Chapter 5

Preliminaries on Semiparametric Theory and Missing Data Problem

5.1 Introduction

The goal of this chapter is to introduce some vocabulary as well as technical results on semiparametric theory and their application to the problem of estimating the conditional expectation model in presence of missing data. Specifically, we will refer to the case that the missingness is in the covariates vector and the non-response is non-ignorable (Little and Rubin, 1987). This chapter is mainly devoted to the exposition of tools and concepts that will be necessary in the development of the semiparametric approach in Chapter 6.

Most of the ideas that we develop correspond to the recent work done by Professor James Robins and Professor Andrea Rotnitzky, from the Harvard School of Public Health. In particular, the main references are Rotnitzky (1996), Rotnitzky and Robins (1997) and Rotnitzky, Robins and Scharfstein (1998).

The chapter is organized as follows. In Section 5.2 we present a review of the literature about semiparametric modeling where we go from the complete case analysis to the non-ignorable non-response pattern model. Section 5.3 is devoted to basic definitions and properties. In Section 5.4 we introduce the generalized method of moments (henceforth GMM) as a general framework for the GEE estimators. In Section 5.5 we define the inverse probability of being observed weighted generalized estimating equations (henceforth IPWGEE) and we discuss their efficiency properties in Section 5.6. Finally, in Section 5.7, we introduce how to perform a sensitivity analysis of the inferences that we get for different values of parameters measuring the non-ignorability of the non-response probabilities.

5.2 State of the art

For the complete case analysis Liang and Zeger (1986) introduced an extension of the generalized linear models (McCullagh and Nelder, 1983) to the analysis of longitudinal data. They proposed a class of generalized estimating equations (henceforth GEE) whose solutions are consistent estimates of the regression parameters. For the variance of the estimates they proposed what is known as sandwich estimator. Both estimators are $\sqrt{n}$-consistent only if the mean regression model is correctly specified, no matter what is the choice of the "working" correlation matrix. These estimating equations are an extension of quasi-likelihood methods to the multivariate regression setting and they reduce to the score equations for multivariate Gaussian outcomes. Their approach results in iteratively reweighted least squares estimators of the regression coefficients. In presence of missing data, Liang and Zeger remarked that inferences using GEE are valid only under the stronger assumption that the missingness is completely at random (Rubin, 1976).

When the non-response pattern is missing at random (MAR) and only the outcome, Y, can be not observed, Robins, Rotnitzky and Zhao (1995) and Rotnitzky and Robins (1995a,b) proposed a semiparametric estimation procedure for estimating the regression of the outcome, measured at the end of a fixed follow-up period, on baseline covariates, $\boldsymbol{X}$, measured before the beginning of the follow-up. This conditional mean model is a semiparametric model in the sense that the distribution of the regressors and the conditional distribution of the residuals are left completely unspecified. The methodology is based on the IPWGEE. It is important to note that when they refer to censored individuals they are referring to subjects that drop out of the study prior to the end of follow-up (*i.e.,* Y is missing), in a different sense from the corresponding to survival analysis. If other completely observed covariates, $\boldsymbol{V}$, are available, then, they can be used as a surrogate variables in order to gain efficiency. The methodology is also developed for longitudinal data and is presented under a monotone missing data pattern. However, some extensions to arbitrary missing data patterns are also studied. In particular, for a longitudinal study Robins and Rotnitzky (1995) show that, when some dropouts do not later return to the study, there is no more information in the observed data about the regression coefficients than the existing in the monotone part of the data. Software tools developing these procedures, in terms of marginal regression, have been implemented by Kastner, Fieger and Heumann (1997).

More research has been done under the MAR assumption but with possibly missing covariates. Robins, Rotnitzky and Zhao (1994) proposed the IPWGEE methodology for estimating the parameter in a conditional mean model when the data are MAR and the missingness probabilities are either known or can be parametrically modeled. They showed that this estimation problem is a special case of the general problem of parameter estimation in an arbitrary semiparametric model. Because the optimal estimator depends on the unknown probability law generating the data, they proposed locally and globally adaptive semiparametric efficient estimators (Bickel et al., 1993). Nielsen (1997) showed that a strengthened version of the MAR assumption is sufficient to yield ordinary large sample results and (Nielsen, 1998) he extended the semiparametric theory to the coarsening at random case (Heitjan and Rubin, 1991; Heitjan, 1993). Recently, Lipsitz, Ibrahim and Zhao (1999) proposed weighted estimating equations (Robins et al., 1994; Zhao et al., 1996) and an EM-type algorithm (Dempster et al., 1977) to solve them with properties similar to a maximum likelihood approach.

Concerning the application of the above ideas to survival data, to the best of our knowledge, only the MAR assumption on missing covariates and the Cox proportional hazards model (Cox, 1972) have been considered (it is important to note that neither the non-response pattern nor the Cox model can be validated). As an example of semiparametric model Robins and Rotnitzky (1992), Robins (1993) and Robins, Rotnitzky and Zhao (1994) used the IPWGEE methodology to estimate the survival function under the Cox model, they described the influence functions of the resulting estimators and its efficiency in the sense that its asymptotic variance attains the semiparametric variance bound of the model.

