5 Methods Based on Inverse Probability Weighting Under MAR

The likelihood-based and multiple imputation methods we considered for inference under MAR in Chapters 3 and 4 are based, either directly or indirectly, on integrating over the distribution of the missing data. In this chapter, we consider a different class of methods that instead uses directly the missingness mechanism itself. In ***EXAMPLE 1*** of Section 1.4 of Chapter 1, we introduced the idea of ***inverse probability weighting of complete cases***, which forms the basis for this general class of methods.

The use of inverse probability weighting was proposed in the context of surveys in a famous paper by Horvitz and Thompson (1952). In a landmark paper decades later, Robins, Rotnitzky, and Zhao (1994) used the ***theory of semiparametrics*** to derive the class of all consistent and asymptotically normal ***semiparametric estimators*** for parameters in a ***semiparametric full data model*** when data are MAR and to identify the ***efficient*** estimator in the class. It turns out that estimators in this class can be expressed as solutions to ***estimating equations*** that involve inverse probability weighting. A detailed account of this theory and of these estimators is given in Tsiatis (2006).

From a practical point of view, this theory is the basis for the class of estimators for full data model parameters using what are often called in the context of missing data problems ***weighted estimating equations***, or ***WGEEs***. We begin by returning to the simple case of estimation of a single mean as in ***EXAMPLE 1*** of Section 1.4 to illustrate the fundamental features of weighted estimating equations, including the notion of ***double robustness***, and then generalize to more complex models.

We continue to assume that MAR holds.

5.1 Inverse probability weighted estimators for a single mean

SIMPLE INVERSE PROBABILITY WEIGHTED ESTIMATORS: Recall the situation in ***EXAMPLE 1*** of Section 1.4, in which the full data are $Z = (Z_1, Z_2) = (Y, V)$, where Y is some scalar outcome of interest, and V is a set of additional variables. The objective is to estimate

$$\mu = E(Y).$$

Note that this is a ***nonparametric*** (and thus semiparametric) model, as we have specified nothing about the distribution of Z.

As noted in Section 1.4, the obvious estimator for μ if we had a sample of full data (Y_i, V_i), $i = 1, \ldots, N$, would be the sample mean of the Y_i

$$\widehat{\mu}^{full} = N^{-1} \sum_{i=1}^{N} Y_i.$$

Note that $\widehat{\mu}^{full}$ is, equivalently, the solution to the (full data) ***estimating equation***

$$\sum_{i=1}^{N} (Y_i - \mu) = 0. \tag{5.1}$$

Even though V is available, it is not needed in (5.1).

Now consider the case of missing data. As usual, $R = (R_1, R_2)^T$. Now suppose as in that example that V is always observed while Y can be missing, so that the two possible values of R are $(1,1)^T$ and $(0,1)^T$. Let $C = 1$ if $R = (1,1)^T$ and $C = 0$ if $R = (0,1)^T$. Thus, the observed data can be summarized as (C, CY, V), and a sample of observed data on N individuals can be written (C_i, C_iY_i, V_i), $i = 1, \ldots, N$.

As in (1.21), if we are willing to assume that missingness of Y depends only on V and not on Y, i.e.,

$$\text{pr}(C = 1 | Y, V) = \text{pr}(C = 1 | V) = \pi(V), \qquad \pi(v) > 0 \text{ for all } v, \tag{5.2}$$

equivalently, $C \perp\!\!\!\perp Y | V$, and the missingness mechanism is MAR. As demonstrated in Section 1.4, under these conditions, the ***complete case*** estimator, the sample mean of the Y_i for the individuals on whom Y is observed,

$$\widehat{\mu}^{cc} = \frac{\sum_{i=1}^{N} C_i Y_i}{\sum_{i=1}^{N} C_i},$$

is in general ***not*** a consistent estimator for μ.

It is straightforward to see that, equivalently, $\widehat{\mu}^{cc}$ solves the estimating equation

$$\sum_{i=1}^{N} C_i (Y_i - \mu) = 0. \tag{5.3}$$

An inverse probability weighted estimator that is consistent can be derived by ***weighting*** the complete case estimating equation (5.3). Consider the ***inverse probability weighted complete case estimating equation***

$$\sum_{i=1}^{N} \frac{C_i}{\pi(V_i)} (Y_i - \mu) = 0; \tag{5.4}$$

(5.4) weights the contribution of each complete case i by the inverse (reciprocal) of $\pi(V_i)$.

Recall from our review of estimating equations in Section 1.5 that, to show that (5.4) leads to a consistent estimator for μ, it suffices to show that (5.4) is a ***unbiased estimating equation***; i.e., that the ***estimating function***

$$\frac{C}{\pi(V)}(Y-\mu)$$

satisfies

$$E_\mu\left\{\frac{C}{\pi(V)}(Y-\mu)\right\} = 0. \tag{5.5}$$

To show (5.5), we use the law of iterated conditional expectation as follows:

$$\begin{aligned} E_\mu\left\{\frac{C}{\pi(V)}(Y-\mu)\right\} &= E_\mu\left[E\left\{\frac{C}{\pi(V)}(Y-\mu)|Y,V\right\}\right] \\ &= E_\mu\left\{\frac{E(C|Y,V)}{\pi(V)}(Y-\mu)\right\} \\ &= E_\mu\left\{\frac{\pi(V)}{\pi(V)}(Y-\mu)\right\} &(5.6) \\ &= E_\mu(Y-\mu) = 0, &(5.7) \end{aligned}$$

where (5.6) follows because

$$E(C|Y,V) = \text{pr}(C=1|Y,V) = \text{pr}(C=1|V) = \pi(V)$$

under MAR, and $\pi(V)/\pi(V) = 1$ because $\pi(V) > 0$ almost surely leads to (5.7).

REMARKS:

- The estimator solving (5.4) is

$$\widehat{\mu}^{ipw2} = \left\{\sum_{i=1}^N \frac{C_i}{\pi(V_i)}\right\}^{-1} \sum_{i=1}^N \frac{C_i Y_i}{\pi(V_i)}. \tag{5.8}$$

  Comparing (5.8) to the estimator

$$\widehat{\mu}^{ipw} = N^{-1}\sum_{i=1}^N \frac{C_i Y_i}{\pi(V_i)}$$

  given in (1.22) shows that they are not the same. The estimator $\widehat{\mu}^{ipw2}$ in (5.8) is a ***weighted average*** of the observed Y_i and is thus guaranteed to be a value between the minimum and maximum of these Y values, whereas for $\widehat{\mu}^{ipw}$ this need not be the case.

- Both of these estimators treat $\pi(v)$ as if it is a known function of v. If missingness is ***by design***, as for the nutrition study discussed in Chapter 1 in which subjects are selected for a ***validation sample*** according to some known mechanism, then $\pi(v)$ would indeed be known.

  However, in most situations, $\pi(v)$ ***is not*** known. Accordingly, to use such inverse probability weighted estimators in practice, we need to deduce $\pi(v)$ based on the data. In particular, we generally can posit a ***model*** for $\text{pr}(C = 1|V)$; for example, a fully parametric model

$$\pi(V; \psi),$$

  depending on a finite-dimensional parameter ψ, say, and estimate ψ from the data. Because C is ***binary***, a natural approach is to posit a ***logistic regression model*** such as

$$\text{logit}\{\text{pr}(C = 1|V)\} = \psi_0 + \psi_1^T V,$$

  say, where recall that $\text{logit}(p) = \log\{p/(1-p)\}$.

  Using the data (C_i, V_i), $i = 1, \ldots, N$, we can then estimate ψ by ***maximum likelihood***, obtaining the MLE $\widehat{\psi}$ by maximizing

$$\prod_{i=1}^{N} \{\pi(V_i; \psi)\}^{I(C_i=1)}\{1 - \pi(V_i; \psi)\}^{I(C_i=0)} = \prod_{i=1}^{N} \{\pi(V_i; \psi)\}^{C_i}\{1 - \pi(V_i; \psi)\}^{1-C_i}. \tag{5.9}$$

  In this case, the ***inverse probability weight complete case estimator*** based on (5.8) is given by

$$\widehat{\mu}^{ipw2} = \left\{ \sum_{i=1}^{N} \frac{C_i}{\pi(V_i; \widehat{\psi})} \right\}^{-1} \sum_{i=1}^{N} \frac{C_i Y_i}{\pi(V_i; \widehat{\psi})}. \tag{5.10}$$

- It is also clear from this development that $\widehat{\mu}^{ipw2}$ can be an ***inconsistent*** estimator if the model $\pi(v; \psi)$ is ***misspecified***; i.e., if there is no value of ψ for which $\text{pr}(C = 1|V = v) = \pi(v; \psi)$.

- Moreover, by construction, inverse probability weighted complete case estimators such as (5.8) use data ***only*** from the complete cases $\{i : C_i = 1\}$ and disregard data from individuals for whom Y_i is missing. Intuitively, this is likely to result in ***inefficiency***.

AUGMENTED INVERSE PROBABILITY WEIGHTED ESTIMATORS: It turns out, as shown by Robins et al. (1994), that there is a class of estimators that involves ***augmenting*** the simple inverse probability weighted complete case estimating equation for μ. Estimators in this class can yield improved efficiency.

The ***optimal*** estimator for μ within this class is the solution to the estimating equation

$$\sum_{i=1}^{N}\left[\frac{C_i}{\pi(V_i;\widehat{\psi})}(Y_i-\mu)-\frac{C_i-\pi(V_i;\widehat{\psi})}{\pi(V_i;\widehat{\psi})}E\{(Y_i-\mu)|V_i\}\right]=0,$$

which, after some algebra, can be written as

$$\sum_{i=1}^{N}\left\{\frac{C_iY_i}{\pi(V_i;\widehat{\psi})}-\frac{C_i-\pi(V_i;\widehat{\psi})}{\pi(V_i;\widehat{\psi})}E(Y_i|V_i)-\mu\right\}=0, \qquad (5.11)$$

and leads to the estimator

$$\widehat{\mu}=N^{-1}\sum_{i=1}^{N}\left\{\frac{C_iY_i}{\pi(V_i;\widehat{\psi})}-\frac{C_i-\pi(V_i;\widehat{\psi})}{\pi(V_i;\widehat{\psi})}E(Y_i|V_i)\right\}.$$

In (5.11) and consequently the expression for $\widehat{\mu}$, the conditional expectation $E(Y|V)$, the regression of Y on V, is not known. Accordingly, to implement (5.11), $E(Y|V)$ must be modeled and fitted based on the data. Because of MAR, we have

$$C \perp\!\!\!\perp Y|V,$$

from which it follows that

$$E(Y|V)=E(Y|V,C=1).$$

That is, the conditional expectation of Y given V is ***the same*** as that among individuals on whom Y is observed. Thus, we can base positing and fitting a model for $E(Y|V=v)$,

$$m(v;\xi),$$

say, involving a finite-dimensional parameter ξ, on the complete cases $\{i : C_i = 1\}$. Specifically, if Y is continuous, for example, we might derive an estimator $\widehat{\xi}$ for ξ by OLS, solving in ξ

$$\sum_{i=1}^{N}C_i\frac{\partial}{\partial\xi}\{m(V_i;\xi)\}\{Y_i-m(V_i;\xi)\}=0, \qquad (5.12)$$

the OLS estimating equation based on the complete cases.

Substituting in (5.11), the resulting ***augmented inverse probability weighted estimator*** for μ is

$$\widehat{\mu}^{aipw}=N^{-1}\sum_{i=1}^{N}\left\{\frac{C_iY_i}{\pi(V_i;\widehat{\psi})}-\frac{C_i-\pi(V_i;\widehat{\psi})}{\pi(V_i;\widehat{\psi})}m(V_i;\widehat{\xi})\right\}. \qquad (5.13)$$

It can be shown that $\widehat{\mu}^{aipw}$ in (5.13) relatively more efficient than the simple inverse probability weighted estimator (5.10). Moreover, it also has the property of ***double robustness***.

DOUBLE ROBUSTNESS: It can be shown that the estimator $\widehat{\mu}^{aipw}$ in (5.13) is a ***consistent*** estimator for μ if ***EITHER***

- the model $\pi(v;\psi)$ for pr$(C=1|V=v)$ is ***correctly specified***, ***OR***
- The model $m(v;\xi)$ for $E(Y|V=1)$ is ***correctly specified***

(or both). This property is referred to as ***double robustness***, and the estimator $\widehat{\mu}^{aipw}$ is said to be ***doubly robust*** because its consistency is ***robust to*** misspecification of either of these models.

A heuristic demonstration of this double robustness property is as follows. Under regularity conditions, $\widehat{\mu}^{aipw}$ in (5.13) converges in probability to

$$E\left\{\frac{CY}{\pi(V;\psi^*)} - \frac{C-\pi(V;\psi^*)}{\pi(V;\psi^*)}m(V;\xi^*)\right\}, \tag{5.14}$$

where ψ^* and ξ^* are the limits in probability of $\widehat{\psi}$ and $\widehat{\xi}$. Adding and subtracting common terms in (5.14), (5.14) can be written as

$$E\left[Y + \left\{\frac{C-\pi(V;\psi^*)}{\pi(V;\psi^*)}\right\}\{Y-m(V;\xi^*)\}\right] = \mu + E\left[\left\{\frac{C-\pi(V;\psi^*)}{\pi(V;\psi^*)}\right\}\{Y-m(V;\xi^*)\}\right].$$

Consequently, $\widehat{\mu}^{aipw}$ is a consistent estimator for μ if we can show that

$$E\left[\left\{\frac{C-\pi(V;\psi^*)}{\pi(V;\psi^*)}\right\}\{Y-m(V;\xi^*)\}\right] = 0. \tag{5.15}$$

Using iterated conditional (on V) expectation, (5.15) can be written as

$$\begin{aligned} &E\left(E\left[\left\{\frac{C-\pi(V;\psi^*)}{\pi(V;\psi^*)}\right\}\{Y-m(V;\xi^*)\}\middle| V\right]\right) \\ &\qquad = E\left[E\left\{\frac{C-\pi(V;\psi^*)}{\pi(V;\psi^*)}\middle| V\right\}E\{Y-m(V;\xi^*)|V\}\right], \end{aligned} \tag{5.16}$$

where (5.16) is a consequence of MAR, so that $C \perp\!\!\!\perp Y|V$.

