CHAPTER 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 $\mu$ if we had a sample of full data $(Y_i, V_i), i = 1, \ldots, N$, would be the sample mean of the $Y_i$

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

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

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

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 $(C_i, C_i Y_i, V_i), i = 1, \ldots, N$.

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

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

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

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

is in general **not** a consistent estimator for $\mu$.

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

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

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**

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

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

114
Recall from our review of estimating equations in Section 1.5 that, to show that (5.4) leads to a consistent estimator for \( \mu \), 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:

\[
E_\mu \left\{ \frac{C}{\pi(V)}(Y - \mu) \right\} = E_\mu \left[ E \left\{ \frac{C}{\pi(V)}(Y - \mu) \mid Y, V \right\} \right] \\
= E_\mu \left\{ \frac{E(C \mid Y, V)}{\pi(V)}(Y - \mu) \right\} \\
= E_\mu \left\{ \frac{\pi(V)}{\pi(V)}(Y - \mu) \right\} \\
= E_\mu (Y - \mu) = 0, 
\tag{5.6}
\]

where (5.6) follows because

\[
E(C \mid Y, V) = \text{pr}(C = 1 \mid Y, V) = \text{pr}(C = 1 \mid V) = \pi(V)
\]

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

\[
\hat{\mu}_{ipw}^{ipw2} = \left\{ \sum_{i=1}^{N} \frac{C_i}{\pi(V_i)} \right\}^{-1} \sum_{i=1}^{N} \frac{C_i Y_i}{\pi(V_i)}. 
\tag{5.8}
\]

Comparing (5.8) to the estimator

\[
\hat{\mu}_{ipw} = N^{-1} \sum_{i=1}^{N} \frac{C_i Y_i}{\pi(V_i)}
\]

given in (1.22) shows that they are not the same. The estimator \( \hat{\mu}_{ipw2} \) 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 \( \hat{\mu}_{ipw} \) 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 \textit{by design}, as for the nutrition study discussed in Chapter 1 in which subjects are selected for a \textit{validation sample} according to some known mechanism, then \( \pi(v) \) would indeed be known.

However, in most situations, \( \pi(v) \) is \textit{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 \textit{model} for \( \Pr(C = 1|V) \); for example, a fully parametric model

\[
\pi(V; \psi),
\]

depending on a finite-dimensional parameter \( \psi \), say, and estimate \( \psi \) from the data. Because \( C \) is \textit{binary}, a natural approach is to posit a \textit{logistic regression model} such as

\[
\logit\left\{ \Pr(C = 1|V) \right\} = \psi_0 + \psi^T V,
\]

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

Using the data \((C_i, V_i), i = 1, ..., N\), we can then estimate \( \psi \) by \textit{maximum likelihood}, obtaining the MLE \( \hat{\psi} \) by maximizing

\[
\prod_{i=1}^{N} \left\{ \pi(V_i; \hat{\psi}) \right\}^{C_i} \{1 - \pi(V_i; \hat{\psi})\}^{1-C_i} = \prod_{i=1}^{N} \left\{ \pi(V_i; \hat{\psi}) \right\}^{C_i} \{1 - \pi(V_i; \hat{\psi})\}^{1-C_i}. \tag{5.9}
\]

In this case, the \textit{inverse probability weight complete case estimator} based on (5.8) is given by

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

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

• Moreover, by construction, inverse probability weighted complete case estimators such as (5.8) use data \textit{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 \textit{inefficiency}.

\textbf{AUGMENTED INVERSE PROBABILITY WEIGHTED ESTIMATORS:} It turns out, as shown by Robins et al. (1994), that there is a class of estimators that involves \textit{augmenting} the simple inverse probability weighted complete case estimating equation for \( \mu \). Estimators in this class can yield improved efficiency.
The \textit{optimal} estimator for \( \mu \) within this class is the solution to the estimating equation

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

which, after some algebra, can be written as

\[
\sum_{i=1}^{N} \left\{ \frac{C_i Y_i}{\pi(V_i; \hat{\psi})} - \frac{C_i - \pi(V_i; \hat{\psi})}{\pi(V_i; \hat{\psi})} E(Y_i | V_i) - \mu \right\} = 0, \tag{5.11}
\]

and leads to the estimator

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

In (5.11) and consequently the expression for \( \hat{\mu} \), 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 \textit{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 \( \xi \), on the complete cases \( \{ i : C_i = 1 \} \). Specifically, if \( Y \) is continuous, for example, we might derive an estimator \( \hat{\xi} \) for \( \xi \) by OLS, solving in \( \xi \)

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

the OLS estimating equation based on the complete cases.