In this same situation (MAR covariates and Cox modeling), other techniques can be applied. Paik and Tsai (1997) suggested to impute the conditional expectation of any statistic in the partial likelihood equations involving missing covariates given the available information. They proved that the proposed estimator is more efficient than the based on the modified Cox partial likelihood score equations proposed by Lin and Ying (1993) for the MCAR case. When the missing data pattern is monotone, Cox score estimating equations such, as those proposed by Pugh (1993) and Reilly and Pepe (1995), are useful. Lipsitz and Ibrahim (1998) extend this methodology to the non-monotone case but only for categorical covariates.

When the missing mechanism is non-ignorable, and in the context of estimating the conditional mean, Rotnitzky and Robins (1997) extended the IPWGEE class of estimators to allow the non-response to depend on partially non-observed variables. The methodology provides consistent and asymptotically normal estimates and it works for missing outcome as well as for missing covariates. The proposed estimators do not require full specification of a parametric likelihood but the non-response probabilities have to be parametrically modeled. They showed that the asymptotic variance of the optimal estimator in the class attains the semiparametric variance bound for the model. Rotnitzky, Robins and Scharfstein (1998) generalized these results to the repeated outcomes subject to non-ignorable non-response case. A general presentation of the semiparametric modeling under semiparametric non-ignorable dropouts with interesting comments by Freedman, Fan and Zhang, van der Laan, Diggle, Little and Rubin, and Laird and Pauler can be found in Scharfstein, Rotnitzky and Robins (1999). However, a semiparametric approach to survival data with non-ignorable missing covariates has not yet been considered in the literature.

5.3 From a parametric to a semiparametric point of view

As we illustrated in Chapter 4, an important inconvenience of the likelihood-based methods is that they are not robust to wrong specifications of the distribution functions (neither of the scientific interest part of the data nor of the non-interest part one). As discussed in Laird (1988) inferences about the regression coefficients can be very sensitive to misspecification of the parametric model for the join law for residuals and covariates. An interesting example pointing out the dependence of the inferences on the departure assumptions can be found in Rotnitzky and Robins (1997).

The main idea in the semiparametric estimators is that they will be consistent if the scientific interest part of the model is correctly specified but it will not depend on the correct specification of other parts of the distribution generating the data. Some of the main references about semiparametric modeling theory are Begun, Hall, Huang and Wellner (1983), Bickel, Klaasen, Ritov and Wellner (1993), Newey (1990) and Newey and McFadden (1994).

Definition 5.3.1 .

a) *A semiparametric model is a class of density functions (with respect to the same measure)*

$$\mathcal{P} = \{f(\boldsymbol{L}; \boldsymbol{\beta}, \boldsymbol{\eta}); \boldsymbol{\beta} \in R^p; \boldsymbol{\eta} \textit{ infinite dimensional}\}$$

containing the true density function, $f(\boldsymbol{L}; \boldsymbol{\beta}^*, \boldsymbol{\eta}^*)$*, generating the data.*

b) *A parametric submodel* $\mathcal{P}_{\boldsymbol{\phi}}$ *in* $\mathcal{P}$ *is a class*

$$\mathcal{P}_{\boldsymbol{\phi}} = \{f(\boldsymbol{L}; \boldsymbol{\beta}, \boldsymbol{\phi}); \boldsymbol{\beta} \in R^p; \boldsymbol{\phi} \in R^q\} \subset \mathcal{P}$$

containing the true density $f(\boldsymbol{L}; \boldsymbol{\beta}^*, \boldsymbol{\eta}^*)$*.*

Indeed, the density function $f(\boldsymbol{L}; \boldsymbol{\beta}, \boldsymbol{\eta})$ has a parametric part $\boldsymbol{\beta}$ (finite dimensional) and a non-parametric one $\boldsymbol{\eta}$ (infinite dimensional). For example, the conditional expectation model (GEE) is a semiparametric model in the sense that if the data are $(\boldsymbol{Y}_i, \boldsymbol{X}_i)$, $i = 1, \dots, n$, we only specify the relation $E(\boldsymbol{Y}_i|\boldsymbol{X}_i) = g(\boldsymbol{X}_i; \boldsymbol{\beta}^*)$. There are no assumptions neither on the distribution $f(\epsilon_i|\boldsymbol{X}_i; \boldsymbol{\eta}_1)$ of the residuals $\epsilon_i = E(\boldsymbol{Y}_i|\boldsymbol{X}_i) - g(\boldsymbol{X}_i; \boldsymbol{\beta}^*)$ nor on the marginal distribution of the covariates $f(\boldsymbol{X}_i; \boldsymbol{\eta}_2)$. In this example $\boldsymbol{\eta} = (\boldsymbol{\eta}_1', \boldsymbol{\eta}_2')'$ will be the indices of the functions $f(\epsilon_i|\boldsymbol{X}_i)$ and $f(\boldsymbol{X}_i)$, and the semiparametric model will be

$$\mathcal{P}_{GEE} = \left\{ f(y|x; \boldsymbol{\beta}, \boldsymbol{\eta}) : \int y f(y|x; \boldsymbol{\beta}, \boldsymbol{\eta}) dy = g(x; \boldsymbol{\beta}) \right\}.$$