Consider two cases:

(a) $\pi(v;\psi)$ is ***correctly specified.*** Then $\widehat{\psi}$ converges in probability to the true value of ψ, so that

$$\pi(V;\psi^*) = \text{pr}(C=1|V).$$

Under this condition,

$$E\left\{\frac{C-\pi(V;\psi^*)}{\pi(V;\psi^*)}\middle| V\right\} = E\left\{\frac{E(C|V)-\text{pr}(C=1|V)}{\text{pr}(C=1|V)}\right\} = 0$$

using $E(C|V) = \text{pr}(C=1|V)$, and (5.15) follows.

(b) $m(v;\xi)$ is ***correctly specified***. Then $\widehat{\xi}$ converges in probability to the true value of ξ, and thus

$$m(V;\xi^*) = E(Y|V).$$

In this case, $E\{Y - m(V;\xi^*)|V\} = E\{Y - E(Y|V)|V\} = 0$, and (5.15) follows.

The results in (a) and (b) thus confirm the double robustness property.

5.2 Inverse probability weighted estimators in regression

Recall ***EXAMPLE 2*** of Section 1.4 of Chapter 1, involving missingness in ***regression analysis***. We now consider how the foregoing principles can be used to derive inverse probability weighted and doubly robust, augmented inverse probability weighted estimators for the regression parameter in a regression model of interest.

Suppose that the full data $Z = (Y, X, V)$, where Y is a scalar outcome, and X is a vector of covariates. As in the previous example, V is a set of additional, auxiliary variables. Interest focuses on a regression model for $E(Y|X = x)$, given by

$$\mu(x;\beta).$$

Suppose that this model is ***correctly specified***. This is a ***semiparametric*** model, as the distribution of Z beyond the form of $E(Y|X)$ is unspecified.

Assume that (X, V) are ***always observed***, but that Y can be ***missing***, and, as in the previous example, let $C = 1$ if Y is observed and $C = 0$ if it is missing. The observed data are thus (C, CY, X, V), and the full sample of observed data can be written as (C_i, C_iY_i, X_i, V_i), $i = 1, \dots, N$.

Here, although the variables in V are not involved in the model of interest for $E(Y|X = x)$, suppose they are needed to make the assumption of MAR tenable. Specifically, assume that

$$\text{pr}(C = 1|Y, X, V) = \text{pr}(C = 1|X, V) = \pi(X, V), \qquad (5.17)$$

say. That is, we are unable to assume that

$$\text{pr}(C = 1|Y, X) = \text{pr}(C = 1|X),$$

which would have allowed us to use the usual, ***complete case*** estimator for β as described in Section 1.4. However, the availability of V makes the MAR assumption (5.17) viable.

Suppose that Y is continuous. Recall from (1.27) that the complete case OLS estimator for β is the solution to the estimating equation

$$\sum_{i=1}^{N} C_i \frac{\partial}{\partial\beta}\{\mu(X_i;\beta)\}\{Y_i - \mu(X_i;\beta)\} = 0. \tag{5.18}$$

We can examine the consistency of the complete case estimator under these conditions by looking at the ***estimating function*** in (5.18). Specifically, using (5.17),

$$\begin{aligned} E_\beta &\left[C\frac{\partial}{\partial\beta}\{\mu(X;\beta)\}\{Y - \mu(X;\beta)\}\right] \\ &= E_\beta\left(E_\beta\left[C\frac{\partial}{\partial\beta}\{\mu(X;\beta)\}\{Y - \mu(X;\beta)\}\middle| Y, X, V\right]\right) \\ &= E_\beta\left[\pi(X, V)\frac{\partial}{\partial\beta}\{\mu(X;\beta)\}\{Y - \mu(X;\beta)\}\right], \end{aligned} \tag{5.19}$$

which is not equal to zero in general, so that (5.18) is not an ***unbiased estimating equation***.

However, using the same ideas as for the case of a single mean in Section 5.1, consider the ***inverse probability weighted complete case*** estimating equation

$$\sum_{i=1}^{N} \frac{C_i}{\pi(X_i, V_i)} \frac{\partial}{\partial\beta}\{\mu(X_i;\beta)\}\{Y_i - \mu(X_i;\beta)\} = 0. \tag{5.20}$$

Using a conditioning argument similar to that leading to (5.19) (try it), we have

$$\begin{aligned} E_\beta &\left[\frac{C}{\pi(X, V)}\frac{\partial}{\partial\beta}\{\mu(X;\beta)\}\{Y - \mu(X;\beta)\}\right] \\ &= E_\beta\left[\frac{\partial}{\partial\beta}\{\mu(X;\beta)\}\{Y - \mu(X;\beta)\}\right] \\ &= E_\beta\left(E_\beta\left[\frac{\partial}{\partial\beta}\{\mu(X;\beta)\}\{Y - \mu(X;\beta)\}\middle| X\right]\right) \\ &= E_\beta\left[\frac{\partial}{\partial\beta}\{\mu(X;\beta)\}\{E(Y|X) - \mu(X;\beta)\}\right] = 0, \end{aligned}$$

as the model $\mu(X;\beta)$ for $E(Y|X)$ is ***correctly specified***. Thus, the inverse probability weighted complete case estimator for β solving (5.20) is ***consistent*** for β.

To implement these ideas in practice, as in the previous example, because $\pi(x, v) = \text{pr}(C = 1|X = x, V = v)$ is not known, we must posit a model for it and fit it using the data. As in Section 5.1, a ***binary regression*** model

$$\pi(x, v;\psi)$$

can be specified; e.g., a logistic regression model.

Analogous to (5.9), ψ can be estimated by the MLE $\widehat{\psi}$ maximizing

$$\prod_{i=1}^{N}\{\pi(X_i, V_i;\psi)\}^{I(C_i=1)}\{1-\pi(X_i, V_i;\psi)\}^{I(C_i=0)} = \prod_{i=1}^{N}\{\pi(X_i, V_i;\psi)\}^{C_i}\{1-\pi(X_i, V_i;\psi)\}^{1-C_i}, \qquad (5.21)$$

using the data (C_i, X_i, V_i), $i = 1, \dots, N$. The fitted $\pi(X_i, V_i;\widehat{\psi})$ can then be substituted in (5.20), and β can be estimated by solving

$$\sum_{i=1}^{N}\frac{C_i}{\pi(X_i, V_i;\widehat{\psi})}\frac{\partial}{\partial\beta}\{\mu(X_i;\beta)\}\{Y_i-\mu(X_i;\beta)\} = 0.$$

A ***doubly robust***, ***augmented inverse probability weighted complete case estimator*** for β can also be derived by considering the estimating equation

$$\sum_{i=1}^{N}\left(\frac{C_i}{\pi(X_i, V_i;\widehat{\psi})}\frac{\partial}{\partial\beta}\{\mu(X_i;\beta)\}\{Y_i-\mu(X_i;\beta)\}\right.$$
$$\left.-\left\{\frac{C_i-\pi(X_i, V_i;\widehat{\psi})}{\pi(X_i, V_i;\widehat{\psi})}\right\}E\left[\frac{\partial}{\partial\beta}\{\mu(X_i;\beta)\}\{Y_i-\mu(X_i;\beta)\}|X_i, V_i\right]\right) = 0.$$

This equation is equal to

$$\sum_{i=1}^{N}\left[\frac{C_i}{\pi(X_i, V_i;\widehat{\psi})}\frac{\partial}{\partial\beta}\{\mu(X_i;\beta)\}\{Y_i-\mu(X_i;\beta)\}\right.$$
$$\left.-\left\{\frac{C_i-\pi(X_i, V_i;\widehat{\psi})}{\pi(X_i, V_i;\widehat{\psi})}\right\}\frac{\partial}{\partial\beta}\{\mu(X_i;\beta)\}\{E(Y_i|X_i, V_i)-\mu(X_i;\beta)\}\right] = 0. \qquad (5.22)$$

Note that, in (5.22), $E(Y|X, V)$ is ***not known***. As in the previous example, we posit a model

$$m(x, v;\xi)$$

for $E(Y|X=x, V=v)$. By MAR, we have

$$E(Y|X, V) = E(Y|X, V, C=1),$$

so that this model can be developed and fitted using the data on the ***complete cases*** only, $\{i : C_i = 1\}$.

For example, analogous to (5.12), an estimator $\widehat{\xi}$ for ξ can be obtained by solving the OLS estimating equation

$$\sum_{i=1}^{N}C_i\frac{\partial}{\partial\xi}\{m(X_i, V_i;\xi)\}\{Y_i-m(X_i, V_i;\xi)\} = 0 \qquad (5.23)$$

using the data on the complete cases.

Substituting $m(X_i, V_i; \widehat{\xi})$ in (5.22) for each i, we obtain the estimating equation to be solved to obtain the ***doubly robust augmented inverse probability weighted complete case estimator*** for β, namely

$$\sum_{i=1}^{N} \left[\frac{C_i}{\pi(X_i, V_i; \widehat{\psi})} \frac{\partial}{\partial \beta} \{\mu(X_i; \beta)\} \{Y_i - \mu(X_i; \beta)\} \right. \\ \left. - \left\{ \frac{C_i - \pi(X_i, V_i; \widehat{\psi})}{\pi(X_i, V_i; \widehat{\psi})} \right\} \frac{\partial}{\partial \beta} \{\mu(X_i; \beta)\} \{m(X_i, V_i; \widehat{\xi}) - \mu(X_i; \beta)\} \right] = 0. \tag{5.24}$$

REMARK: An issue that arises in (5.24) is the ***compatibility*** of the models $m(x, v; \xi)$ for $E(Y|X = x, V = v)$ and $\mu(x; \beta)$ for $E(Y|X = x)$. That is, for these models to be compatible, it must be that

$$\mu(X; \beta) = E(Y|X) = E\{E(Y|X, V)|X\} = E\{m(X, V; \xi)|X\}.$$

One way to develop such compatible models is to assume that the ***centered residual*** $\{Y - \mu(X; \beta)\}$ and the ***centered*** V, $\{V - E(V|X)\}$, are, conditional on X, ***multivariate normal*** with mean zero and covariance matrix that can depend on X.

To demonstrate, for simplicity, assume that the conditional covariance matrix is ***independent*** of X. In this case,

$$\left(\begin{array}{c} Y - \mu(X; \beta) \\ V - E(V|X) \end{array} \middle| X \right) \sim \mathcal{N} \left\{ 0, \begin{pmatrix} \Sigma_{YY} & \Sigma_{YV} \\ \Sigma_{VY} & \Sigma_{VV} \end{pmatrix} \right\},$$

and

$$E(Y|X, V) = \mu(X, ; \beta) + \Sigma_{VY} \Sigma_{VV}^{-1} \{V - E(V|X)\}.$$

If we are also willing to assume that $E(V|X)$ is ***linear*** in X for all V, then

$$E(Y|X, V) = \mu(X, ; \beta) + \xi_0 + \xi_1^T X_i + \xi_2^T V_i.$$

We can then estimate β and ξ ***simultaneously*** by solving jointly the estimating equations

$$\sum_{i=1}^{N} \left[\frac{C_i}{\pi(X_i, V_i; \widehat{\psi})} \frac{\partial}{\partial \beta} \{\mu(X_i; \beta)\} \{Y_i - \mu(X_i; \beta)\} \right. \\ \left. - \left\{ \frac{C_i - \pi(X_i, V_i; \widehat{\psi})}{\pi(X_i, V_i; \widehat{\psi})} \right\} \frac{\partial}{\partial \beta} \{\mu(X_i; \beta)\} \{\xi_0 + \xi_1^T X_i + \xi_2^T V_i\} \right] = 0$$

and

$$\sum_{i=1}^{N} C_i \begin{pmatrix} 1 \\ X_i \\ V_i \end{pmatrix} \{Y_i - \mu(X_i; \beta) - \xi_0 + \xi_1^T X_i + \xi_2^T V_i\} = 0.$$

That these models are compatible is clear because we could generate data as follows. Choose X from an arbitrary distribution. Take $W \sim \mathcal{N}(0, I)$; i.e., generate the components of W as standard normal. Then

$$\begin{pmatrix} Y \\ V \end{pmatrix} = \begin{pmatrix} \mu(X;\beta) \\ E(V|X) \end{pmatrix} + \Sigma^{-1/2} W.$$

(Y, X, V) generated in this fashion guarantee that

$$E(Y|X, V) = \mu(X;\beta) + \Sigma_{VY}\Sigma_{VV}^{-1}\{V - E(V|X)\},$$

and $E(Y|X) = \mu(X;\beta)$.

5.3 Weighted generalized estimating equations for longitudinal data subject to dropout

POPULATION AVERAGE MODELS FOR LONGITUDINAL DATA: For the remainder of this chapter, we consider a more general formulation of the situation of ***longitudinal regression modeling and analysis*** discussed in ***EXAMPLE 3*** of Section 1.4 of Chapter 1. This framework is ***widely used*** for inference from longitudinal data under a ***semiparametric*** model.