Substituting in (5.11), the resulting \textit{augmented inverse probability weighted estimator} for \( \mu \) is

\[
\hat{\mu}^{\text{aipw}} = N^{-1} \sum_{i=1}^{N} \left\{ \frac{C_i Y_i}{\pi(V_i; \hat{\psi})} - \frac{C_i - \pi(V_i; \hat{\psi})}{\pi(V_i; \hat{\psi})} m(V_i; \hat{\xi}) \right\}. \tag{5.13}
\]

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

---

117
**DOUBLE ROBUSTNESS:** It can be shown that the estimator \( \hat{\mu}_{\text{aipw}} \) in (5.13) is a **consistent** estimator for \( \mu \) if **EITHER**

- the model \( \pi(v; \psi) \) for \( \text{pr}(C = 1|V = v) \) 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 \( \hat{\mu}_{\text{aipw}} \) 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, \( \hat{\mu}_{\text{aipw}} \) in (5.13) converges in probability to

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

where \( \psi^* \) and \( \xi^* \) are the limits in probability of \( \hat{\psi} \) and \( \hat{\xi} \). 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, \( \hat{\mu}_{\text{aipw}} \) is a consistent estimator for \( \mu \) if we can show that

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

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

\[
E \left( E \left[ \left\{ \frac{C - \pi(V; \psi^*)}{\pi(V; \psi^*)} \right\} \{ Y - m(V; \xi^*) \} \right| V \right) \right) = E \left[ E \left\{ \frac{C - \pi(V; \psi^*)}{\pi(V; \psi^*)} \right\} \{ Y - m(V; \xi^*) \} \right| V \right],
\] (5.16)

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

Consider two cases:

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

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

Under this condition,

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

using \( E(C|V) = \text{pr}(C = 1|V) \), and (5.15) follows.
(b) \( m(\nu; \xi) \) is **correctly specified**. Then \( \hat{\xi} \) converges in probability to the true value of \( \xi \), and thus

\[
m(\nu; \xi^*) = E(Y|V).
\]

In this case, \( E \{ Y - m(\nu; \xi^*)|V \} = E \{ Y - E(Y|V)|V \} = 0 \), 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 \((C_i, CY_i, X_i, V_i), i = 1, \ldots, N\).

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

\[
pr(C = 1|Y, X, V) = pr(C = 1|X, V) = \pi(X, V), \tag{5.17}
\]

say. That is, we are unable to assume that

\[
pr(C = 1|Y, X) = pr(C = 1|X),
\]

which would have allowed us to use the usual, **complete case** estimator for \( \beta \) 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 $\beta$ is the solution to the estimating equation

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

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),

$$E_{\beta} \left[ C_{i} \frac{\partial}{\partial \beta} \{\mu(X; \beta)\} \{Y - \mu(X; \beta)\} \right] = E_{\beta} \left( E_{\beta} \left[ C_{i} \frac{\partial}{\partial \beta} \{\mu(X; \beta)\} \{Y - \mu(X; \beta)\} \middle| Y, X, V \right] \right) = E_{\beta} \left[ \pi(X, V) \frac{\partial}{\partial \beta} \{\mu(X; \beta)\} \{Y - \mu(X; \beta)\} \right], \quad (5.19)$$