Definition 5.3.2 *A estimator $\widehat{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}^*$ is semiparametric in $\mathcal{P}$, if for all distribution in $\mathcal{P}$, $\sqrt{n}\left(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}^*\right)$ converges in distribution ($\xrightarrow{\mathcal{D}}$) to a multivariate random variable distributed as $\mathcal{N}(\mathbf{0},\boldsymbol{\Sigma})$, that is,*

$$\sqrt{n}\left(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}^*\right)\xrightarrow{\mathcal{D}}\mathcal{N}(\mathbf{0},\boldsymbol{\Sigma}).$$

In other words, if $\widehat{\boldsymbol{\beta}}$ is $\sqrt{n}$-consistent and asymptotically normal no matter the distribution in $\mathcal{P}$ generating the data.

Definition 5.3.3 .

a) A parametric model $\mathcal{P}_{\boldsymbol{\phi}}$ is regular if the information matrix is well defined.

b) A semiparametric model $\mathcal{P}$ is regular if every parametric submodel in $\mathcal{P}$ is regular.

Definition 5.3.4 *A parametric estimator $\widehat{\boldsymbol{\beta}_n}$ of $\boldsymbol{\beta}$ is regular if $\widehat{\boldsymbol{\beta}_n}$ locally uniformly converges to $\boldsymbol{\beta}$. That is, for every sequence $(\boldsymbol{\beta}_n,\boldsymbol{\phi}_n)$ in a closed neighbourhood of $(\boldsymbol{\beta}^*,\boldsymbol{\phi}^*)$, $P_{(\boldsymbol{\beta}_n,\boldsymbol{\phi}_n)}\left\{\sqrt{n}\left(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}^*\right)\leq z\right\}$ converges to $P_{(\boldsymbol{\beta}^*,\boldsymbol{\phi}^*)}\left\{\sqrt{n}\left(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}^*\right)\leq z\right\}$, for each $z\in R^p$.*

In fact, regular estimators are those that allow us to build asymptotic confidence intervals. In order to define the regularity of a semiparametric estimator, we discuss first the concept of *efficiency* in the parametric modeling framework.

As it is well known, in a parametric model the maximum likelihood estimator ($\widehat{\boldsymbol{\beta}}_{MLE}$) is *efficient* in the sense that its asymptotic variance is the minimum of the asymptotic variances of any regular estimator. In particular, the asymptotic variance-covariance matrix of the MLE is the inverse of the information matrix. That is, if the parametric model has parameters $\boldsymbol{\beta}$ and $\boldsymbol{\phi}$, then the information matrix is

$$I=E\begin{pmatrix} S_{\boldsymbol{\beta}}S'_{\boldsymbol{\beta}} & S_{\boldsymbol{\beta}}S'_{\boldsymbol{\phi}} \\ S_{\boldsymbol{\phi}}S'_{\boldsymbol{\beta}} & S_{\boldsymbol{\phi}}S'_{\boldsymbol{\phi}} \end{pmatrix}$$

where

$$S_{\boldsymbol{\beta}}=S_{\boldsymbol{\beta}}(\boldsymbol{L};\boldsymbol{\beta}^*,\boldsymbol{\phi}^*)=\{\partial\log f(\boldsymbol{L};\boldsymbol{\beta},\boldsymbol{\phi})/\partial\boldsymbol{\beta}\}\,|_{(\boldsymbol{\beta}^*,\boldsymbol{\phi}^*)}$$

and

$$S_{\boldsymbol{\phi}} = S_{\boldsymbol{\phi}}(\boldsymbol{L}; \boldsymbol{\beta}^*, \boldsymbol{\phi}^*) = \{\partial \log f(\boldsymbol{L}; \boldsymbol{\beta}, \boldsymbol{\phi})/\partial \boldsymbol{\phi}\} \mid_{(\boldsymbol{\beta}^*, \boldsymbol{\phi}^*)},$$

and the asymptotic variance of the MLE for $\boldsymbol{\beta}$, $\mathrm{Var}^A_{\boldsymbol{\phi}}(\widehat{\boldsymbol{\beta}}_{MLE})$, is the upper-left squared matrix of $V = I^{-1}$

$$I^{-1} = V = \begin{pmatrix} V_{\boldsymbol{\beta}\boldsymbol{\beta}} & V_{\boldsymbol{\beta}\boldsymbol{\phi}} \\ V_{\boldsymbol{\phi}\boldsymbol{\beta}} & V_{\boldsymbol{\phi}\boldsymbol{\phi}} \end{pmatrix}$$

i.e., $V_{\boldsymbol{\beta}\boldsymbol{\beta}} = \left(E(S_{\boldsymbol{\beta}}S'_{\boldsymbol{\beta}}) - E(S_{\boldsymbol{\beta}}S'_{\boldsymbol{\phi}})(E(S_{\boldsymbol{\phi}}S'_{\boldsymbol{\phi}}))^{-1}E(S_{\boldsymbol{\phi}}S'_{\boldsymbol{\beta}})\right)^{-1}$. It is important to observe that $V_{\boldsymbol{\beta}\boldsymbol{\beta}}$ is the inverse of the residual variance of regressing $S_{\boldsymbol{\beta}}$ on $S_{\boldsymbol{\phi}}$; that is, $V_{\boldsymbol{\beta}\boldsymbol{\beta}}$ is the inverse of the variance of