Suppose that longitudinal data are to be collected at T time points $t_1 < \cdots < t_T$, where t_1 represents ***baseline***. Specifically, let Y_j be the scalar outcome of interest, X_j be a vector of covariates, and V_j be a vector of additional variables recorded at time t_j, $j = 1, \dots, T$. The full data are then

$$Z = \{(Y_1, X_1, V_1), \dots, (Y_T, X_T, V_T)\}.$$

Letting $Y = (Y_1, \dots, Y_T)^T$, the $(T \times 1)$ outcome vector, and collecting all T covariate vectors as

$$\overline{X}_T = \{X_1, \dots, X_T\},$$

in its most general form, the regression model of interest is

$$E(Y|\overline{X}_T).$$

We consider a model of the form

$$E(Y|\overline{X}_T = \overline{x}_T) = \mu(\overline{x}_T;\beta) = \begin{pmatrix} \mu_1(\overline{x}_T;\beta) \\ \vdots \\ \mu_T(\overline{x}_T;\beta) \end{pmatrix}. \tag{5.25}$$

In (5.25), $\mu_j(\overline{x}_T;\beta)$ is a model for $E(Y_j|\overline{X}_T = \overline{x}_T)$, $j = 1, \dots, T$, and accordingly depends on time t_j as well as the $(p \times 1)$ parameter β.

As in the regression situation in Section 5.2, this is a ***semiparametric*** model, as the distribution of Z is left unspecified beyond the form of $E(Y|\overline{X}_T)$.

Given a sample of data from N individuals, the goal is to estimate β in (5.25).

Models for the expectation of an outcome vector given covariates, as in (5.25), are referred to as ***population average*** or ***population averaged*** models, as they describe average outcome for all individuals in the population of interest having a given covariate value and thus the relationship between average outcome and covariates in the population.

Under the model in (5.25), the expected value of Y_j given the ***entire collection*** of covariates $\overline{X}_T$ over all T time points is taken to potentially depend on ***all*** of these covariates for each $j = 1, \dots, T$. Thus, under this model, the mean outcome at time j can depend on covariates collected in the past (prior to time j), at time j, or in the ***future*** (after time j).

Although in principle possible, adoption of a model that allows mean outcome to depend on ***future covariates*** is rare in practice, as it is difficult to conceive a scientific rationale for such dependence. A special case that is more intuitive and justifiable in practice is to take each component $\mu_j(\overline{x}_T; \beta)$ to depend only on the ***covariates available through time*** t_j. That is, the model for $E(Y_j|\overline{X}_T = \overline{x}_T)$ depends on $\overline{x}_T$ only through $x_1, \dots, x_j$.

In many practical situations, the covariates collected over $t_1, \dots, t_T$ are ***exogenous***. That is, the covariates are such that their values can be determined ***external*** to the evolution of information on the individuals being followed. ***Baseline covariates*** are a key example of exogenous covariates; the values of baseline covariates are available through all T time points.

If ***all*** covariates are exogenous, then write X, with no subscript, to denote the collection of these covariates. Interest then focuses on a model for $E(Y|X)$, and (5.25) simplifies to

$$E(Y|X = x) = \mu(x; \beta) = \begin{pmatrix} \mu_1(x; \beta) \\ \vdots \\ \mu_T(x; \beta) \end{pmatrix}. \tag{5.26}$$

In a model of the form in (5.26), the components $\mu_j(x; \beta)$ depend on x and t_j, so can involve, for example, main effects in the components of x and time and interactions thereof, e.g.,

$$\mu_j(x; \beta) = \beta_0 + \beta_1 t_j + \beta_2^T x + (\beta_3^T x) t_j.$$

A comprehensive account of issues associated with such population average modeling of ***full*** longitudinal data and the implications of different modeling assumptions for the properties of estimators for β obtained by solving ***GEEs*** is beyond our scope here. We remark that one must be very careful when adopting models of the general form in (5.25) that involve ***endogenous*** covariates that change over time; e.g., see the classic paper by Pepe and Anderson (1994).

In the remainder of this chapter, we focus on situations in which the covariates are ***exogenous***, and we write the full data as

$$Z = \{(Y_1, V_1), \dots, (Y_T, V_T), X\}. \tag{5.27}$$

Models of interest are of the form in (5.26), and the goal is to estimate β. Note that, because X is exogenous, its value is known throughout all T time points, and thus is available even on individuals who ***drop out***, as we discuss shortly.

GENERALIZED ESTIMATING EQUATION (GEE) FOR FULL DATA: If a sample of full data Z_i, $i = 1, \dots, N$, is available, then it is well known that the ***optimal GEE*** to be solved to estimate β is given by

$$\sum_{i=1}^{N} \mathcal{D}^T(X_i;\beta)\mathcal{V}^{-1}(X_i;\beta) \begin{pmatrix} Y_{i1} - \mu_1(X_i;\beta) \\ \vdots \\ Y_{iT} - \mu_T(X_i;\beta) \end{pmatrix} = 0, \tag{5.28}$$

where

$$\mathcal{D}(x;\beta) = \frac{\partial}{\partial \beta^T}\{\mu(x;\beta)\}$$

is the $(T \times p)$ matrix of partial derivatives of the T elements of $\mu(x;\beta)$ with respect to the p components of β; and $\mathcal{V}(x;\beta)$ is a $(T \times T)$ ***working covariance matrix***, a model for $\text{var}(Y|X = x)$. Ordinarily, the working covariance matrix also depends on additional ***covariance parameters*** that are estimated from the data by solving additional estimating or moment equations. For brevity, we suppress this in the notation, but be aware that this is a standard, additional feature of fitting population average models like that in (5.26). The GEE (5.28) can be seen to be a generalization of (1.32) discussed in Section 1.4.

From the review of estimating equations in Section 1.5, (5.28) has associated ***estimating function***

$$M(Z;\beta) = \mathcal{D}^T(X;\beta)\mathcal{V}^{-1}(X;\beta) \begin{pmatrix} Y_1 - \mu_1(X;\beta) \\ \vdots \\ Y_T - \mu_T(X;\beta) \end{pmatrix} = \sum_{j=1}^{T} \mathcal{A}_j(X)\{Y_j - \mu_j(X;\beta)\}, \tag{5.29}$$

where $\mathcal{A}_j(x)$, $j = 1, \dots, T$, is the $(p \times 1)$ vector such that the $(p \times T)$ matrix with columns $\mathcal{A}_j(x)$

$$\{\mathcal{A}_1(x), \dots, \mathcal{A}_T(x)\} = \mathcal{D}^T(x;\beta)\mathcal{V}^{-1}(x;\beta).$$

As discussed shortly, among the ***class*** of estimating functions having the form of the rightmost expression in (5.29), this choice of the $\mathcal{A}_j(x)$ is ***optimal***; other choices of the $\mathcal{A}_j(x)$ would lead to different (and not optimal) GEEs. $M(Z;\beta)$ in (5.29) is easily seen to satisfy $E_\beta\{M(Z;\beta)\} = 0$, so that (5.28) is an ***unbiased estimating equation***.

DROPOUT: The foregoing development demonstrates how inference proceeds when a sample of full data is available. We now consider inference on β in the model (5.26) when some individuals ***drop out***, so that the missingness induced is ***monotone***.

Because X is ***exogenous***, it is always observed for all individuals. Under our convention, if an individual drops out at time t_{j+1}, s/he is last seen at time t_j, and we assume in this case that we ***observe*** $(Y_1, V_1), \ldots, (Y_j, V_j)$ but that $(Y_{j+1}, V_{j+1}), \ldots, (Y_T, V_T)$ are ***missing***. As usual, then, let

$$R = (R_1, \ldots, R_T, R_{T+1})^T,$$

corresponding to the $T+1$ components of Z in (5.27). Clearly, $R_{T+1} = 1$ for all individuals. In addition, we assume that ***all individuals are observed at baseline***, so that $R_1 = 1$.

With ***dropout***, if $R_j = 1$, then this implies that $R_2, \ldots, R_{j-1}$ also all are equal to 1. Define as usual

$$D = 1 + \sum_{j=1}^{T} R_j,$$

so that $D = j+1$ implies that the individual is last seen at t_j. Because $R_1 = 1$ always, D thus takes on values $2, \ldots, T+1$, where $D = T+1$ corresponds to the situation where full data are observed.

As in Section 2.3 of Chapter 2, we describe the stochastic dropout process using the ***cause-specific hazard function*** of dropout,

$$\lambda_j(Z) = \text{pr}(D = j | D \geq j, Z), \quad j = 2, \ldots, T. \tag{5.30}$$

Note that $\lambda_1(Z) = \text{pr}(D = 1 | D \geq 1, Z) = 0$ because (Y_1, V_1) are always observed, and $\lambda_{T+1}(Z) = \text{pr}(D = T+1 | D \geq T+1, Z) = 1$ by construction. We then can deduce that (verify)

$$\overline{\pi}_j(Z) = \text{pr}(R_j = 1 | Z) = \prod_{\ell=1}^{j} \{1 - \lambda_\ell(Z)\}, \quad j = 2, \ldots, T \tag{5.31}$$

and

$$\text{pr}(D = j+1 | Z) = \overline{\pi}_j(Z)\lambda_{j+1}(Z), \quad j = 1, \ldots, T. \tag{5.32}$$

Note that, because all individuals are observed at baseline, $\overline{\pi}_1(Z) = \text{pr}(R_1 = 1 | Z) = \text{pr}(R_1 = 1) = 1$.

MAR ASSUMPTION: We assume that the dropout mechanism is ***MAR***. It is convenient to define

$$H_j = \{X, (Y_1, V_1), \ldots, (Y_j, V_j)\}, \quad j = 1, \ldots, T,$$

the ***history*** available through time t_j.

MAR implies that the cause-specific hazard of dropout (5.30) can be written as

$$\lambda_j(Z) = \text{pr}(D = j|D \geq j, Z) = \text{pr}(D = j|D \geq j, H_{j-1}) = \lambda_j(H_{j-1}), \quad j = 2, \dots, T; \tag{5.33}$$

that is, the hazard of dropping out at time t_j (i.e., last being seen at time t_{j-1}) depends only on the ***observed history*** through time t_{j-1}. Likewise, (5.31) and (5.32) become

$$\overline{\pi}_j(Z) = \overline{\pi}_j(H_{j-1}) = \text{pr}(R_j = 1|H_{j-1}) = \prod_{\ell=1}^{j} \{1 - \lambda_\ell(H_{\ell-1})\}, \quad j = 2, \dots, T, \tag{5.34}$$

and

$$\text{pr}(D = j + 1|Z) = \text{pr}(D = j + 1|H_j) = \overline{\pi}_j(H_{j-1})\lambda_{j+1}(H_j), \quad j = 1, \dots, T, \tag{5.35}$$

By convention in formulæ to follow, when $j = 1$, $\overline{\pi}_j(H_{j-1}) = \overline{\pi}_1 = 1$. We will use this and (5.33), (5.34), and (5.35) in the sequel.

AUXILIARY VARIABLES: As in Sections 5.1 and 5.2, although the auxiliary variables V_j, $j = 1, \dots, T$, are not relevant to the longitudinal regression model $\mu(x; \beta)$ of interest, as the foregoing development shows, they may be implicated in the dropout mechanism and thus are necessary to render the assumption of MAR ***plausible***.

WEIGHTED GENERALIZED ESTIMATING EQUATIONS (WGEEs) UNDER MAR DROPOUT: We now discuss how the usual full data GEE (5.28) can be ***modified*** in the case of dropout to lead to estimators for β based on a sample of observed data subject to dropout. Approaches include

(i) ***Inverse probability weighting at the occasion level.*** This approach was first proposed by Robins, Rotnitzky, and Zhao (1995) and involves using weights specific to each time point. These methods are applicable only in the situation we discuss here, where there are only ***exogenous covariates***, as in (5.26).

(ii) ***Inverse probability weighting at the subject level.*** This approach was proposed by Fitzmaurice, Molenberghs, and Lipsitz (1995) and is applicable more generally to models of the form (5.25) that depend on $\overline{x}_T$ only through $x_1, \dots, x_j$.

(iii) ***Doubly robust methods***.

We discuss (ii) first, followed by (i), and defer (iii) to the next section.

In each case, we discuss the associated WGEEs by presenting their corresponding ***estimating functions***, which depend on the observed data. Note that the observed data on an individual, $(R, Z_{(R)})$, can be expressed as R (equivalently, D), and if $D = j+1$, H_j. The estimating functions are expressed in terms of (5.33), (5.34), and (5.35) as if these functions were known. We discuss modeling and fitting of the dropout hazards at the end of this section.

INVERSE PROBABILITY WEIGHTING AT THE SUBJECT LEVEL: The general form of the estimating function involving ***subject level*** weighting is given by

$$\sum_{j=1}^{T} \left\{ \frac{I(D=j+1)}{\overline{\pi}_j(H_{j-1})\lambda_{j+1}(H_j)} \right\} \Big[G_{j1}(X)\{Y_1 - \mu_1(X;\beta)\} + \cdots + G_{jj}(X)\{Y_j - \mu_j(X;\beta)\} \Big], \tag{5.36}$$

where $G_{j\ell}(x)$, $\ell = 1, \dots, j$, $j = 1, \dots, T$, are arbitrary $(p \times 1)$ functions of x. Thus, note that the estimating function at the jth level, say, has coefficients in x that vary by j. Assuming that the dropout model is correctly specified, it is straightforward to show that (5.36) is an ***unbiased estimating function*** by first conditioning on the full data and then on X (try it).

For individual i for whom $D_i = j+1$ for fixed $j = 1, \dots, T$, his/her contribution to the WGEE is

$$\left\{ \frac{I(D_i=j+1)}{\overline{\pi}_j(H_{i,j-1})\lambda_{j+1}(i, H_j)} \right\} \Big[G_{j1}(X_i)\{Y_{i1} - \mu_1(X_i;\beta)\} + \cdots + G_{jj}(X_i)\{Y_{ij} - \mu_j(X_i;\beta)\} \Big]$$

corresponding to this j. Thus, for each individual, there is a single, ***subject level*** weight,

$$\left\{ \frac{I(D=j+1)}{\overline{\pi}_j(H_{j-1})\lambda_{j+1}(H_j)} \right\}$$

applied to the linear combination of his/her $\{Y_\ell - \mu_\ell(X;\beta)\}$, $\ell = 1, \dots, j$.