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

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

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

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

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

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), \( \psi \) can be estimated by the MLE \( \hat{\psi} \) maximizing
\[
N \prod_{i=1}^{N} \left\{ \pi(X_i, V_i; \psi) \{1 - \pi(X_i, V_i; \psi)\} \right\}^{C_i} \implies \prod_{i=1}^{N} \left\{ \pi(X_i, V_i; \psi) \{1 - \pi(X_i, V_i; \psi)\} \right\}^{1-C_i}, \tag{5.21}
\]
using the data \((C_i, X_i, V_i)\), \(i = 1, \ldots, N\). The fitted \( \pi(X_i, V_i; \hat{\psi}) \) can then be substituted in (5.20), and \( \beta \) can be estimated by solving
\[
\sum_{i=1}^{N} C_i \pi(X_i, V_i; \hat{\psi}) \frac{\partial}{\partial \beta} \{ \mu(X_i; \beta) \} \{ Y_i - \mu(X_i; \beta) \} = 0.
\]

A doubly robust, augmented inverse probability weighted complete case estimator for \( \beta \) can also be derived by considering the estimating equation
\[
\sum_{i=1}^{N} \left( \frac{C_i}{\pi(X_i, V_i; \hat{\psi})} \frac{\partial}{\partial \beta} \{ \mu(X_i; \beta) \} \{ Y_i - \mu(X_i; \beta) \} - \left\{ \frac{C_i - \pi(X_i, V_i; \hat{\psi})}{\pi(X_i, V_i; \hat{\psi})} \right\} \frac{\partial}{\partial \beta} \{ \mu(X_i; \beta) \} \{ Y_i - \mu(X_i; \beta) \} \right) = 0.
\]
This equation is equal to
\[
\sum_{i=1}^{N} \left[ \frac{C_i}{\pi(X_i, V_i; \hat{\psi})} \frac{\partial}{\partial \beta} \{ \mu(X_i; \beta) \} \{ Y_i - \mu(X_i; \beta) \} - \left\{ \frac{C_i - \pi(X_i, V_i; \hat{\psi})}{\pi(X_i, V_i; \hat{\psi})} \right\} \frac{\partial}{\partial \beta} \{ \mu(X_i; \beta) \} \{ Y_i - \mu(X_i; \beta) \} \right] = 0. \tag{5.22}
\]
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 \( \hat{\xi} \) for \( \xi \) can be obtained by solving the OLS estimating equation
\[
\sum_{i=1}^{N} C_i \frac{\partial}{\partial \xi} \{ m(X_i, V_i; \xi) \} \{ Y_i - m(X_i, V_i; \xi) \} = 0 \tag{5.23}
\]
using the data on the complete cases.
Substituting \( m(X_i, V_i; \hat{\xi}) \) 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 \( \beta \), namely

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

**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,

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

and

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

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

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

We can then estimate \( \beta \) and \( \xi \) **simultaneously** by solving jointly the estimating equations

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

and

\[
\sum_{i=1}^{N} C_i \left( \begin{array}{c} 1 \\ X_i \\ V_i \end{array} \right) \{ Y_i - \mu(X_i; \beta) - \xi_0 + \xi_1^T X + \xi_2^T V \} = 0.
\]
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

\[
\begin{pmatrix}
Y \\
V
\end{pmatrix} = \left( \begin{array}{c}
\mu(X; \beta) \\
E(V|X)
\end{array} \right) + \Sigma^{-1/2} W.
\]

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

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

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 \( t_j, j = 1, \ldots, T \). The full data are then

\[ Z = \{(Y_1, X_1, V_1), \ldots, (Y_T, X_T, V_T)\}. \]

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

\[ \overline{X}_T = \{X_1, \ldots, X_T\}, \]

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

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

We consider a model of the form

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

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

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 \( \beta \) 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 \( \bar{X}_T \) over all \( T \) time points is taken to potentially depend on all of these covariates for each \( j = 1, \ldots, T \). 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 \( \mu_j(\bar{x}_T; \beta) \) to depend only on the covariates available through time \( t_j \). That is, the model for \( E(Y_j|\bar{X}_T = \bar{x}_T) \) depends on \( \bar{x}_T \) only through \( x_1, \ldots, x_j \).