$$S_{\boldsymbol{\beta},eff(\boldsymbol{\phi})} = S_{\boldsymbol{\beta}} - E(S_{\boldsymbol{\beta}}S'_{\boldsymbol{\phi}})(E(S_{\boldsymbol{\phi}}S'_{\boldsymbol{\phi}}))^{-1}S_{\boldsymbol{\phi}}. \tag{5.1}$$

Definition 5.3.5 *$S_{\boldsymbol{\beta},eff(\boldsymbol{\phi})}$ defined as in (5.1) is called the efficient score for $\boldsymbol{\beta}$ in the model $\mathcal{P}_{\boldsymbol{\phi}}$.*

Returning to the semiparametric modeling, we define

Definition 5.3.6 *A semiparametric estimator $\widehat{\boldsymbol{\beta}}$ is regular in $\mathcal{P}$ if the estimator is regular in any parametric submodel $\mathcal{P}_{\boldsymbol{\phi}}$.*

It means that the set of regular semiparametric estimators for a semiparametric model $\mathcal{P}$ is a subset of the regular parametric estimators for every parametric submodel $\mathcal{P}_{\boldsymbol{\phi}} \subset \mathcal{P}$. Then, for every parametric submodel $\mathcal{P}_{\boldsymbol{\phi}} \subset \mathcal{P}$,

$$\mathrm{Var}^A(\widehat{\boldsymbol{\beta}}) \geq \mathrm{Var}^A_{\boldsymbol{\phi}}(\widehat{\boldsymbol{\beta}}_{MLE})$$

i.e.,

$$\mathrm{Var}^A(\widehat{\boldsymbol{\beta}}) \geq C = \sup_{\mathcal{P}_{\boldsymbol{\phi}}} \left\{ \left[\mathrm{Var}\left(S_{\boldsymbol{\beta},eff(\boldsymbol{\phi})}\right)\right]^{-1} \right\}. \tag{5.2}$$

Definition 5.3.7 *The constant C in (5.2) is defined as the semiparametric variance bound for $\boldsymbol{\beta}$ and C^{-1} as the semiparametric information matrix of $\boldsymbol{\beta}$.*

We provide now a geometric description of the parametric and semiparametric information matrices. In a parametric model $\mathcal{P}_{\boldsymbol{\phi}}$, the amount of information about $\boldsymbol{\beta}$ in $\mathcal{P}_{\boldsymbol{\phi}}$ is measured by $\mathrm{Var}(S_{\boldsymbol{\beta},eff(\boldsymbol{\phi})})$.

In the Hilbert's space integrated by all p-dimensional random vectors with mean $\mathbf{0}$ and finite variance with the usual covariance inner product (*i.e.,* with the variance norm) we consider the following definition.

Definition 5.3.8 *In a parametric model $\mathcal{P}_{\boldsymbol{\phi}}$, denote by $\Lambda_{\boldsymbol{\phi}}$ the nuisance tangent space for $\boldsymbol{\phi}$ defined as*

$$\Lambda_{\boldsymbol{\phi}} = \{KS_{\boldsymbol{\phi}} : K \textit{ is a } p \times q \textit{ constant matrix}\},$$

i.e., the vector subspace of linear combinations of $S_{\boldsymbol{\phi}}$.

If, for a closed subspace Φ, we denote by $\prod(.|\Phi)$ the orthogonal projection onto Φ operator and by $\Phi^{\perp}$ the orthogonal subspace of Φ, then we can derive the following geometric representation.

Lemma 5.3.1 *In a parametric model $\mathcal{P}_{\boldsymbol{\phi}}$,*

$$S_{\boldsymbol{\beta},eff(\boldsymbol{\phi})} = S_{\boldsymbol{\beta}} - \prod(S_{\boldsymbol{\beta}}|\Lambda_{\boldsymbol{\phi}}) = \prod(S_{\boldsymbol{\beta}}|\Lambda_{\boldsymbol{\phi}}^{\perp}). \tag{5.3}$$

Proof: By construction $\prod(S_{\boldsymbol{\beta}}|\Lambda_{\boldsymbol{\phi}})$ is the unique vector in $\Lambda_{\boldsymbol{\phi}}$ that verifies that $S_{\boldsymbol{\beta}} - \prod(S_{\boldsymbol{\beta}}|\Lambda_{\boldsymbol{\phi}})$ is orthogonal with all random vectors in $\Lambda_{\boldsymbol{\phi}}$. That is, $\prod(S_{\boldsymbol{\beta}}|\Lambda_{\boldsymbol{\phi}})$ corresponds to the choice of the K matrix that minimizes the norm of $S_{\boldsymbol{\beta}} - KS_{\boldsymbol{\phi}}$, *i.e.,*

$$E\left\{\left(S_{\boldsymbol{\beta}} - KS_{\boldsymbol{\phi}}\right)'\left(S_{\boldsymbol{\beta}} - KS_{\boldsymbol{\phi}}\right)\right\}. \tag{5.4}$$

But, by the Gauss-Markov's Theorem (Luenberger, 1969), (5.4) is minimized by $K = E(S_{\boldsymbol{\beta}}S'_{\boldsymbol{\phi}})(E(S_{\boldsymbol{\phi}}S'_{\boldsymbol{\phi}}))^{-1}$ then, according to the expression (5.1), (5.3) holds. □

This representation can be extended to semiparametric models.