Fitzmaurice et al. (1995) suggest taking the $(p \times j)$ matrix

$$\{G_{j1}(X), \dots, G_{jj}(X)\} = \mathcal{D}_j^T(X;\beta)\mathcal{V}_j^{-1}(X;\beta), \tag{5.37}$$

where, as in Section 1.4, $\mathcal{D}_j^T(x;\beta)$ $(p \times j)$ and $\mathcal{V}_j(x;\beta)$ $(j \times j)$ are the corresponding submatrices of $\mathcal{D}^T(x;\beta)$ $(p \times T)$ and $\mathcal{V}(x;\beta)$ $(T \times T)$. Recall from ***EXAMPLE 3*** in Section 1.4 that (5.37) corresponds to what would be used in a naive analysis based on the ***available data***.

Thus, the WGEE based on the estimating function (5.37) using this specification can be interpreted as a ***weighted*** (with a ***single, scalar weight*** for each individual) version of the estimating equations that would be used for the naive, ***available data*** analysis, namely (compare to (1.33))

$$\sum_{i=1}^{N} \left\{ \sum_{j=1}^{T} w_{ij}\mathcal{D}_j^T(X_i;\beta)\mathcal{V}_j^{-1}(X_i;\beta) \begin{pmatrix} Y_{i1} - \mu_1(X_i;\beta) \\ \vdots \\ Y_{ij} - \mu_j(X_i;\beta) \end{pmatrix} \right\} = 0, \quad w_{ij} = \frac{I(D_i=j+1)}{\overline{\pi}_j(H_{i,j-1})\lambda_{j+1}(i, H_j)}.$$

INVERSE PROBABILITY WEIGHTING AT THE OCCASION LEVEL: The general form of the estimating function involving ***occasion-specific*** weighting is

$$\sum_{j=1}^{T} \frac{R_j}{\pi_j(H_{j-1})} B_j(X)\{Y_j - \mu_j(X;\beta)\}, \tag{5.38}$$

where $B_1(x), \ldots, B_T(x)$ are arbitrary $(p \times 1)$ functions of x. That (5.38) is an ***unbiased estimating function*** can also be shown by first conditioning on the full data and then on X (try it).

For an individual i for whom $D_i = j+1$ for fixed $j = 1, \ldots, T$, note that $R_{i1} = \cdots = R_{ij} = 1$, with $R_{i,j+1} = 0$ henceforth. Thus, in contrast to (5.36), for such an individual, his/her contribution to the WGEE is, from (5.38),

$$\frac{R_{i1}}{\pi_1} B_1(X_i)\{Y_{i1} - \mu_1(X_i;\beta)\} + \frac{R_{i2}}{\pi_2(H_{i1})} B_2(X_i)\{Y_{i2} - \mu_2(X_i;\beta)\} + \cdots + \frac{R_{ij}}{\pi_j(H_{i,j-1})} B_j(X_i)\{Y_{ij} - \mu_j(X_i;\beta)\}.$$

This demonstrates that (5.38) involves ***separate***, ***occasion-level*** weighting of ***each*** component of individual i's contribution to the estimating equation, where each component and its corresponding weight is ***specific*** to time (occasion) t_j for all t_j at which i has not yet dropped out.

Robins et al. (1995) suggest taking the $(p \times T)$ matrix

$$\{B_1(X), \ldots, B_T(X)\} = \mathcal{D}^T(X;\beta)\mathcal{V}^{-1}(X;\beta). \tag{5.39}$$

This is, of course, the choice corresponding to the ***optimal***, ***full data*** GEE. Thus, the WGEE corresponding to the estimating function (5.38) with specification (5.39) can be interpreted as a ***weighted version*** of (5.28), namely

$$\sum_{i=1}^{N} \mathcal{D}^T(X_i;\beta)\mathcal{V}^{-1}(X_i;\beta)\mathcal{W}_i \begin{pmatrix} Y_{i1} - \mu_1(X_i;\beta) \\ \vdots \\ Y_{iT} - \mu_T(X_i;\beta) \end{pmatrix} = 0,$$

where $\mathcal{W}_i$ is the $(T \times T)$ diagonal ***weight matrix*** with diagonal elements

$$\frac{R_{i1}}{\pi_1}, \frac{R_{i2}}{\pi_2(H_{i1})}, \ldots, \frac{R_{iT}}{\pi_T(H_{i,T-1})}.$$

SOFTWARE: The SAS procedure `proc gee` in SAS/STAT 13.2 implements both the subject level and occasion level weighted methods using (5.37) for the former and (5.39) for the latter. The weighting method is chosen and the hazard model is specified through the `missmodel` statement. There is an R package, `CRTgeeDR`, that implements an augmented version of occasion-level weighting.

SEMIPARAMETRIC THEORY PERSPECTIVE: We can place the above approaches in the context of the implications of ***semiparametric theory***.

With ***full data***, semiparametric theory shows that the ***class of all estimating functions*** leading to ***consistent and asymptotically normal estimators*** for β in the (semiparametric) model (5.26) is

$$\sum_{j=1}^{T} \mathcal{A}_j(X)\{Y_j - \mu_j(X;\beta)\}$$

for arbitrary $(p\times 1)$ $\mathcal{A}_j(x)$, $j = 1,\ldots,T$. As noted in (5.29), the ***optimal*** choice of the $\mathcal{A}_j(x)$, that leading to the estimator for β with ***smallest asymptotic variance*** among all in this class, is such that the $(p \times T)$ matrix $\{\mathcal{A}_1(x),\ldots,\mathcal{A}_T(x)\} = \mathcal{D}^T(x;\beta)\mathcal{V}^{-1}(x;\beta)$.

When we have observed data that involve MAR ***monotone dropout***, semiparametric theory can likewise be used to derive the ***class of all estimating functions*** leading to consistent and asymptotically normal estimators for β based on the ***observed data***. These estimating functions turn out to have the ***augmented inverse probability weighted complete case*** form

$$\frac{R_T}{\overline{\pi}_T(H_{T-1})}\sum_{j=1}^{T} \mathcal{A}_j(X)\{Y_j - \mu_j(X;\beta)\} + \sum_{j=1}^{T-1}\left\{\frac{R_j}{\overline{\pi}_j(H_{j-1})} - \frac{R_{j+1}}{\overline{\pi}_{j+1}(H_j)}\right\} f_j(H_j), \tag{5.40}$$

where $f_j(H_j)$, $j = 1,\ldots,T-1$, are arbitrary $(p\times 1)$ functions of the histories H_j; and $\mathcal{A}_j(X)$, $j = 1,\ldots,T$, are arbitrary $(p \times 1)$ functions of X. Showing that (5.40) is an ***unbiased estimating function*** is straightforward using conditioning arguments (try it).

We now demonstrate that the estimating functions for the subject level and occasion level inverse weighting approaches in (5.36) and (5.38) are ***special cases*** of (5.40) .

SUBJECT LEVEL: Consider the subject level estimating function (5.36). It is straightforward to deduce that

$$I(D = j + 1) = R_j - R_{j+1};$$

thus, (5.36) can be written as

$$\begin{aligned}&\frac{R_T}{\overline{\pi}_T(H_{T-1})}\Big[G_{T1}(X)\{Y_1 - \mu_1(X;\beta)\} + \cdots + G_{TT}(X)\{Y_T - \mu_T(X;\beta)\}\Big]\\ &\qquad + \sum_{j=1}^{T-1}\frac{R_j - R_{j+1}}{\overline{\pi}_j(H_{j-1})\lambda_{j+1}(H_j)}\Big[G_{j1}(X)\{Y_1 - \mu_1(X;\beta)\} + \cdots + G_{jj}(X)\{Y_j - \mu_j(X;\beta)\}\Big],\end{aligned} \tag{5.41}$$

We now recursively relate (5.40) to (5.41).

Note first that

$$\frac{R_1}{\overline{\pi}_1} f_1(H_1) = \frac{R_1}{\overline{\pi}_1 \lambda_2(H_1)} G_{11}(X)\{Y_1 - \mu_1(X;\beta)\},$$

which implies that

$$f_1(H_1) = \frac{G_{11}(X)\{Y_1 - \mu_1(X;\beta)\}}{\lambda_2(H_1)}.$$

Adopting the shorthand notation

$$\mathcal{G}_{j\ell} = G_{j\ell}(X)\{Y_\ell - \mu_\ell(X;\beta)\},$$

$\overline{\pi}_j = \overline{\pi}_j(H_{j-1})$, and $\lambda_j = \lambda_j(H_{j-1})$, we then have

$$\frac{R_2}{\overline{\pi}_2}\{f_2(H_2) - f_1(H_1)\} = \frac{R_2}{\overline{\pi}_2}\left\{\frac{\mathcal{G}_{21} + \mathcal{G}_{22}}{\lambda_3} - \left(\frac{1-\lambda_2}{\lambda_2}\right)\mathcal{G}_{11}\right\},$$

and thus

$$f_2(H_2) = \frac{\mathcal{G}_{21} + \mathcal{G}_{22}}{\lambda_3} + \mathcal{G}_{11} - \frac{\mathcal{G}_{11}}{\lambda_2} + \frac{\mathcal{G}_{11}}{\lambda_2} = \frac{\mathcal{G}_{21} + \mathcal{G}_{22}}{\lambda_3} + \mathcal{G}_{11}.$$

Next,

$$\frac{R_3}{\overline{\pi}_3}\{f_3(H_3) - f_2(H_2)\} = \frac{R_3}{\overline{\pi}_3}\left\{\frac{\mathcal{G}_{31} + \mathcal{G}_{32} + \mathcal{G}_{33}}{\lambda_4} - \left(\frac{1-\lambda_3}{\lambda_3}\right)(\mathcal{G}_{21} + \mathcal{G}_{22})\right\},$$

so that, solving for $f_3(H_3)$,

$$f_3(H_3) = \frac{\mathcal{G}_{31} + \mathcal{G}_{32} + \mathcal{G}_{33}}{\lambda_4} + (\mathcal{G}_{11} + \mathcal{G}_{21}) + \mathcal{G}_{22}.$$

Continuing with this recursion, we have for $j = 1, \ldots, T-1$,

$$\begin{aligned}
f_j(H_j) \quad &= \quad \frac{G_{j1}(X)\{Y_1 - \mu_1(X;\beta)\} + \cdots + G_{jj}(X)\{Y_j - \mu_j(X;\beta)\}}{\lambda_{j+1}(H_j)} \\
&\quad + \{G_{11}(X) + \cdots + G_{j-1,1}(X)\}\{Y_1 - \mu_1(X;\beta)\} \\
&\quad + \{G_{22}(X) + \cdots + G_{j-1,2}(X)\}\{Y_2 - \mu_2(X;\beta)\} \\
&\quad \vdots \\
&\quad + G_{j-1,j-1}(X)\{Y_j - \mu_j(X;\beta)\},
\end{aligned}$$

and

$$\begin{aligned}
\mathcal{A}_1(X) \quad &= \quad G_{11}(X) + \cdots + G_{T1}(X) \\
\mathcal{A}_2(X) \quad &= \quad G_{22}(X) + \cdots + G_{T2}(X) \\
&\vdots \\
\mathcal{A}_T(X) \quad &= \quad G_{TT}(X).
\end{aligned}$$

OCCASION LEVEL: Consider the occasion level estimating function (5.38). Here, it is straightforward to deduce that (try it) this is a special case of (5.40), with

$$\begin{aligned} f_1(H_1) &= B_1(X)\{Y_1 - \mu_1(X;\beta)\} \\ &\vdots \\ f_j(H_j) &= B_1(X)\{Y_1 - \mu_1(X;\beta)\} + \cdots + B_j(X)\{Y_j - \mu_j(X;\beta)\}, \end{aligned}$$

$j = 2, \ldots, T-1$, and $\mathcal{A}_j(X) = B_j(X)$, $j = 1, \ldots, T$.

ESTIMATING THE DROPOUT HAZARDS: In practice, of course, the dropout hazards

$$\lambda_j(H_{j-1}) = \text{pr}(D = j | D \geq j, H_{j-1}), \quad j = 2, \ldots, T,$$

(5.33), which determine

$$\overline{\pi}_j(H_{j-1}) = \prod_{\ell=1}^{j}\{1 - \lambda_\ell(H_{\ell-1})\}, \quad j = 2, \ldots, T,$$

in (5.34), are not known. We now discuss how these hazards can be modeled and fitted based on the observed data. The resulting fitted models can then be substituted in (5.36), (5.38), or, indeed, any estimating function in the class (5.40).

Suppose we posit models

$$\lambda_j(H_{j-1};\psi)$$

for the $\lambda_j(H_{j-1})$, $j = 2, \ldots, T$, in terms of a vector of parameters ψ. From (5.35) and above, this implies models for $\text{pr}(D = j|Z) = \text{pr}(D = j|H_{j-1})$ of the form

$$\overline{\pi}_{j-1}(H_{j-2};\psi)\lambda_j(H_{j-1};\psi) = \prod_{\ell=1}^{j-1}\{1 - \lambda_\ell(H_{\ell-1};\psi)\}\lambda_j(H_{j-1};\psi),$$

$j = 2, \ldots, T-1$. Then the ***likelihood*** for ψ based on a sample of observed data can be written as

$$\prod_{i=1}^{N}\prod_{j=2}^{T+1}\left[\prod_{\ell=1}^{j-1}\{1 - \lambda_\ell(H_{i,\ell-1};\psi)\}\lambda_j(H_{i,j-1};\psi)\right]^{I(D_i=j)}. \tag{5.42}$$

Rearranging terms, it can be shown (verify) that the likelihood (5.42) can be written as

$$\prod_{j=2}^{T}\prod_{i:R_{i,j-1}=1}\{\lambda_j(H_{i,j-1};\psi)\}^{I(D_i=j)}\{1 - \lambda_j(H_{i,j-1};\psi)\}^{I(D_i>j)},$$

where in fact $I((D > j) = I(R_j = 1)$.