In many practical situations, the covariates collected over \( t_1, \ldots, t_T \) 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

\[
E(Y|X = x) = \mu(x; \beta) = \left( \begin{array}{c} \mu_1(x; \beta) \\ \vdots \\ \mu_T(x; \beta) \end{array} \right). \tag{5.26}
\]

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 \( \beta \) 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

\[ Z = \{(Y_1, V_1), \ldots, (Y_T, V_T), X\}. \]  

(5.27)

Models of interest are of the form in (5.26), and the goal is to estimate \( \beta \). 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 \( Z_i, i = 1, \ldots, N \), is available, then it is well known that the optimal GEE to be solved to estimate \( \beta \) is given by

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

where

\[ D(x; \beta) = \frac{\partial}{\partial \beta} \{ \mu(x; \beta) \} \]

is the \((T \times p)\) matrix of partial derivatives of the \( T \) elements of \( \mu(x; \beta) \) with respect to the \( p \) components of \( \beta \); and \( V(x; \beta) \) 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**

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

where \( A_j(x), j = 1, \ldots, T, \) is the \((p \times 1)\) vector such that the \((p \times T)\) matrix with columns \( A_j(x) \)

\[ \{ A_1(x), \ldots, A_T(x) \} = D^T(X; \beta) V^{-1}(X; \beta). \]

As discussed shortly, among the **class** of estimating functions having the form of the rightmost expression in (5.29), this choice of the \( A_j(x) \) is **optimal**; other choices of the \( A_j(x) \) 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**.
**CHAPTER 5  ST 790, MISSING DATA**

**DROPOUT**: The foregoing development demonstrates how inference proceeds when a sample of full data is available. We now consider inference on $\beta$ 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,

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

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

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

and

$$\text{pr}(D = j + 1 \mid Z) = \pi_j(Z)\lambda_{j+1}(Z), \quad j = 1, \ldots, T. \tag{5.32}$$

Note that, because all individuals are observed at baseline, $\pi_1(Z) = \text{pr}(R_1 = 1 \mid Z) = \text{pr}(R_1 = 1) = 1$.

**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 $t_j$. 

---

126
MAR implies that the cause-specific hazard of dropout (5.30) can be written as
\[
\lambda_j(Z) = \Pr(D = j | D \geq j, Z) = \Pr(D = j | D \geq j, H_{j-1}) = \lambda_j(H_{j-1}), \quad j = 2, \ldots, T; \tag{5.33}
\]
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 \textit{observed history} through time \( t_{j-1} \). Likewise, (5.31) and (5.32) become
\[
\pi_j(Z) = \pi_j(H_{j-1}) = \Pr(R_j = 1 | H_{j-1}) = \prod_{\ell=1}^{j} \{1 - \lambda_\ell(H_{\ell-1})\}, \quad j = 2, \ldots, T, \tag{5.34}
\]
and
\[
\Pr(D = j + 1 | Z) = \Pr(D = j + 1 | H_{j}) = \pi_j(H_{j-1}) \lambda_{j+1}(H_{j}), \quad j = 1, \ldots, T, \tag{5.35}
\]
By convention in formulæ to follow, when \( j = 1 \), \( \pi_j(H_{j-1}) = \pi_1 = 1 \). We will use this and (5.33), (5.34), and (5.35) in the sequel.

\textbf{AUXILIARY VARIABLES:} As in Sections 5.1 and 5.2, although the auxiliary variables \( V_j, j = 1, \ldots, T \), 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 \textit{plausible}.

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

(i) \textit{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 \textit{exogenous covariates}, as in (5.26).

(ii) \textit{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 \( x_T \) only through \( x_1, \ldots, x_j \).