Definition 5.3.9 *In a semiparametric model $\mathcal{P}$, we define the nuisance tangent space, denoted by Λ, as the closure of the set of all the $\Lambda_{\boldsymbol{\phi}}$ spaces for every $\mathcal{P}_{\boldsymbol{\phi}}$ parametric submodel in $\mathcal{P}$, that is,*

$$\Lambda = \overline{\bigcup_{\mathcal{P}_{\boldsymbol{\phi}} \subset \mathcal{P}} \Lambda_{\boldsymbol{\phi}}}.$$

Robins, Rotnitzky and Zhao (1994) proved that the semiparametric information about $\boldsymbol{\beta}$ in $\mathcal{P}$ is the variance of the projection of $S_{\boldsymbol{\beta}}$ onto the space $\Lambda^{\perp}$.

Theorem 5.3.1 *With the previous definitions*

$$C^{-1} = Var\left(\prod (S_{\boldsymbol{\beta}}|\Lambda^{\perp})\right). \tag{5.5}$$

This result allows to define the efficient semiparametric score as well.

Definition 5.3.10 *According to the definitions, we call efficient semiparametric score, $S_{\boldsymbol{\beta},eff}$, to the projection in (5.5), i.e.,*

$$S_{\boldsymbol{\beta},eff} = S_{\boldsymbol{\beta}} - \prod(S_{\boldsymbol{\beta}}|\Lambda) = \prod(S_{\boldsymbol{\beta}}|\Lambda^{\perp}).$$

Consequently, the semiparametric variance bound is the inverse of the variance of the efficient semiparametric score.

Definition 5.3.11 *Let $\widehat{\boldsymbol{\beta}}$ be a regular semiparametric estimator for $\boldsymbol{\beta}$ in a semiparametric model $\mathcal{P}$.*

a) $\widehat{\boldsymbol{\beta}}$ *is efficient if* $Var^{A}(\widehat{\boldsymbol{\beta}}) = C$.

b) If $\mathcal{P}'$ *denotes the semiparametric model* $\mathcal{P}$ *with an additional restriction, and* $C_{\mathcal{P}'}$ *and* $C_{\mathcal{P}}$, $(C_{\mathcal{P}'} \leq C_{\mathcal{P}})$, *the respective semiparametric variance bounds,* $\widehat{\boldsymbol{\beta}}$ *is locally efficient semiparametric estimator if* $Var^{A}(\widehat{\boldsymbol{\beta}}) = C_{\mathcal{P}'}$.

In practice, locally efficient semiparametric estimators are efficient if the assumed restriction is true. However, they still provide consistent and asymptotically normal estimates if the additional restriction is false.

In the missing data problem, since neither the distribution of interest nor the non-response probabilities can be validated, in order to derive locally efficient semiparametric estimators we will suppose a parametric model for the missing data pattern. The resulting estimators will be efficient if the parametric assumption is true. Otherwise, they will be consistent and asymptotically normal distributed. Sharfstein, Rotnitzky and Robins (1999) also study the missing data problem allowing a semiparametric model for the non-response mechanism.

5.4 GMM class of estimators

In this section we introduce the GMM class of estimators as a more general framework to deal with the GEE estimators. Many of the results introduced in this section can be found in Newey and McFadden (1994). Basically, we summarize properties concerning the identification, consistency and asymptotic normality of the resulting estimators.

Suppose that there is a "moment function" vector for the i-th observation and parameters, $\boldsymbol{U}_i(\boldsymbol{\theta})$, such that the population moments satisfy $E(\boldsymbol{U}_i(\boldsymbol{\theta}^*)) = 0$. A generalized method of moments estimator is one that minimizes a squared Euclidean distance of sample moments from their population counterpart of zero, in the sense that we specify here below. Let $\hat{\boldsymbol{W}}$ be a positive semi-definite matrix, so that $(\boldsymbol{m}'\hat{\boldsymbol{W}}\boldsymbol{m})^{1/2}$ is a measure of the distance of a vector $\boldsymbol{m}$ from $\boldsymbol{0}$. A GMM estimator, $\widehat{\boldsymbol{\theta}}$, is such that

$$\widehat{\boldsymbol{\theta}} \text{ maximizes } \widehat{Q_n}(\boldsymbol{\theta}) \text{ subject to } \boldsymbol{\theta} \in \boldsymbol{\Theta} \tag{5.6}$$

with

$$\widehat{Q_n}(\theta) = -\left[n^{-1}\sum_{i=1}^{n}\boldsymbol{U}_i(\boldsymbol{\theta})\right]' \hat{\boldsymbol{W}} \left[n^{-1}\sum_{i=1}^{n}\boldsymbol{U}_i(\boldsymbol{\theta})\right]. \tag{5.7}$$