Suppose we adopt a ***logistic*** model for each $\lambda_j(H_{j-1};\psi)$; i.e.,

$$\lambda_j(H_{j-1};\psi) = \frac{\exp\{\alpha_j(H_{j-1};\psi)\}}{1+\exp\{\alpha_j(H_{j-1};\psi)\}},$$

or, equivalently, $\text{logit}\{\lambda_j(H_{j-1};\psi)\} = \alpha_j(H_{j-1};\psi)$, for some functions $\alpha_j(H_{j-1};\psi)$, $j = 1, \ldots, T$. Then it is straightforward to demonstrate that the MLE $\widehat{\psi}$ for ψ maximizing the above likelihood is the solution to the ***estimating equation***

$$\sum_{i=1}^{N}\sum_{j=2}^{T} R_{i,j-1}\frac{\partial}{\partial\psi}\{\alpha_j(H_{i,j-1};\psi)\}\{I(D_i=j) - \lambda_j(H_{i,j-1};\psi)\} = 0. \qquad (5.43)$$

In practice, it would be unusual for an analyst to posit models $\lambda_j(H_{j-1};\psi)$ that share a ***common parameter*** ψ across different j. Rather, a standard approach is to take the model for each occasion j to have a separate parameter ψ_j, say, so that $\psi = (\psi_2^T, \ldots, \psi_T^T)^T$, and the ψ_j are ***variation independent***. Thus, one would take $\lambda_j(H_{j-1};\psi) = \lambda_j(H_{j-1};\psi_j)$ for each j. In this case, solving (5.43) for ψ boils down to solving

$$\sum_{i=1}^{N} R_{i,j-1}\frac{\partial}{\partial\psi_j}\{\alpha_j(H_{i,j-1};\psi_j)\}\{I(D_i=j) - \lambda_j(H_{i,j-1};\psi_j)\} = 0$$

separately for $j = 2, \ldots, T$. That is, to estimate ψ_j, compute the MLE among individuals who have still not dropped out at time $j-1$, using as the response the indicator of whether or not such an individual drops out at time j (or continues beyond j).

Standard software can be used to fit these models; e.g., if the $\alpha_j(H_{j-1};\psi_j)$ are ***linear*** in ψ_j, software such as SAS `proc logistic` or any generalized linear model software can be used.

COMPARISON OF SUBJECT AND OCCASION LEVEL WEIGHTING:

- The subject level approach has greater ***practical appeal*** because it is easier to implement using ***standard software*** for solving GEEs. Many such programs, such as SAS `proc genmod`, allow the user to specify a fixed ***weight*** for each individual. Thus, the user can model and fit the dropout hazards, form weights, and incorporate them straightforwardly in a call to such software to solve the subject level WGEE. The advent of software implementing the occasion level approach, as discussed previously, lessens this appeal.

- Theoretically, it is not straightforward to deduce if the subject level or occasion level approach is ***preferred in general*** on the basis of efficiency.

- Preisser, Lohman, and Rathouz (2002) carried out extensive ***simulation studies*** comparing the two approaches under various MCAR, MAR, and MNAR missingness mechanisms and under correct and misspecified dropout models. They concluded that, overall, under MAR, the occasion level WGEE is to be preferred on efficiency grounds, but noted that both methods can be sensitive to misspecification of the associated weights.

5.4 Doubly robust estimation

We now examine more closely the class of ***augmented inverse probability weighted complete case*** estimating functions (5.40), namely,

$$\frac{R_T}{\pi_T(H_{T-1})}\sum_{j=1}^{T}\mathcal{A}_j(X)\{Y_j-\mu_j(X;\beta)\}+\sum_{j=1}^{T-1}\left\{\frac{R_j}{\pi_j(H_{j-1})}-\frac{R_{j+1}}{\pi_{j+1}(H_j)}\right\}f_j(H_j). \tag{5.44}$$

From the theory of semiparametrics for missing data problems, it can be shown (Tsiatis, 2006) that, for ***fixed*** $\{\mathcal{A}_1(x),\ldots,\mathcal{A}_T(x)\}$, the ***optimal*** choice of $f_j(H_j)$ in (5.44) is

$$\begin{aligned} f_j(H_j) &= E\left[\left.\sum_{\ell=1}^{T}\mathcal{A}_\ell(X)\{Y_\ell-\mu_\ell(X;\beta)\}\right| H_j\right], \\ &= \sum_{\ell=1}^{j}\mathcal{A}_\ell(X)\{Y_\ell-\mu_\ell(X;\beta)\}+\sum_{\ell=j+1}^{T}\mathcal{A}_\ell(X)E[\{Y_\ell-\mu_\ell(X;\beta)\}|H_j], \end{aligned} \tag{5.45}$$

for $j=1,\ldots,T-1$. That is, for ***fixed*** $\mathcal{A}_j(x)$, $j=1,\ldots,T$, using this choice of $f_j(H_j)$, which depends on the particular fixed $\mathcal{A}_j(x)$, will yield the estimating function of form (5.44) with this fixed $\mathcal{A}_j(x)$ leading to the estimator for β with smallest asymptotic variance.

REMARKS:

- To compute $f_j(H_j)$, $j=1,\ldots,T-1$, in (5.45), we must be able to estimate

  $$E[\{Y_\ell-\mu_\ell(X;\beta)\}|H_j],\quad \ell>j$$

  based on the observed data. We discuss this shortly.

- In the ***special case*** where $(Y_1,V_1),\ldots,(Y_T,V_T)$ are ***conditionally independent*** given X,

  $$E[\{Y_\ell-\mu_\ell(X;\beta)\}|H_j]=0\quad\text{for } \ell>j,$$

  so that the ***optimal*** choice is

  $$f_j(H_j)=\sum_{\ell=1}^{j}\mathcal{A}_\ell(X)\{Y_\ell-\mu_\ell(X;\beta)\}.$$

  This leads to the ***occasion level*** WGEE in (5.38) with $\mathcal{A}_j(x)=B_j(x)$, $j=1,\ldots,T$.

However, if $(Y_1, V_1), \ldots, (Y_T, V_T)$ are ***correlated*** conditional on X, then it is possible to take advantage of this correlation by choosing the $f_j(H_j)$ judiciously to ***gain efficiency***. Essentially, this correlation allows us to gain back some ***information*** regarding the missing data from the observed data.

- Moreover, the resulting estimator will be ***doubly robust*** in the sense that it will lead to a consistent estimator for β if ***EITHER*** $\overline{\pi}_j(H_{j-1})$ (equivalently, the $\lambda_j(H_{j-1})$), $j = 2, \ldots, T$, are ***consistently estimated*** (i.e., we have a ***correct model*** for the dropout process), ***OR*** if
$$E[\{Y - \mu(X;\beta)|H_j]$$
is ***consistently estimated*** for all $j = 1, \ldots, T$.

  This doubly robust method is advocated by Seaman and Copas (2009) using the ***fixed*** choice
$$\{\mathcal{A}_1(x), \ldots, \mathcal{A}_T(x)\} = \mathcal{D}^T(x;\beta)\mathcal{V}^{-1}(x;\beta).$$
  As we demonstrated in the last section, for the ***occasion level*** WGEE, $\mathcal{A}_j(x) = B_j(x)$, $j = 1, \ldots, T$, and this is the choice suggested by Robins et al. (1995) in (5.39).

- For both the subject and occasion level WGEEs in (5.36) and (5.38), with the implied choices for the $\mathcal{A}_j(x)$ given in the previous section, the corresponding $f_j(H_j)$ are ***not the same*** as the optimal choice in (5.45). This suggests that it is possible to improve on both of these estimators.

- In this discussion, we have restricted attention to ***fixed*** $\{\mathcal{A}_1(x), \ldots, \mathcal{A}_T(x)\}$; as noted above, in many cases, these are taken to be the choice leading to the optimal GEE for ***full data***. However, it turns out that, from semiparametric theory, that this is ***not*** the optimal choice with observed data subject to ***dropout***. Shortly, we discuss the globally optimal choice and whether or not it is even ***feasible*** to implement this choice in practice.

ESTIMATING THE AUGMENTATION TERM FOR DOUBLY ROBUST ESTIMATORS: As we have just seen, for ***fixed*** $\{\mathcal{A}_1(x), \ldots, \mathcal{A}_T(x)\}$, from (5.45), the optimal choice for $f_j(H_j)$ for construction of a doubly robust, augmented inverse probability weighted estimating function for β requires determining
$$E[\{Y - \mu(X;\beta)|H_j], \quad j = 1, \ldots, T - 1.$$
Because these conditional expectations are not known, they must be estimated from the observed data. We now examine how this might be carried out in practice.

For convenience, denote the ***centered*** data as

$$\epsilon = Y - \mu(X;\beta),$$

with elements $\epsilon_j = Y_j - \mu_j(X;\beta)$. By MAR, it can be shown that, in obvious notation,

$$p_{\epsilon_{j+1},V_{j+1}|H_j}(\epsilon_{j+1}, v_{j+1}|h_j) = p_{\epsilon_{j+1},V_{j+1}|H_j,R_{j+1}}(\epsilon_{j+1}, v_{j+1}|h_j, 1); \tag{5.46}$$

that is, the conditional density of $(\epsilon_{j+1}, V_{j+1})$ given H_j is the same as this conditional density among those individuals who still have not dropped out at time t_{j+1}, at which time we observe $(\epsilon_{j+1}, V_{j+1}, H_j)$. To see this, write the right hand side of (5.46) as

$$\frac{\text{pr}(R_{j+1} = 1|\epsilon_{j+1}, V_{j+1} = v_{j+1}, H_j = h_j)p_{\epsilon_{j+1},V_{j+1},H_j}(\epsilon_{j+1}, v_{j+1}, h_j)}{\text{pr}(R_{j+1} = 1|H_j = h_j)p_{H_j}(h_j)}. \tag{5.47}$$

Because of MAR, $\text{pr}(R_{j+1} = 1|\epsilon_{j+1}, V_{j+1} = v_{j+1}, H_j = h_j) = \text{pr}(R_{j+1} = 1|H_j = h_j)$, so that (5.47) becomes

$$\frac{p_{\epsilon_{j+1},V_{j+1},H_j}(\epsilon_{j+1}, v_{j+1}, h_j)}{p_{H_j}(h_j)} = p_{\epsilon_{j+1},V_{j+1}|H_j}(\epsilon_{j+1}, v_{j+1}|h_j).$$

It is convenient to write H_j equivalently as the ordered column vector

$$H_j = (1, X^T, \epsilon_1, V_1^T, \ldots, \epsilon_j, V_j^T)^T, \quad j = 1, \ldots, T-1.$$

We wish to estimate $E(\epsilon|H_j)$ for each j.

In general, this is a ***numerical challenge***. One possible approach is based on making the approximation that $(\epsilon_1, \ldots, \epsilon_T, V_1^T, \ldots, V_T^T, X^T)^T$ is ***multivariate normal***.

To see how this works, let q_j be the dimension of H_j and r_j be the dimension of V_j, $j = 1, \ldots, T$. Note that $q_{j+1} = q_j + r_j + 1$. Under the normality assumption,

$$E\left(\begin{array}{c|c} \epsilon_{j+1} & \\ V_{j+1} & \end{array} H_j\right) = \begin{pmatrix} \Lambda_j \\ \Gamma_j \end{pmatrix} H_j,$$

where Λ_j $(1 \times q_j)$ and Γ_j $(r_j \times q_j)$ are constants. That is, the conditional expectation of $(\epsilon_{j+1}, V_{j+1}^T)^T$ given H_j is a ***linear combination*** of elements of H_j for each j.

Note that we can also write

$$E(H_{j+1}|H_j) = \begin{pmatrix} I_{q_j} \\ \Lambda_j \\ \Gamma_j \end{pmatrix} H_j = \Delta_j H_j,$$

say. Here Δ_j is $(q_{j+1} \times q_j)$.

As noted earlier, because of MAR, in fact

$$E(H_{j+1}|H_j) = E(H_{j+1}|H_j, R_{j+1} = 1);$$

consequently, the elements of Δ_j can be estimated using the individuals ***at risk*** at time t_{j+1}; i.e, those individuals who have still not dropped out at t_{j+1}.

This can be accomplished using ***least squares*** for each j by regressing ϵ_{j+1} and each element of V_{j+1} on those of H_j. Note that

$$H_T = (1, X^T, \epsilon_1, V_1^T, \dots, \epsilon_T, V_T^T)^T;$$

thus, by the laws of iterated conditional expectations,

$$E(H_T|H_j) = \Delta_{T-1}\Delta_{T-2}\cdots\Delta_j H_j.$$

Consequently, $f_j(H_j) = E(\epsilon|H_j)$ can be obtained by "picking off" the elements $E(\epsilon|H_j)$ from $E(H_T|H_j)$.

In fact, if we are willing to assume multivariate normality of $C = (\epsilon_1, \dots, \epsilon_T, V_1^T, \dots, V_T^T)^T$ given X, then it is possible to use the EM algorithm with monotone missingness to estimate the parameters in the covariance matrix of C given X and thereby obtain an estimate of the working covariance matrix corresponding to the submatrix var$(\epsilon_1, \dots, \epsilon_T|X)$.

REMARK: We do not elaborate on this scheme further, as, admittedly it is quite involved.

LOCALLY EFFICIENT DOUBLY ROBUST ESTIMATOR: We now examine how semiparametric theory can be used in principle to obtain the ***globally optimal*** WGEE.

Consider again the class of ***augmented inverse probability weighted complete case*** estimating functions (5.40),

$$\frac{R_T}{\overline{\pi}_T(H_{T-1})}\sum_{j=1}^{T}\mathcal{A}_j(X)\{Y_j - \mu_j(X;\beta)\} + \sum_{j=1}^{T-1}\left\{\frac{R_j}{\overline{\pi}_j(H_{j-1})} - \frac{R_{j+1}}{\overline{\pi}_{j+1}(H_j)}\right\} f_j(H_j). \tag{5.48}$$

We have discussed the ***optimal*** choice of the $f_j(H_j)$ when the $\mathcal{A}_j(x)$ are ***fixed***. ***However***, as we remarked previously, this will not lead to the globally optimal estimating function of form (5.48) unless the optimal choice of the $\mathcal{A}_j(x)$ is used.