(iii) \textit{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

\[
\sum_{j=1}^{T} \left\{ \frac{I(D = j + 1)}{\pi_j(H_j - 1) \lambda_{j+1}(H_j)} \right\} \left[ G_j(X) \{ Y_1 - \mu_1(X; \beta) \} + \cdots + G_j(X) \{ Y_J - \mu_J(X; \beta) \} \right], \tag{5.36}
\]

where \(G_j(x), \ell = 1, \ldots, j, j = 1, \ldots, T\), are arbitrary \((p \times 1)\) functions of \(x\). Thus, note that the estimating function at the \(j\)th 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 \(j = 1, \ldots, T\), his/her contribution to the WGEE is

\[
\left\{ \frac{I(D_i = j + 1)}{\pi_j(H_i - 1) \lambda_{j+1}(H_i)} \right\} \left[ G_j(X_i) \{ Y_{i1} - \mu_{1}(X_i; \beta) \} + \cdots + G_j(X_i) \{ Y_{ij} - \mu_{j}(X_i; \beta) \} \right]
\]

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

\[
\left\{ \frac{I(D_i = j + 1)}{\pi_j(H_i - 1) \lambda_{j+1}(H_i)} \right\}
\]

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

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

\[
\{ G_{i1}(X), \ldots, G_{ij}(X) \} = D_j^T (X; \beta) \nu_j^{-1}(X; \beta), \tag{5.37}
\]

where, as in Section 1.4, \(D_j^T (X; \beta) \) \((p \times j)\) and \(\nu_j(X; \beta) \) \((j \times j)\) are the corresponding submatrices of \(D^T (X; \beta) \) \((p \times T)\) and \(\nu(X; \beta) \) \((T \times T)\). 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))

\[
\sum_{j=1}^{T} \left\{ \sum_{i=1}^{N} w_{ij} D_j^T (X_i; \beta) \nu_j^{-1}(X_i; \beta) \begin{pmatrix} Y_{i1} - \mu_{1}(X_i; \beta) \\ \vdots \\ Y_{ij} - \mu_{j}(X_i; \beta) \end{pmatrix} \right\} = 0, \quad w_{ij} = \frac{I(D_i = j + 1)}{\pi_j(H_i - 1) \lambda_{j+1}(i, H_i)}.
\]
**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 \( R_{ij+1} = 0 \) 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

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

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

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

where \( W_i \) is the \((T \times T)\) diagonal *weight matrix* with diagonal elements

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

**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 does not seem to be a R package implementing these approaches at present.
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 $\beta$ in the (semiparametric) model (5.26) is

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

for arbitrary $(p \times 1) A_j(x)$, $j = 1, \ldots, T$. As noted in (5.29), the optimal choice of the $A_j(x)$, that leading to the estimator for $\beta$ with smallest asymptotic variance among all in this class, is such that the $(p \times T)$ matrix $\{A_1(x), \ldots, A_T(x)\} = D^T(x; \beta)V^{-1}(x; \beta)$.

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 $\beta$ based on the observed data. These estimating functions turn out to have the augmented inverse probability weighted complete case form

$$
\frac{R_T}{\pi_T(H_{T-1})} \sum_{j=1}^{T} A_j(X) \{ Y_j - \mu_j(X; \beta) \} + \sum_{j=1}^{T-1} \left\{ \frac{R_j}{\pi_j(H_{j-1})} - \frac{R_{j+1}}{\pi_{j+1}(H_j)} \right\} f_j(H_j), \tag{5.40}
$$

where $f_j(H_j)$, $j = 1, \ldots, T - 1$, are arbitrary $(p \times 1)$ functions of the histories $H_j$; and $A_j(X)$, $j = 1, \ldots, T$, 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

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

thus, (5.36) can be written as

$$
\frac{R_T}{\pi_T(H_{T-1})} \left[ G_{T1}(X) \{ Y_1 - \mu_1(X; \beta) \} + \cdots + G_{TT}(X) \{ Y_T - \mu_T(X; \beta) \} \right] + \sum_{j=1}^{T-1} \left[ \frac{R_j - R_{j+1}}{\pi_j(H_{j-1})\lambda_{j+1}(H_j)} \right] \left[ G_{1j}(X) \{ Y_1 - \mu_1(X; \beta) \} + \cdots + G_{jj}(X) \{ Y_j - \mu_j(X; \beta) \} \right], \tag{5.41}
$$