The GMM class of estimators is large enough to include MLE and nonlinear least squares estimators as particular cases of extremum estimators; for instance by defining $\boldsymbol{U}_i(\boldsymbol{\theta})$ as the derivatives of the log-density or the derivatives of the minus least squared values. The GMM class itself is included in the class of minimum distance estimators, that is, the class of estimators that satisfy (5.6) for $\widehat{Q_n}(\boldsymbol{\theta}) =$

$-\widehat{\boldsymbol{U}_n}(\boldsymbol{\theta})'\hat{\boldsymbol{W}}\widehat{\boldsymbol{U}_n}(\boldsymbol{\theta})$ where $\widehat{\boldsymbol{U}_n}(\boldsymbol{\theta})$ is a vector of the data and parameters such that $\widehat{\boldsymbol{U}_n}(\boldsymbol{\theta}^*)$ converges in probability ($\xrightarrow{\mathcal{P}}$) to $\mathbf{0}$ and $\hat{\boldsymbol{W}}$ is positive semi-definite. More precisely, GMM corresponds to the case $\widehat{\boldsymbol{U}_n}(\boldsymbol{\theta}) = n^{-1}\sum_{i=1}^{n}\boldsymbol{U}_i(\boldsymbol{\theta})$.

By the law of large numbers, $n^{-1}\sum_{i=1}^{n}\boldsymbol{U}_i(\boldsymbol{\theta}) \xrightarrow{\mathcal{P}} E(\boldsymbol{U}_i(\boldsymbol{\theta}))$, so that if $\hat{\boldsymbol{W}} \xrightarrow{\mathcal{P}} \boldsymbol{W}$ for some positive semi-definite matrix $\boldsymbol{W}$, then by the continuity of the product, $\widehat{Q_n}(\boldsymbol{\theta}) \xrightarrow{\mathcal{P}} -E(\boldsymbol{U}_i(\boldsymbol{\theta}))'\boldsymbol{W}E(\boldsymbol{U}_i(\boldsymbol{\theta}))$. This function has a maximum equals to zero at $\boldsymbol{\theta} = \boldsymbol{\theta}^*$, so $\boldsymbol{\theta}^*$ will be identified if $-E(\boldsymbol{U}_i(\boldsymbol{\theta}))'\boldsymbol{W}E(\boldsymbol{U}_i(\boldsymbol{\theta})) < 0$ for $\boldsymbol{\theta} \neq \boldsymbol{\theta}^*$.

Lemma 5.4.1 *(Lemma 2.3, GMM identification, in Newey and McFadden (1994)) If $\boldsymbol{W}$ is positive semi-definite and, $E(\boldsymbol{U}_i(\boldsymbol{\theta}^*)) = \mathbf{0}$ and $\boldsymbol{W}E(\boldsymbol{U}_i(\boldsymbol{\theta})) \neq \mathbf{0}$ for $\boldsymbol{\theta} \neq \boldsymbol{\theta}^*$ then $-E(\boldsymbol{U}_i(\boldsymbol{\theta}))'\boldsymbol{W}E(\boldsymbol{U}_i(\boldsymbol{\theta}))$ has a unique maximum at $\boldsymbol{\theta}^*$.*

That is, $\boldsymbol{\theta}^*$ is identified if $\boldsymbol{\theta} \neq \boldsymbol{\theta}^*$ implies that $E(\boldsymbol{U}_i(\boldsymbol{\theta}))$ is not in the null space of $\boldsymbol{W}$. In particular, if $\boldsymbol{W}$ is nonsingular, this condition is equivalent to $E(\boldsymbol{U}_i(\boldsymbol{\theta})) \neq \mathbf{0}$ if $\boldsymbol{\theta} \neq \boldsymbol{\theta}^*$. A necessary order condition for GMM identification is that the dimension of moment functions would be at least the dimension of parameters.

A consistency result for GMM can be formulated as follows:

Theorem 5.4.1 *(Theorem 2.6, consistency of GMM, in Newey and McFadden (1994)) Suppose that data are i.i.d., $\hat{\boldsymbol{W}} \xrightarrow{\mathcal{P}} \boldsymbol{W}$, and*

1. *$\boldsymbol{W}$ is positive semi-definite,*
2. *$\boldsymbol{W}E(\boldsymbol{U}_i(\boldsymbol{\theta})) = \mathbf{0}$ only if $\boldsymbol{\theta} = \boldsymbol{\theta}^*$,*
3. *$\boldsymbol{\theta}^* \in \boldsymbol{\Theta}$, which is compact,*
4. *$\boldsymbol{U}_i(\boldsymbol{\theta})$ is continuous at each $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ with probability one,*
5. *$E(\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \|\boldsymbol{U}_i(\boldsymbol{\theta})\|) < \infty$.*

Then the GMM solution in (5.6) consistently estimates $\boldsymbol{\theta}^*$, that is, $\widehat{\boldsymbol{\theta}} \xrightarrow{\mathcal{P}} \boldsymbol{\theta}^*$.