In Tsiatis (2006), it is shown that the ***optimal*** $\{\mathcal{A}_1(x), \dots, \mathcal{A}_T(x)\}$ is given by

$$\{\mathcal{A}_1(x), \dots, \mathcal{A}_T(x)\} = \mathcal{D}^T(x;\beta)\{\mathcal{V}^*(x)\}^{-1},$$

where

$$\mathcal{V}^*(X) = E\left\{ \left. \frac{\epsilon\epsilon^T}{\overline{\pi}_T(H_{T-1})} \right| X \right\} - \sum_{j=1}^{T-1} E\left\{ \left. \frac{\lambda_{j+1}(H_j)}{\overline{\pi}_{j+1}(H_j)} E(\epsilon|H_j)\epsilon^T \right| X \right\}, \qquad (5.49)$$

which can be written equivalently as

$$\mathcal{V}^*(X) = E\left\{ \left. \frac{\epsilon\epsilon^T}{\overline{\pi}_T(H_{T-1})} \right| X \right\} - \sum_{j=1}^{T-1} \left[E\left\{ \left. \frac{E(\epsilon|H_j)}{\overline{\pi}_j(H_{j-1})}\epsilon^T \right| X \right\} - E\left\{ \left. \frac{E(\epsilon|H_j)}{\overline{\pi}_{j+1}(H_j)}\epsilon^T \right| X \right\} \right].$$

Of course, (5.49) is obviously very difficult to evaluate. However, under the approximation of multivariate normality already discussed, this should be feasible in principle.

In fact, this could be carried out by simulation. If we were to simulate realizations $(\epsilon^{(s)\,T}, V_1^{(s)\,T}, \dots, V_T^{(s)\,T})^T$, $s = 1, \dots, S$, for large S, from a multivariate normal model for $(\epsilon^T, V_1^T, \dots, V_T^T)^T$ given X, we can estimate $\mathcal{V}^*(x)$ as

$$S^{-1}\sum_{s=1}^{S}\left\{ \frac{\epsilon^{(s)}\epsilon^{(s)\,T}}{\overline{\pi}_T(H_{T-1}^{(s)})} - \sum_{j=1}^{T-1} \frac{\lambda_{j+1}(H_j^{(s)})}{\overline{\pi}_{j+1}(H_j^{(s)})} E(\epsilon|H_j^{(s)})\epsilon^{(s)\,T} \right\}.$$

Whether or not this is a feasible strategy, and whether or not going to all of this trouble will yield an estimator for β that offers a ***nonnegligible gain in relative efficiency*** over the methods discussed previously is an ***open research problem***.

5.5 Discussion

Inverse probability weighted methods are a natural approach to analysis under MAR dropout in problems where a ***semiparametric full data model*** is of interest. As long as the dropout mechanism is ***correctly modeled*** (so that the dropout hazards are correct for each time point), methods based on inverse weighted (non-augmented) estimating functions will lead to consistent estimators for parameters in the semiparametric model of interest. These can be implemented straightforwardly in practice, and there is emerging software to do so.

Where doubly robust, augmented inverse probability weighted estimators are ***feasible*** in practice, these offer protection against misspecification of these models. However, as we have seen, in all but the simplest settings, these can be rather challenging to implement. See Seaman and Copas (2009) and Vansteelandt, Carpenter, and Kenward (2010) for discussion and simulation studies of the extent of improvement possible.

STANDARD ERRORS: We have not discussed how to obtain ***standard errors*** for any of the estimators. ***In principle***, because all of these estimators are ***M-estimators***, as reviewed in Section 1.5 of Chapter 1, it is possible to derive the form of the asymptotic covariance matrix for the estimator for the parameter of interest using the sandwich technique as in (1.42) and (1.43).

- Here, one must take account of the fact that the parameter of interest is estimated ***jointly*** with the parameters in the dropout models and working covariance model by solving accompanying estimating equations for these parameters.

  Thus, application of the sandwich technique should be to ***all*** of these equations, "***stacked***;" see, for example, Theorem 1 of Robins et al. (1995) and Section 2 of Preisser et al. (2002).

- According to the documentation for SAS `proc gee`, this is implemented in this procedure when the `missmodel` statement is invoked.

- It is well known from semiparametric theory that ***ignoring*** the fact that the ***weights*** are ***estimated*** and treating them as fixed (as would be the default for usual GEE software such as SAS `proc genmod`) leads to standard errors that are ***conservative*** and thus ***understate*** the precision with which parameters of interest are estimated.

- As always, an alternative to all of this is to employ a ***nonparametric bootstrap***.

5.6 Semiparametric theory

As we have discussed, the foundation for the methods discussed in this chapter is the ***theory of semiparametrics***. A comprehensive account of how application of this theory leads to the classes of estimators involving inverse probability weighting we have reviewed is the subject of an entire course. The book by Tsiatis (2006) is a seminal reference on this topic and provides a detailed account of semiparametric theory and its application to deriving such estimators.

In this section, we present a basic introduction to the geometric principles underlying this powerful theory. We draw upon brief accounts given by Davidian, Tsiatis, and Leon (2005) and Kennedy (2016), which can be consulted for more details.

PARAMETRIC AND SEMIPARAMETRIC MODELS: We first review the notion of a ***statistical model*** introduced in Section 1.3. A statistical model is a class of probability distributions that is assumed to have generated the data (and thus is assumed to contain the true distribution).

In the context of generic data Z, with density $p_Z(z)$, a ***parametric model*** is one in which the class of probability distributions is indexed by a finite-dimensional parameter θ ($q \times 1$), so involving densities of the form $p_Z(z;\theta)$, and the goal is to make inference on θ or, partitioning θ as $\theta = (\beta^T, \eta^T)^T$, on a subset β of interest. In a ***semiparametric model*** for $p_z(z)$, the class of probability distributions is indexed by finite-dimensional and infinite-dimensional components, so involves densities of the form $p_Z\{z;\beta,\eta(\cdot)\}$, where β is a finite-dimensional parameter of interest and $\eta(\cdot)$ is an infinite-dimensional component.

Semiparametric theory leads to a class of estimators for β based on iid data, from which the (asymptotically) efficient estimator in the class can be deduced. In both parametric and semiparametric models as defined here, we view the finite-dimensional parameter η in a parametric model or the infinite-dimensional $\eta(\cdot)$ as a ***nuisance parameter***, which is not of central interest but must be dealt with in making inference on β.

INFLUENCE FUNCTIONS: In this theory, attention is restricted to estimators that are ***regular and asymptotically linear*** (RAL), as introduced briefly in Section 4.9. ***Regularity*** is a technical condition that rules out "pathological" estimators, such as "superefficient" estimators (see Tsiatis, 2006, Section 3.1). Generically, if we have a parametric or semiparametric model for Z that contains the true distribution generating the data, and if β_0 is the associated true value of β, an ***asymptotically linear estimator*** for β based on iid data Z_i, $i = 1, \ldots, N$, satisfies

$$N^{1/2}(\widehat{\beta} - \beta_0) = N^{-1/2}\sum_{i=1}^{N} \varphi(Z_i) + o_p(1), \tag{5.50}$$

where $o_p(1)$ represents terms that converge in probability to zero as $N \to \infty$, and the function $\varphi(Z)$ is referred to as the ***influence function*** of $\widehat{\beta}$. The influence function satisfies

$$E\{\varphi(Z)\} = 0, \qquad E\{\varphi(Z)^T\varphi(Z)\} < \infty,$$

where expectation is with respect to the true distribution of Z.

It is immediate from (5.50) by the central limit theorem that an asymptotically linear estimator with influence function $\varphi(Z)$ is consistent and asymptotically normal with mean zero and covariance matrix

$$E\{\varphi(Z)\varphi(Z)^T\}.$$

There exists an influence function $\varphi^{eff}(Z)$, say, such that

$$E\{\varphi(Z)\varphi(Z)^T\} - E\{\varphi^{eff}(Z)\varphi^{eff}(Z)^T\}$$

is nonnegative definite for any other influence function $\varphi(Z)$, and $\varphi^{eff}(Z)$ is referred to as the ***efficient influence function***.

RELATIONSHIP BETWEEN ASYMPTOTICALLY LINEAR ESTIMATORS AND INFLUENCE FUNCTIONS: It is straightforward to show by ***contradiction*** that a RAL estimator satisfying (5.50) has a ***unique*** (almost surely) influence function. If this were not the case, then there would exist another influence function $\varphi^*(Z)$, say, with $E\{\varphi^*(Z)\} = 0$ that also satisfies (5.50). If (5.50) holds for both $\varphi(Z)$ and $\varphi^*(Z)$, then it must be that

$$A_N = N^{-1/2}\sum_{i=1}^{N}\{\varphi(Z_i) - \varphi^*(Z_i)\} = o_p(1);$$

that is, $A_N \xrightarrow{p} 0$. However, by the central limit theorem, it is also the case that

$$A_N \xrightarrow{\mathcal{L}} \mathcal{N}\Big(0, E\Big[\{\varphi(Z) - \varphi^*(Z)\}\{\varphi(Z) - \varphi^*(Z)\}^T\Big]\Big).$$

For A_N to converge to probability to zero and in distribution, it must be that

$$E\Big[\{\varphi(Z) - \varphi^*(Z)\}\{\varphi(Z) - \varphi^*(Z)\}^T\Big] = 0,$$

which implies that $\varphi(Z) = \varphi^*(Z)$ almost surely.

This result demonstrates that there is a ***one-to-one correspondence*** between asymptotically linear estimators and influence functions. This suggests that, by identifying influence functions, we can deduce estimators. Thus, characterizing the ***space of all influence functions*** is the fundamental premise underlying semiparametric theory. Once this space is deduced, we can then construct estimators and contrast them on the basis of efficiency.

GEOMETRIC PERSPECTIVE ON THE PARAMETRIC MODEL: We introduce the ideas first in the context of a ***parametric model*** $p_Z(z;\theta)$, $\theta = (\beta^T, \eta^T)^T$ as above, with β $(p \times 1)$ and η $(r \times 1)$, which we assume to be ***correctly specified*** in the usual sense that there exists $\theta_0 = (\beta_0^T, \eta_0^T)^T$ such that $p_Z(z;\theta_0)$ is the true density of Z. To simplify the notation, we take the parameter β to be ***one-dimensional***. Although ultimately we are concerned with our missing data problem, the following demonstration treats Z and the iid data Z_i, $i = 1, \dots, N$, as generic, so we do not, for example, use notation distinguishing full versus observed data.

MAXIMUM LIKELIHOOD IN A PARAMETRIC MODEL: We demonstrate the basic ideas of this geometric perspective in the context of the familiar setting of maximum likelihood. We take $\theta = (\beta, \eta^T)^T$, so that $q = 1 + r$, and define the score vector

$$S_\theta(Z;\theta) = \{S_\beta(Z;\theta), S_\eta(Z;\theta)^T\}^T = \left[\frac{\partial \log\{p_Z(Z;\theta)\}}{\partial\beta}, \frac{\partial \log\{p_Z(Z;\theta)\}}{\partial\eta^T}\right]^T.$$

As is well known, the ***expected information matrix*** is of the form

$$\mathcal{I}(\theta_0) = E\{S_\theta(Z;\theta_0)S_\theta(Z;\theta_0)^T\} = \begin{pmatrix} \mathcal{I}_{\beta\beta} & \mathcal{I}_{\beta\eta} \\ \mathcal{I}_{\beta\eta}^T & \mathcal{I}_{\eta\eta} \end{pmatrix}, \quad \mathcal{I}_{\eta\eta}\ (r \times r),\ \ \mathcal{I}_{\beta\eta}\ (1 \times r). \tag{5.51}$$

Letting $\widehat{\theta} = (\widehat{\beta}, \widehat{\eta}^T)^T$ denote the ***maximum likelihood estimator*** for θ maximizing $\sum_{i=1}^N \log\{p_Z(Z_i;\theta)\}$, it is well known that, under the usual regularity conditions,

$$N^{1/2}(\widehat{\beta} - \beta_0) = N^{-1/2}\sum_{i=1}^N \varphi^{eff}(Z_i) + o_p(1), \tag{5.52}$$

$$\varphi^{eff}(Z) = \mathcal{I}_{\beta\beta\bullet\eta}^{-1}\{S_\beta(Z;\theta_0) - \mathcal{I}_{\beta\eta}\mathcal{I}_{\eta\eta}^{-1}S_\eta(Z;\theta_0)\}, \quad \mathcal{I}_{\beta\beta\bullet\eta} = \mathcal{I}_{\beta\beta} - \mathcal{I}_{\beta\eta}\mathcal{I}_{\eta\eta}^{-1}\mathcal{I}_{\beta\eta}^T. \tag{5.53}$$

Clearly, $E\{\varphi^{eff}(Z)\} = 0$, and thus the maximum likelihood estimator $\widehat{\beta}$ for β has ***influence function*** $\varphi^{eff}(Z)$. The quantity in braces in (5.53),

$$S^{eff}(Z) = S_\beta(Z;\theta_0) - \mathcal{I}_{\beta\eta}\mathcal{I}_{\eta\eta}^{-1}S_\eta(Z;\theta_0),$$

is often called the ***efficient score***. It is straightforward to show that (try it) $S^{eff}(Z)$ has mean zero and variance $\mathcal{I}_{\beta\beta\bullet\eta}$. It then follows from (5.52) that $\widehat{\beta}$ is asymptotically normal with asymptotic variance

$$E\{\varphi^{eff}(Z)^2\} = \mathcal{I}_{\beta\beta\bullet\eta}^{-1},$$

which is the well known ***Cramér-Rao lower bound*** , the smallest possible variance. Thus, as is also well known, $\widehat{\beta}$ is the ***efficient estimator*** for β, and we have accordingly referred to its influence function as $\varphi^{eff}(Z)$.