We now recursively relate (5.40) to (5.41).
Note first that
\[
R_1 f_1(H_1) = \frac{R_1}{\pi_1 \lambda_2(H_1)} G_{11}(X) \{ Y_1 - \mu_1(X; \beta) \},
\]
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
\[
\pi_j = \pi_j(H_{j-1}), \quad \lambda_j = \lambda_j(H_{j-1}),
\]
we then have
\[
R_2 \frac{f_2(H_2) - f_1(H_1)}{\pi_2} = \frac{R_2}{\pi_2} \left\{ \frac{g_{21} + g_{22}}{\lambda_3} \right\},
\]
and thus
\[
f_2(H_2) = \frac{g_{21} + g_{22}}{\lambda_3} + g_{11} - \frac{g_{11}}{\lambda_2} = \frac{g_{21} + g_{22}}{\lambda_3} + g_{11}.
\]

Next,
\[
R_3 \frac{f_3(H_3) - f_2(H_2)}{\pi_3} = \frac{R_3}{\pi_3} \left\{ \frac{g_{31} + g_{32} + g_{33}}{\lambda_4} - \frac{1 - \lambda_3}{\lambda_3} (g_{21} + g_{22}) \right\},
\]
so that, solving for \( f_3(H_3) \),
\[
f_3(H_3) = \frac{g_{31} + g_{32} + g_{33}}{\lambda_4} + (g_{11} + g_{21}) + g_{22}.
\]

Continuing with this recursion, we have for \( j = 1, \ldots, T - 1, \)
\[
f_j(H_j) = \frac{G_{j1}(X) \{ Y_1 - \mu_1(X; \beta) \} + \cdots + G_{j\ell}(X) \{ Y_\ell - \mu_\ell(X; \beta) \}}{\lambda_{j-1}(H_j)}
+ \{ G_{11}(X) + \cdots + G_{j-1,1}(X) \} \{ Y_1 - \mu_1(X; \beta) \}
+ \{ G_{22}(X) + \cdots + G_{j-1,2}(X) \} \{ Y_2 - \mu_2(X; \beta) \}
\vdots
+ G_{j-1,\ell-1}(X) \{ Y_\ell - \mu_\ell(X; \beta) \},
\]
and
\[
\mathcal{A}_1(X) = G_{11}(X) + \cdots + G_{T1}(X)
\]
\[
\mathcal{A}_2(X) = G_{22}(X) + \cdots + G_{T2}(X)
\]
\vdots
\[
\mathcal{A}_T(X) = G_{TT}(X).
\]
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

\[
f_1(H_1) = B_1(X)\{Y_1 - \mu_1(X; \beta)\}
\]

\[
\vdots
\]

\[
f_j(H_j) = B_1(X)\{Y_1 - \mu_1(X; \beta)\} + \cdots + B_j(X)\{Y_j - \mu_j(X; \beta)\},
\]

\(j = 2, \ldots, T - 1\), and \(A_j(X) = B_j(X), j = 1, \ldots, T\).

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

\[
\pi_j(H_{j-1}) = \prod_{i=1}^{j} \{1 - \lambda_i(H_{i-1})\}, \quad j = 2, \ldots, T,
\]

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 \(\psi\). From (5.35) and above, this implies models for \(\text{pr}(D = j | Z) = \text{pr}(D = j | H_{j-1})\) of the form

\[
\pi_{j-1}(H_{j-2}; \psi)\lambda_j(H_{j-1}; \psi) = \prod_{i=1}^{j-1} \{1 - \lambda_i(H_{i-1}; \psi)\}\lambda_j(H_{j-1}; \psi),
\]

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

\[
\prod_{i=1}^{N} \prod_{j=2}^{T+1} \left[ \prod_{i=1}^{j-1} \{1 - \lambda_i(H_{i-1}; \psi)\}\lambda_j(H_{j-1}; \psi) \right]^{\mathbb{I}(D_i = j)}. \quad (5.42)
\]