The asymptotically normal behavior of the GMM estimators is established in the next theorem.

Theorem 5.4.2 *(Theorem 3.4, asymptotic normality for GMM, in Newey and McFadden (1994)) Suppose that the hypotheses in Theorem 5.4.1 are fulfilled and*

1. $\boldsymbol{\theta}^* \in$ *interior*$(\boldsymbol{\Theta})$,
2. $\boldsymbol{U}_i(\boldsymbol{\theta})$ *is continuously differentiable in a neighborhood* N *of* $\boldsymbol{\theta}^*$, *with probability approaching one,*
3. $E(\boldsymbol{U}_i(\boldsymbol{\theta}^*)) = \mathbf{0}$ *and* $E(\|\boldsymbol{U}_i(\boldsymbol{\theta}^*)\|^2)$ *is finite,*
4. $E(\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \|\partial \boldsymbol{U}_i(\boldsymbol{\theta})/\partial\boldsymbol{\theta}\|) < \infty$
5. $\boldsymbol{G}'\boldsymbol{W}\boldsymbol{G}$ *is nonsingular for* $\boldsymbol{G} = E(\partial\boldsymbol{U}_i(\boldsymbol{\theta}^*)/\partial\boldsymbol{\theta})$.

Then for $\boldsymbol{\Omega} = E(\boldsymbol{U}_i(\boldsymbol{\theta}^*)\boldsymbol{U}_i(\boldsymbol{\theta}^*)')$,

$$\sqrt{n}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*) \xrightarrow{\mathcal{D}} \mathcal{N}(\mathbf{0}, (\boldsymbol{G}'\boldsymbol{W}\boldsymbol{G})^{-1}\boldsymbol{G}'\boldsymbol{W}\Omega\boldsymbol{W}\boldsymbol{G}(\boldsymbol{G}'\boldsymbol{W}\boldsymbol{G})^{-1}).$$

An important issue about the proof of this theorem is that the fulfillment of the hypotheses in Theorem 5.4.1 is only required in order to ensure the consistency of the resulting estimators. Other issue is to note that the expression for the asymptotic variance simplifies to $\boldsymbol{G}^{-1}\boldsymbol{\Omega}\boldsymbol{G}^{-1'}$ when we use $\boldsymbol{W} = \boldsymbol{\Omega}^{-1}$ as a metrics or weighting matrix.

The asymptotic variance of a GMM estimator can be consistently estimated by substituting estimators for each of $\boldsymbol{G}$, $\boldsymbol{W}$ and $\boldsymbol{\Omega}$. Next theorem summarizes this result.

Theorem 5.4.3 *(Theorem 4.5, asymptotic variance estimation for GMM, in Newey and McFadden (1994)) Suppose that the hypotheses in Theorem 5.4.2 are satisfied and* $E(\sup_{\boldsymbol{\theta}\in N} \|\boldsymbol{U}_i(\boldsymbol{\theta})\|^2)$, *in a neighborhood* N *of* $\boldsymbol{\theta}^*$, *then*

$$(\hat{\boldsymbol{G}}'\hat{\boldsymbol{W}}\hat{\boldsymbol{G}})^{-1}\hat{\boldsymbol{G}}'\hat{\boldsymbol{W}}\hat{\Omega}\hat{\boldsymbol{W}}\hat{\boldsymbol{G}}(\hat{\boldsymbol{G}}'\hat{\boldsymbol{W}}\hat{\boldsymbol{G}})^{-1} \xrightarrow{\mathcal{P}} (\boldsymbol{G}'\boldsymbol{W}\boldsymbol{G})^{-1}\boldsymbol{G}'\boldsymbol{W}\Omega\boldsymbol{W}\boldsymbol{G}(\boldsymbol{G}'\boldsymbol{W}\boldsymbol{G})^{-1}.$$

5.5 IPWGEE class of estimators

Liang and Zeger (1986) proposed the GEE to obtain consistent and asymptotically normal estimates of the regression coefficient, $\boldsymbol{\beta}^*$, in the conditional expectation model

$$E(Y|\boldsymbol{X}) = g(\boldsymbol{X}; \boldsymbol{\beta}^*), \tag{5.8}$$

solving

$$\boldsymbol{U}(\boldsymbol{\beta}) = \sum_{i=1}^{n} \boldsymbol{U}_i(\boldsymbol{\beta}) = \sum_{i=1}^{n} d(\boldsymbol{X}_i; \boldsymbol{\beta}) \left(Y_i - g(\boldsymbol{X}_i; \boldsymbol{\beta})\right) = \boldsymbol{0} \tag{5.9}$$

where

$$d(\boldsymbol{X}_i; \boldsymbol{\beta}) = \left(\partial g(\boldsymbol{X}_i; \boldsymbol{\beta}) / \partial \boldsymbol{\beta}\right) \left(\mathrm{Var}(Y_i|\boldsymbol{X}_i)\right)^{-1}.$$

Indeed, (5.9) is not a estimating equation because $\mathrm{Var}(Y_i|\boldsymbol{X}_i)$ depends on the unknown distribution generating the data. So, Liang and Zeger solved recursively these equations by modeling $\mathrm{Var}(Y_i|\boldsymbol{X}_i)$ and replacing it by a current estimate based on the current estimate $\widehat{\boldsymbol{\beta}}$. This procedure is referred to as *adaptive* method. Chamberlain (1987) proved that in the absence of missing data the asymptotic variance of the solution of (5.9) achieves the semiparametric variance bound in the semiparametric model (5.8). So, the adaptive procedure proposed by Liang and Zeger provides locally efficient semiparametric estimators for $\boldsymbol{\beta}^*$.