Having stated these fundamental results, we now derive them from the ***geometric perspective*** that underlies semiparametric theory.

HILBERT SPACE: A ***Hilbert space*** $\mathcal{H}$ is a linear vector space, so has the property that

$$ah_1 + bh_2 \in \mathcal{H} \quad \text{for} \quad h_1, h_2 \in \mathcal{H}$$

and any real a, b, equipped with an ***inner product***.

The key feature underlying the geometric perspective is that ***influence functions*** based on data Z for estimators for a p-dimensional parameter β in a (parametric or semiparametric) statistical model can be viewed as elements in the particular Hilbert space $\mathcal{H}$ of all ***p-dimensional, mean zero (measurable) functions*** $h(Z)$, so with $E\{h(Z)\} = 0$, such that $E\{h(Z)^T h(Z)\} < \infty$, with ***covariance inner product***

$$E\{h_1(Z)^T h_2(Z)\} \quad \text{for} \quad h_1, h_2 \in \mathcal{H},$$

and corresponding ***norm***, measuring "distance" from $h \equiv 0$, given by

$$\|h\| = [E\{h(Z)^T h(Z)\}]^{1/2}.$$

A Hilbert space can thus be viewed as a ***generalization*** of usual Euclidean space and provides notions of ***distance*** and ***direction*** for spaces whose elements are potentially ***infinite-dimensional functions***. The geometry of Hilbert spaces provides a unified framework for deducing results regarding influence functions in both parametric and semiparametric models.

We state some important general results concerning Hilbert spaces;

- A space $M \subset \mathcal{H}$ is a ***linear subspace*** if $m_1, m_2 \in M$ implies that $am_1 + bm_2 \in M$ for all scalar a, b. The origin is included by default ($a = b = 0$). For example, if $h_1, \ldots, h_k$ are arbitrary elements of $\mathcal{H}$, then the space comprising elements of the form $a_1 h_1 + \cdots + a_k h_k$ for all $(a_1, \ldots, a_k) \in \mathbb{R}^k$ is a linear subspace spanned by $\{h_1, \ldots, h_k\}$.
- For any ***linear subspace*** M of $\mathcal{H}$, the set of all elements of $\mathcal{H}$ ***orthogonal*** to those in M, denoted $M^\perp$ (i.e., such that if $h_1 \in M$ and $h_2 \in M^\perp$, the inner product of h_1, h_2 is zero), is ***also*** a linear subspace of $\mathcal{H}$.
- For two linear subspaces M and N, $M \oplus N$ is the ***direct sum*** of M and N if every element in $M \oplus N$ has a unique representation of the form $m + n$ for $m \in M$, $n \in N$. Intuitively, it is the case that the entire Hilbert space $\mathcal{H} = M \oplus M^\perp$.
- An essential concept is the notion of a ***projection***. The projection of $h \in \mathcal{H}$ onto a closed linear subspace M of $\mathcal{H}$ is the element in M, denoted by $\Pi(h|M)$, such that

  $$\|h - \Pi(h|M)\| < \|h - m\| \quad \text{for all} \quad m \in M.$$

  The ***residual*** $h - \Pi(h|M)$ is orthogonal to all $m \in M$.
- The ***Projection Theorem for Hilbert Spaces*** states that such a projection is ***unique***; this is shown in Luenberger (1969, Section 3.3). The book by Luenberger (1969) is a seminal reference on Hilbert spaces.

An important subspace of $\mathcal{H}$ is the ***tangent space***. For our parametric model, the tangent space Γ is defined as the linear subspace of $\mathcal{H}$ spanned by the entire score vector $S_\theta(Z;\theta_0)$; i.e.,

$$\Gamma = \{BS_\theta(Z;\theta_0) \quad \text{for all} \quad (1 \times q) \;\; B\},$$

the space of all linear combinations of $S_\theta(Z;\theta_0)$. The tangent space can be decomposed as

$$\Gamma = \Gamma_\beta \oplus \Lambda, \qquad \Gamma_\beta = \{BS_\beta(Z;\theta_0) \text{ for all real-valued } B\} \tag{5.54}$$

and

$$\Lambda = \{BS_\eta(Z;\theta_0) \quad \text{for all} \quad (1 \times r) \;\; B\}, \tag{5.55}$$

the linear subspace spanned by the score vector of the ***nuisance parameter*** η. Accordingly, the space Λ in (5.55) is referred to as the ***nuisance tangent space***.

FUNDAMENTAL RESULT: All influence functions for RAL estimators for β lie in the subspace $\Lambda^\perp$ ***orthogonal to the nuisance tangent space***.

Although the proof of this result is beyond our scope here, it is straightforward to provide an ***example*** by demonstrating that the efficient influence function φ^{eff} in (5.52) lies in $\Lambda^\perp$. We must show that

$$E\{\varphi^{\mathit{eff}}(Z)^T BS_\eta(Z;\theta_0)\} = E[\{S_\beta(Z;\theta_0) - \mathcal{I}_{\beta\eta}\mathcal{I}_{\eta\eta}^{-1}S_\eta(Z;\theta_0)\}^T BS_\eta(Z;\theta_0)]/\mathcal{I}_{\beta\beta\bullet\eta} = 0$$

for all B $(1 \times r)$. By taking B to successively be a $(1 \times r)$ vector with a "1" in one component and "0"s elsewhere, this may be seen to be equivalent to showing that

$$E[\{S_\beta(Z;\theta_0) - \mathcal{I}_{\beta\eta}\mathcal{I}_{\eta\eta}^{-1}S_\eta(Z;\theta_0)\}S_\eta(Z;\theta_0)^T] = 0,$$

which follows immediately (try it).

Thus, one approach to identifying influence functions for a particular model with $\theta = (\beta, \eta^T)^T$ is to characterize the form of elements in $\Lambda^\perp$ directly. Alternatively, the following result suggests another approach to characterizing influence functions.

REPRESENTATION OF INFLUENCE FUNCTIONS: All influence functions for RAL estimators for β can be represented as

$$\varphi(Z) = \varphi^*(Z) + \psi(Z), \tag{5.56}$$

where $\varphi^*(Z)$ is any influence function and $\psi(Z) \in \Gamma^\perp$, the subspace of $\mathcal{H}$ ***orthogonal*** to Γ.

DEFINING PROPERTIES OF INFLUENCE FUNCTIONS: We can demonstrate this representation by appealing to two ***defining properties*** of influence functions $\varphi(Z)$, which are related to regularity and whose proofs are beyond our scope here (see Tsiatis, 2006, Sections 3.1 and 3.2).

(i) $E\{\varphi(Z)S_\beta(Z;\theta_0)\} = 1$

(ii) $E\{\varphi(Z)S_\eta(Z;\theta_0)^T\} = 0$, $(1 \times r)$.

We used the general version of property (i) in the derivation of the large sample distribution of the improper imputation estimator in Section 4.9. If $\varphi(Z)$ is an influence function, (i) and (ii) must hold; conversely, if (i) and (ii) hold, $\varphi(Z)$ must be an influence function.

To show (5.56), we first demonstrate that if $\varphi(Z)$ can be written as $\varphi^*(Z) + \psi(Z)$ as in (5.56) , where $\varphi^*(Z)$ is any influence function and $\psi(Z) \in \Gamma^\perp$, then $\varphi(Z)$ is an influence function. From the definition of Γ_β in (5.54), if $\psi \in \Gamma^\perp$, then $\psi(Z)$ is orthogonal to functions in both Λ and Γ_β, so that $E\{\psi(Z)S_\beta(Z;\theta_0)\} = 0$ and $E\{\psi(Z)S_\eta(Z;\theta_0)^T\} = 0$ $(1 \times r)$. Moreover, because $\varphi^*(Z)$ is an influence function, it satisfies (i) and (ii), from whence it follows that $\varphi(Z)$ also satisfies (i) and (ii) and thus is itself an influence function.

Conversely, we can demonstrate that if $\varphi(Z)$ is an influence function, it can be represented as in (5.56). If $\varphi(Z)$ is an influence function, it must satisfy (i) and (ii), and, writing $\varphi(Z) = \varphi^*(Z) + \{\varphi(Z) - \varphi^*(Z)\}$ for some other influence function $\varphi^*(Z)$, it is straightforward to use (i) and (ii) to show that (try it) $\psi(Z) = \{\varphi(Z) - \varphi^*(Z)\} \in \Gamma^\perp$.

The representation (5.56) also implies a useful characterization of the ***efficient influence function*** $\varphi^{eff}(Z)$, which here satisfies $E\{\varphi(Z)^2\} - E\{\varphi^{eff}(Z)^2\} \geq 0$ for all influence functions $\varphi(Z)$.

REPRESENTATION OF THE EFFICIENT INFLUENCE FUNCTION: $\varphi^{eff}(Z)$ can be represented as

$$\varphi^{eff}(Z) = \varphi(Z) - \Pi(\varphi|\Gamma^\perp)(Z)$$

for any influence function $\varphi(Z)$.

This follows because, for arbitrary $\varphi(Z)$, $\varphi^{eff}(Z) = \varphi(Z) - \psi(Z)$ for $\psi \in \Gamma^\perp$, and $E\{\varphi^{eff}(Z)^2\} = \|\varphi - \psi\|$ must be as small as possible, it must be that $\psi = \Pi(\varphi|\Gamma^\perp)$.

In our simple parametric model with $\theta = (\beta, \eta^T)^T$, it is possible to identify ***explicitly*** the form of the efficient influence function $\varphi^{eff}(Z)$. Here, the ***efficient score*** is defined as the residual of the score vector for β after ***projecting*** it onto the nuisance tangent space,

$$S^{eff}(Z;\theta_0) = S_\beta(Z;\theta_0) - \Pi(S_\beta|\Lambda),$$

and the ***efficient influence function*** is an appropriately-scaled version of $S^{eff}(Z)$ given by

$$\varphi^{eff}(Z) = \Big[E\{S^{eff}(Z;\theta_0)^2\}\Big]^{-1} S^{eff}(Z;\theta_0).$$

- It is straightforward to observe that $\varphi^{eff}(Z)$ is an influence function by showing it satisfies properties (i) and (ii) above. Specifically, by construction, $S^{eff}(Z) \in \Lambda^\perp$, so that (ii) holds. This implies $E\{\varphi^{eff}(Z)\Pi(S_\beta|\Lambda)(Z)\} = 0$, so that

$$\begin{aligned} E\{\varphi^{eff}(Z)S_\beta(Z;\theta_0)\} &= E\{\varphi^{eff}(Z)S^{eff}(Z;\theta_0)\} + E\{\varphi^{eff}(Z)\Pi(S_\beta|\Lambda)(Z)\} \\ &= \Big[E\{S^{eff\,2}(Z;\theta_0)\}\Big]^{-1} E\{S^{eff\,2}(Z;\theta_0)\} = 1, \end{aligned}$$

  demonstrating property (i).

- That $\varphi^{eff}(Z)$ has ***smallest variance*** among all influence functions can be seen by using the fact that all influence functions can be written as $\varphi(Z) = \varphi^{eff}(Z) + \psi(Z)$ for some $\psi(Z) \in \Gamma^\perp$. Because $S_\beta \in \Gamma_\beta$, $\Pi(S_\beta|\Lambda) \in \Lambda$ are both in Γ, it follows that $E\{\psi(Z)\varphi^{eff}(Z)\} = 0$. Thus, $E\{\varphi(Z)^2\} = E[\{\varphi^{eff}(Z) + \psi(Z)\}^2] = E\{\varphi^{eff}(Z)^2\} + E\{\psi(Z)^2\}$, so that any other influence function $\varphi(Z)$ has variance at least as large as that of $\varphi^{eff}(Z)$, and this smallest variance is immediately seen to be $1/S^{eff}(Z;\theta_0)^2$.

MAXIMUM LIKELIHOOD IN A PARAMETRIC MODEL, REVISITED: We can place the familiar maximum likelihood results when $\theta = (\beta, \eta^T)^T$ above in this framework. By definition, $\Pi(S_\beta|\Lambda) \in \Lambda$ is the ***unique*** element $B_0 S_\eta \in \Lambda$ such that

$$E[\{S_\beta(Z;\theta_0) - B_0 S_\eta(Z;\theta_0)\} B S_\eta(Z;\theta_0)] = 0 \text{ for all } B \;\; (1 \times r).$$

As above, this is equivalent to requiring

$$E[\{S_\beta(Z;\theta_0) - B_0 S_\eta(Z;\theta_0)\} S_\eta(Z;\theta_0)^T] = 0 \;\; (1 \times r),$$

implying that $B_0 = \mathcal{I}_{\beta\eta}\mathcal{I}_{\eta\eta}^{-1}$. Thus, as expected,

$$\Pi(S_\beta|\Lambda) = \mathcal{I}_{\beta\eta}\mathcal{I}_{\eta\eta}^{-1} S_\eta(Z;\theta_0) \quad \text{and} \quad S^{eff}(Z;\theta_0) = S_\beta(Z;\theta_0) - \mathcal{I}_{\beta\eta}\mathcal{I}_{\eta\eta}^{-1} S_\eta(Z;\theta_0).$$

For a parametric model, it is usually ***unnecessary*** to appeal to the foregoing geometric construction to identify the efficient estimator and influence functions. In contrast, in the more complex case of a semiparametric model, such results often can not be derived readily. However, as we now discuss, the geometric perspective may be generalized to semiparametric models, providing a systematic framework for identifying influence functions.

GEOMETRIC PERSPECTIVE ON THE SEMIPARAMETRIC MODEL: We now consider the generalization of the above results for parametric models to the ***semiparametric model*** characterized by the class of distributions $\mathcal{P}$, say, comprising densities of the form $p\{z;\beta,\eta(\,\cdot\,)\}$ depending on an infinite-dimensional parameter $\eta(\,\cdot\,)$. An even more general formulation is discussed by Davidian et al. (2005). We only present a sketch of the main ideas; a precise and detailed account is given in Tsiatis (2006, Chapter 4).