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

\[
\prod_{j=2}^{T} \prod_{l:R_{ij} = 1} \{\lambda_j(H_{ij-1}; \psi)\}^{\mathbb{I}(D_i = j)}\{1 - \lambda_j(H_{ij-1}; \psi)\}^{\mathbb{I}(D_i > j)},
\]

where in fact \(\mathbb{I}(D > j) = \mathbb{I}(R_j = 1)\).
Suppose we adopt a \textbf{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 $\hat{\psi}$ for $\psi$ maximizing the above likelihood is the solution to the \textbf{estimating equation}
\[
\sum_{i=1}^{N} \sum_{j=2}^{T} R_{ij-1} \frac{\partial}{\partial \psi}\{\alpha_j(H_{ij-1}; \psi)\}\{I(D_i = j) - \lambda_j(H_{ij-1}; \psi)\} = 0. \tag{5.43}
\]

In practice, it would be unusual for an analyst to posit models $\lambda_j(H_{j-1}; \psi)$ that share a \textbf{common parameter} $\psi$ across different $j$. Rather, a standard approach is to take the model for each occasion $j$ to have a separate parameter $\psi_j$, say, so that $\psi = (\psi_2^T, \ldots, \psi_T^T)^T$, and the $\psi_j$ are \textbf{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 $\psi$ boils down to solving
\[
\sum_{i=1}^{N} R_{ij-1} \frac{\partial}{\partial \psi_j}\{\alpha_j(H_{ij-1}; \psi_j)\}\{I(D_i = j) - \lambda_j(H_{ij-1}; \psi_j)\} = 0
\]
\textit{separately} for $j = 2, \ldots, T$. That is, to estimate $\psi_j$, 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 \textbf{linear} in $\psi_j$, software such as SAS \texttt{proc logistic} or any generalized linear model software can be used.

\textbf{COMPARISON OF SUBJECT AND OCCASION LEVEL WEIGHTING:}

- The subject level approach has greater \textbf{practical appeal} because it is easier to implement using \textbf{standard software} for solving GEEs. Many such programs, such as SAS \texttt{proc genmod}, allow the user to specify a fixed \textbf{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 \textbf{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,

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

From the theory of semiparametrics for missing data problems, it can be shown (Tsiatis, 2006) that, for fixed \( \{ A_1(x), \ldots, A_T(x) \} \), the optimal choice of \( f_j(H_j) \) in (5.44) is

$$f_j(H_j) = E \left[ \sum_{\ell=1}^{T} A_\ell(X) \{ Y_\ell - \mu_\ell(X; \beta) \} | H_j \right],$$

$$= \sum_{\ell=1}^{j} A_\ell(X) \{ Y_\ell - \mu_\ell(X; \beta) \} + \sum_{\ell=j+1}^{T} A_\ell(X) E[\{ Y_\ell - \mu_\ell(X; \beta) \} | H_j], \quad (5.45)$$

for \( j = 1, \ldots, T - 1 \). That is, for fixed \( A_j(x) \), \( j = 1, \ldots, T \), using this choice of \( f_j(H_j) \), which depends on the particular fixed \( A_j(x) \), will yield the estimating function of form (5.44) with this fixed \( A_j(x) \) leading to the estimator for \( \beta \) 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
  $$f_j(H_j) = \sum_{\ell=1}^{j} A_\ell(X) \{ Y_\ell - \mu_\ell(X; \beta) \}.$$

  This leads to the occasion level WGEE in (5.38) with \( A_j(x) = B_j(x) \), \( j = 1, \ldots, T \).
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 \(\beta\) if EITHER \(\pi_j(H_{j-1})\) (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

\[
\{A_1(x), \ldots, A_T(x)\} = D_T(x; \beta) V^{-1}(x; \beta).
\]

As we demonstrated in the last section, for the occasion level WGEE, \(A_j(x) = B_j(x), j = 1, \ldots, T\), 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 \(A_j(x)\) 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 \(\{A_1(x), \ldots, A_T(x)\}\); 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 when 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 \(\{A_1(x), \ldots, A_T(x)\}\), from (5.45), the optimal choice for \(f_j(H_j)\) for construction of a doubly robust, augmented inverse probability weighted estimating function for \(\beta\) 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,
\[ p_{\epsilon_{j+1}, V_{j+1}|H_j}(\epsilon_{j+1}, V_{j+1}|h_j) = p_{\epsilon_{j+1}, V_{j+1}|H_j}(\epsilon_{j+1}, V_{j+1}|h_j, 1); \tag{5.46} \]
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
\[ H_j = (1, X^T, \epsilon_1, \epsilon_1^T, \ldots, \epsilon_j, V_j^T)^T, \quad j = 1, \ldots, T - 1. \]

