using the function `gee` (you may have to `install.packages(‘‘gee’’)`; alternatively, if you are familiar with GEEs and prefer to use `geepack`, feel free):

```
dent1 <- read.table("dental.dat")
colnames(dent1) <- c("num","child","age","distance","gender")

library(gee)
gee.full <- gee(distance ~ age + gender+ age*gender, id=child, family=gaussian,
                corstr="exchangeable",data=dent1)
summary(gee.full)
```

In interpreting the results, you should use the `Robust S.E.`; the estimate of σ^2 is the `Estimated Scale Parameter`.

In both cases, the estimates of β and ρ are self-explanatory.

(c) In the files `dental_dropout_sas.dat` and `dental_dropout_R.dat`, which are in the same format as the original data, dropout according to a MAR mechanism has been artificially induced; the first file indicates a missing value by "." and the second by "`NA`" for use with SAS and R, respectively.

Using SAS or R (your choice), fit the model (3) using these available data by solving (1) assuming that the working covariance matrix is the same as in (b).

This may be accomplished as follows. In SAS using `proc genmod`:

```
data dent2; infile 'dental_dropout_sas.dat';
  input obsno child age distance gender;
run;

proc genmod data=dent2;
    class child;
    model distance = gender age gender*age /  dist=normal link=identity;
    repeated subject=child / type=exch corrw modelse;
run;
```

In R, using `gee`, the code is

```
dent2 <- read.table("dental_dropout_R.dat")
colnames(dent2) <- c("num","child","age","distance","gender")

library(gee)
gee.avail <- gee(distance ~ age + gender+ age*gender, id=child, family=gaussian,
                 corstr="exchangeable",na.action=na.omit,data=dent2)
summary(gee.avail)
```

(d) Compare the estimates of β between the analyses based on the full and available data. Do they coincide with the implications of the analytical arguments in the notes regarding the behavior of the estimator for β under MAR using available data?

(e) As noted above, interest focuses on the null hypothesis $H_o : \beta_4 = 0$ versus the alternative $H_1 : \beta_4 \neq 0$. Comment on the implications of not having full data for making inference on β_4.

5. *Last Observation Carried Forward.* As discussed in Section 2.3 of the notes, the LOCF approach is controversial, and the analytical results there suggest that it need not lead to a consistent estimator for the mean at the last time point, $E(Y_T)$, in a longitudinal study with dropout. To examine possible problems associated with LOCF in a specific situation, consider again the dental data. In your favorite programming language, do the following using the dental data:

   (a) Based on the full data, calculate the sample means and standard errors at $t_T = 14$ for the boys and girls separately, and carry out a usual two-sample test (e.g., a t-test) of whether or not the means differ.

   (b) From the data set for which dropout according to a MAR mechanism was artificially induced (in the files `dental_dropout_sas.dat` and `dental_dropout_R.dat`), create a new data set in which the LOCF convention is used to impute all missing observations Y_{iT} (hint: It may be easier to do this if you first reconfigure the data set to be in the form of 1 data record/individual, so that the first line is “1 0 21 20 21.5 NA” or “1 0 21 20 21.5 .”). Based on the resulting data set, calculate the sample means and standard errors at $t_T = 14$ for the boys and girls separately, and carry out the same test you performed in (a).

   (c) Comment on the results.