PARAMETRIC SUBMODEL: The key to the generalization is the notion of a ***parametric submodel***. In words, a ***parametric submodel*** is a parametric model contained in the semiparametric model that contains the truth, where the truth is the density $p\{z;\beta_0,\eta_0(\cdot)\} \in \mathcal{P}$ generating the data.

More formally, a parametric submodel is the class $\mathcal{P}_{\beta,\xi}$ of all densities $p(z;\beta,\xi)$ characterized by the finite-dimensional parameter $(\beta^T,\xi^T)^T$ such that

(i) $\mathcal{P}_{\beta,\xi} \subset \mathcal{P}$, so that every density in $\mathcal{P}_{\beta,\xi}$ belongs to the semiparametric model $\mathcal{P}$.

(ii) The parametric submodel contains the truth in the sense that there exists a density in $\mathcal{P}_{\beta,\xi}$ identified by $(\beta_0^T,\xi_0^T)^T$ such that the true density is

$$p\{z;\beta_0,\eta_0(\,\cdot\,)\} = p(z;\beta_0,\xi_0) \in \mathcal{P}_{\beta,\xi}.$$

The dimension r of ξ varies according to the choice of submodel. We again take β to be one-dimensional for simplicity.

A parametric submodel is ***not*** a model one would use for data analysis; given that we ***do not know*** the truth, it is not possible to specify a parametric submodel for use in practice. Rather, it is a ***conceptual device*** that is used to develop theory for semiparametric models.

Because a parametric submodel is a ***parametric model***, albeit one not useful in practice, we know from the development above for parametric models that

(i) Influence functions for RAL estimators for β in a parametric submodel belong to the subspace of the Hilbert space $\mathcal{H}$ whose elements are orthogonal to the ***parametric submodel nuisance tangent space***

$$\Lambda_\xi = \{BS_\xi(Z;\beta_0,\xi_0) \quad \text{for all} \quad (1 \times r) \;\; B\}, \qquad S_\xi(Z;\beta,\xi) = \frac{\partial \log\{p_Z(Z;\beta,\xi)\}}{\partial \xi^T}.$$

(ii) The ***efficient influence function for the parametric submodel*** is given by

$$\varphi^{\mathit{eff}}_{\beta,\xi}(Z) = \Big[E\{S^{\mathit{eff}}_{\beta,\xi}(Z;\beta_0,\xi_0)^2\}\Big]^{-1} S^{\mathit{eff}}_{\beta,\xi}(Z;\beta_0,\xi_0),$$

where the parametric model efficient score is

$$S^{\mathit{eff}}_{\beta,\xi}(Z) = S^{\mathit{eff}}_{\beta,\xi}(Z;\beta_0,\xi_0) = S_\beta(Z;\beta_0,\xi_0) - \Pi\{S_\beta(Z;\beta_0,\xi_0)|\Lambda_\xi\}.$$

(iii) The ***smallest asymptotic variance*** for RAL estimators for β in the parametric submodel is

$$\Big[E\{S^{\mathit{eff}}_{\beta,\xi}(Z)^2\}\Big]^{-1}. \tag{5.57}$$

INFLUENCE FUNCTIONS FOR RAL ESTIMATORS THE SEMIPARAMETRIC MODEL: An estimator is an (RAL) estimator for β under the ***semiparametric model*** if it is a RAL estimator under ***every parametric submodel***. Thus, the class of estimators for β for the semiparametric model ***must be contained in*** the class of estimators for a parametric submodel and thus any influence function for the semiparametric model must be an influence function for a parametric submodel. It follows that

- Any influence function of a RAL estimator for β for the semiparametric model must be ***orthogonal*** to all parametric submodel nuisance tangent spaces.
- Accordingly, the ***semiparametric model nuisance tangent space*** $\Lambda \subset \mathcal{H}$ is defined as the ***mean square closure*** of all parametric submodel nuisance tangent spaces Λ_ξ; that is, $\Lambda = \{h \in \mathcal{H}$ such that there exists a sequence $B_j S_{\xi_j}(Z)$ for which $\|h(Z) - B_j S_{\xi_j}(Z)\|^2 \to 0$ as $j \to \infty\}$. See Tsiatis (2006, Section 4.4) for details. It can then be shown that ***all influence functions*** for the semiparametric model lie in $\Lambda^\perp$, the space orthogonal to the semiparametric model nuisance tangent space.

- Thus, the ***variance*** of any ***semiparametric model influence function*** must be greater than or equal to (5.57) for all parametric submodels $\mathcal{P}_{\beta,\xi}$. That is, the variance of the influence function of any semiparametric estimator for β must be greater than or equal to the ***supremum*** of (5.57) over all parametric submodels; that is,

$$\sup_{\mathcal{P}_{\beta,\xi}} \left[E\{S^{eff}_{\beta,\xi}(Z)^2\} \right]^{-1}. \tag{5.58}$$

  The supremum (5.58) is defined to be the ***semiparametric efficiency bound***; i.e., the supremum over all parametric submodels of the Cramér-Rao lower bounds. Any semiparametric RAL estimator for β with asymptotic variance achieving this bound is said to be ***locally efficient*** .

Geometrically, the parametric submodel efficient score is the ***residual*** of $S_\beta(Z;\beta_0,\xi_0)$ after projecting it onto the parametric submodel nuisance tangent space Λ_ξ. With β one-dimensional as we are taking it to be here, the inverse of the norm squared of this residual is the smallest variance for all influence functions for RAL estimators for β in the parametric submodel. As we consider the linear space spanned by the nuisance tangent spaces of all parametric submodels, the space becomes larger and the norm of the residual becomes smaller, so that the variance (inverse of norm squared) becomes larger. As a result, the efficient semiparametric estimator has variance larger than the efficient estimator for any parametric submodel.

The foregoing developments can be made precise; this is way beyond our scope here. The result is that the key to deriving influence functions for the semiparametric model is to ***identify the semiparametric model nuisance tangent space*** Λ and the form of elements of $\mathcal{H}$ that are orthogonal to it. Tsiatis (2006, Section 4.5) presents a detailed demonstration for the semiparametric model we have discussed previously, namely

$$E(Y|X=x) = \mu(x;\beta). \tag{5.59}$$

FULL DATA INFLUENCE FUNCTIONS: Our development to now has been generic. We now take Z to be the full data in a missing data problem involving a semiparametric full data model $p\{z;\beta,\eta(\cdot)\}$. Denote the Hilbert space of all mean zero functions $h(Z)$ of the full data as $\mathcal{H}^F$. Then we can use the formulation above to identify the class of all ***full data influence functions*** $\varphi^F(Z)$, say, for estimators for β, which reside in the space $\Lambda^{F\perp}$, say, orthogonal to the ***full data nuisance tangent space*** Λ^F.

OBSERVED DATA INFLUENCE FUNCTIONS: Letting as usual $(R, Z_{(R)})$ denote the ***observed data*** under a MAR missingness mechanism, the key advance of Robins, Rotnitzky, and Zhao (1994) was to derive, for a general semiparametric model, the class of all ***observed data influence functions*** and to characterize the ***efficient observed data influence function*** in this class.

Here, the relevant ***Hilbert space*** in which observed data influence functions reside, $\mathcal{H}$, say, is now the space of all ***p-dimensional, mean zero (measurable) functions*** $h(R, Z_{(R)})$, so with

$$E\{h(R, Z_{(R)})\} = 0, \quad E\{h(R, Z_{(R)})^T h(R, Z_{(R)})\} < \infty,$$

with ***covariance inner product***.

$$E\{h_1(R, Z_{(R)})^T h_2(R, Z_{(R)})\} \quad \text{for} \quad h_1, h_2 \in \mathcal{H},$$

and corresponding ***norm***, measuring "distance" from $h \equiv 0$, given by

$$\|h\| = [E\{h(R, Z_{(R)})^T h(R, Z_{(R)})\}]^{1/2}.$$

By the generic formulation, the class of all observed data influence functions $\varphi(R, Z_{(R)})$, say, must lie in the linear subspace $\Lambda^\perp$ of $\mathcal{H}$ orthogonal to the ***observed data nuisance tangent space*** Λ.

OBSERVED DATA NUISANCE TANGENT SPACE: The semiparametric model involves the ***nuisance parameter*** $\eta(\cdot)$. As we know, the likelihood corresponding to the observed data involves the ***missingness mechanism***, for which in practice we might posit a ***model*** involving an additional finite-dimensional parameter ψ. From the current perspective, we can view ψ as an ***additional nuisance parameter***. Because under the ***separability condition*** $\eta(\cdot)$ and ψ are distinct, we expect the observed nuisance tangent space Λ to be the ***direct sum*** of two ***orthogonal spaces***, one involving the space generated by the score vector for ψ Λ_ψ, say; and one for $\eta(\cdot)$ (being the mean square closure of all parametric submodel nuisance tangent spaces spanned by score vector for the submodel parameter ξ), Λ_η, say, so that

$$\Lambda = \Lambda_\eta \oplus \Lambda_\psi.$$

The general derivation of Λ and the space $\Lambda^\perp$ orthogonal to it is ***complicated*** and is discussed in detail in Tsiatis (2006, Chapter 6-11).

FORM OF OBSERVED DATA INFLUENCE FUNCTIONS: The resulting theory shows that there is a ***relationship*** between full and observed data influence functions. The form of observed data influence functions involves an ***inverse probability weighted*** full data influence function plus an ***augmentation***. Here, it must be that the probability of observing full data given Z is > 0 for all Z almost surely.

When the missingness mechanism is ***known***, which is the case if MAR missingness is ***by design***, there is no parameter ψ, in which case $\Lambda^{\perp} = \Lambda_{\eta}^{\perp}$ is the space where observed data influence functions reside. ***Theorem 7.2*** of Tsiatis (2006) presents the generic form of these influence functions in this case. In our notation, with $Z = (Z_1, \ldots, Z_K)$, $R = (R_1, \ldots, R_K)$, and

$$\text{pr}(R = r|Z) = \text{pr}(R = r|Z_{(r)}) = \pi(r, Z_{(r)})$$

under MAR, and letting $\underset{\sim}{1}$ be a K-vector of all ones, under the assumption that

$$\pi(\underset{\sim}{1}, Z) > 0 \quad \text{for all } Z \text{ almost surely,}$$

the space $\Lambda^{\perp} = \Lambda_{\eta}^{\perp}$ consists of all elements of $\mathcal{H}$ that can be written as

$$\frac{\mathcal{I}(R = \underset{\sim}{1})\varphi^F(Z)}{\pi(\underset{\sim}{1}, Z)} + \frac{\mathcal{I}(R = \underset{\sim}{1})}{\pi(\underset{\sim}{1}, Z)} \left\{ \sum_{r \neq \underset{\sim}{1}} \pi(r, Z_{(r)}) L_{2r}(Z_{(r)}) \right\} - \sum_{r \neq \underset{\sim}{1}} \mathcal{I}(R = r) L_{2r}(Z_{(r)}), \tag{5.60}$$

where $\varphi^F(Z)$ is an arbitrary element of $\Lambda^{F\perp}$ (full data influence function) and, for $r \neq \underset{\sim}{1}$, $L_{2r}(Z_{(r)})$ is an arbitrary function of $Z_{(r)}$.

EXAMPLE: SEMIPARAMETRIC REGRESSION MODEL: We demonstrate (5.60) in the special case of regression in Section 5.2, where we now write the ***full data*** as $Z = \{Y, (X, V)\}$, where Y is a scalar outcome, X is a vector of covariates of interest, and V is a set of additional, auxiliary covariates; and we are interested in the semiparametric model characterized by (5.59),

$$E(Y|X = x) = \mu(x; \beta)$$

for β $(p \times 1)$. Assuming as in Section 5.2 that Y can be missing via a MAR mechanism and (X, V) is always observed, $K = 2$, and $R = (R_1, R_2)$. In Section 5.2, we defined $C = 1$ when $R = \underset{\sim}{1}$ and $C = 0$ if $R = (0, 1)$, which are the only two possible values r that R can take on.

It can be shown that all elements of $\Lambda^{F\perp}$, and thus all full data influence functions, are of the form

$$\mathcal{A}(X)\{Y-\mu(X;\beta)\},$$

where $\mathcal{A}(X)$ ($p\times 1$) is an arbitrary p-dimensional function of X; see Tsiatis (2006, Section 4.5). Letting

$$\mathrm{pr}(C=1|X,V)=\pi(X,V)$$

as in Section 5.2, it follows from (5.60) that elements of $\Lambda^{\perp}=\Lambda_{\eta}^{\perp}$, that is, observed data influence functions, are of the form

$$\begin{aligned}\frac{C}{\pi(X,V)}&\Big[\mathcal{A}(X)\{Y-\mu(X;\beta)\}+\{1-\pi(X,V)\}L_2(X,V)\Big]-(1-C)L_2(X,V)\\&=\frac{C}{\pi(X,V)}\mathcal{A}(X)\{Y-\mu(X;\beta)\}+\frac{C-\pi(X,V)}{\pi(X,V)}L_2(X,V),\end{aligned} \tag{5.61}$$

where (5.61) follows by straightforward algebra.

When the missingness mechanism ***unknown*** and is modeled, and the model involves the parameter ψ. when ψ is estimated by maximum likelihood, Theorem 9.1 of Tsiatis (2006) gives the form of the observed data influence functions.

Finding the ***efficient observed data influence function*** in either case is very difficult in general. Theorem 10.1 of Tsiatis (2006) gives the generic form of the optimal observed data influence function if we restrict attention to the class of observed data influence functions involving a ***fixed full data influence function*** $\varphi^F(Z)$. However, it is ***not necessarily the case*** that the efficient observed data influence function involves the efficient full data influence function.

The application of semiparametric theory to the missing data problem is the subject of an ***entire course***. The brief review in this chapter is meant to give a sense of the considerations involved.