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,
\[ E \left( \begin{array}{c} \epsilon_{j+1} \\ V_{j+1} \end{array} \right | H_j) = \left( \begin{array}{c} \Lambda_j \\ \Gamma_j \end{array} \right ) H_j, \]
where \(\Lambda_j (1 \times q_j)\) and \(\Gamma_j (r_j \times q_j)\) 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
\[ E(H_{j+1}|H_j) = \left( \begin{array}{c} I_{q_j} \\ \Lambda_j \\ \Gamma_j \end{array} \right ) H_j = \Delta_j H_j, \]
say. Here \(\Delta_j\) 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 \( \Delta_j \) 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 \( \epsilon_{j+1} \) and each element of \( V_{j+1} \) on those of \( H_j \). Note that

\[ H_T = (1, X^T, \epsilon_1^T, V_1^T, \ldots, \epsilon_T^T, V_T^T)^T; \]

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 \( C = (\epsilon_1, \ldots, \epsilon_T, V_1^T, \ldots, V_T^T)^T \) 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 \( \text{var}(\epsilon_1, \ldots, \epsilon_T | X) \).

**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),

\[
\frac{R_T}{\pi_T(H_{T-1})} \sum_{j=1}^{T} A_j(X) \{ Y_j - \mu_j(X; \beta) \} + \sum_{j=1}^{T-1} \left\{ \frac{R_j}{\pi_j(H_{j-1})} - \frac{R_{j+1}}{\pi_{j+1}(H_j)} \right\} f_j(H_j). \tag{5.48}
\]

We have discussed the optimal choice of the \( f_j(H_j) \) when the \( A_j(x) \) 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 \( A_j(x) \) is used.
In Tsiatis (2006), it is shown that the **optimal** \( \{A_1(x), ..., A_T(x)\} \) is given by

\[
\{A_1(x), ..., A_T(x)\} = D^T(x; \beta) \{V^*(x)\}^{-1},
\]

where

\[
V^*(X) = E \left\{ \frac{\epsilon \epsilon^T}{\pi_T(H_{T-1})} \left| X \right\} - \sum_{j=1}^{T-1} E \left\{ \frac{\lambda_{j+1}(H_j)}{\pi_{j+1}(H_j)} E(\epsilon|H_j) \epsilon^T \left| X \right\} \right\}, \tag{5.49}
\]

which can be written equivalently as

\[
V^*(X) = E \left\{ \frac{\epsilon \epsilon^T}{\pi_T(H_{T-1})} \left| X \right\} - \sum_{j=1}^{T-1} \left[ E \left\{ \frac{E(\epsilon|H_j)}{\pi_j(H_{T-1})} \epsilon^T \left| X \right\} \right\} - E \left\{ \frac{E(\epsilon|H_j)}{\pi_{j+1}(H_j)} \epsilon^T \left| X \right\} \right\}.
\]

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 \((\epsilon^{(s)}^T, V_1^{(s)}^T, ..., V_T^{(s)}^T)^T, s = 1, ..., S, \) for large \( S, \) from a multivariate normal model for \((\epsilon^T, V_1^T, ..., V_T^T)^T \) given \(X,\) we can estimate \(V^*(x)\) as

\[
S^{-1} \sum_{s=1}^{S} \left\{ \frac{\epsilon^{(s)}_s \epsilon^{(s)}_s^T}{\pi_T(H_{T-1}^{(s)})} - \sum_{j=1}^{T-1} \frac{\lambda_{j+1}(H_j^{(s)})}{\pi_{j+1}(H_j^{(s)})} E(\epsilon|H_j^{(s)}) \epsilon^{(s)}_s^T \right\}.
\]

Whether or not this is a feasible strategy, and whether or not going to all of this trouble will yield an estimator for \(\beta\) 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.