In this section these GEE estimators are extended to a more general class, still inside the GMM class of estimators. In what follows, we will assume that $\boldsymbol{\Omega} = \mathrm{Var}(\boldsymbol{U}_i(\boldsymbol{\beta}^*))$ is finite and positive definite, and, as a consequence, we will setup $\boldsymbol{W} = \boldsymbol{\Omega}^{-1}$.

In a missing data context, denote the vector of potential data by $\boldsymbol{L}_i = (Y_i, \boldsymbol{X}_i', \boldsymbol{V}_i')'$ for $i = 1, \ldots, n$. Consider that the vector $\boldsymbol{X}_i$ is p-dimensional and possibly has missing components, and Y_i and $\boldsymbol{V}_i$ are always observed. For the j-th component of $\boldsymbol{X}$, we define R_j as the binary variable that takes value 1 if this component of $\boldsymbol{X}$ has been observed, and 0 otherwise. $\boldsymbol{R} = (R_1, \ldots, R_p)'$ is the indicator vector of response in the random vector $\boldsymbol{X}$. For the i-th individual, $i = 1, \ldots, n$, we consider the realization $\boldsymbol{R}_i$ of the variable $\boldsymbol{R}$, and we denote by $\boldsymbol{L}_{(\boldsymbol{R}_i)i}$ the subvector of $\boldsymbol{L}_i$ formed by the observed components and by $\boldsymbol{L}_{\left(\overline{\boldsymbol{R}_i}\right)i}$ the subvector of $\boldsymbol{L}_i$ formed by the non-observed ones. So, the observed data are $\left\{\boldsymbol{R}_i, \boldsymbol{L}_{(\boldsymbol{R}_i)i}\right\}_{i=1,\ldots,n}$.

We will consider now that the non-response pattern is non-ignorable, in the sense that the probabilities of response might depend on the non-observed variables. It means that for each realization $\boldsymbol{r}$ of the variable $\boldsymbol{R}$, the conditional probability $P(\boldsymbol{R}=\boldsymbol{r}|\boldsymbol{L})$ depends on partially non-observed components of $\boldsymbol{Y}$. We will denote this probability by $\pi(\boldsymbol{r})$.

In this situation, Rotnitzky and Robins (1997) proposed the inverse probability of being observed weighted generalized estimating equations (IPWGEE) as an extension of the previously proposed class of estimators for MAR non-response pattern in covariates (Robins et al., 1994). Later, Rotnitzky, Robins and Scharfstein (1998) derived similar properties for the repeated outcomes subject to non-ignorable non-response case. In their methodology, they considered that probabilities $\pi(\boldsymbol{r})$ can be parametrically specified depending on a unknown q-dimensional parameter $\boldsymbol{\alpha}$ (*i.e.,* they assumed $\pi(\boldsymbol{r})=\pi(\boldsymbol{r};\boldsymbol{\alpha})$).

The IPWGEE class of estimators are the solutions to the estimating equations

$$\boldsymbol{U}(\boldsymbol{\beta},\boldsymbol{\alpha};d,\phi)=\sum_{i=1}^{n}\boldsymbol{U}_i(\boldsymbol{\beta},\boldsymbol{\alpha};d,\phi)=\mathbf{0} \tag{5.10}$$

where

$$\boldsymbol{U}_i(\boldsymbol{\beta},\boldsymbol{\alpha};d,\phi)=\frac{I(\boldsymbol{R}_i=\mathbf{1})}{\pi_i\left(\mathbf{1};\boldsymbol{\alpha}\right)}d(\boldsymbol{X}_i;\boldsymbol{\beta})\left(Y_i-g(\boldsymbol{X}_i;\boldsymbol{\beta})\right)+\boldsymbol{A}_i(\phi) \tag{5.11}$$

and

$$\boldsymbol{A}_i(\phi)=\sum_{\boldsymbol{r}\neq\mathbf{1}}\left(\left\{I(\boldsymbol{R}_i=\boldsymbol{r})-\frac{I(\boldsymbol{R}_i=\mathbf{1})}{\pi_i(\mathbf{1};\boldsymbol{\alpha})}\pi_i(\boldsymbol{r};\boldsymbol{\alpha})\right\}\phi_{\boldsymbol{r}}(\boldsymbol{L}_{(\boldsymbol{r})i})\right) \tag{5.12}$$

with $d(.;.)$ and $\phi_{\boldsymbol{r}}(.)$ arbitrary vectorial functions of dimension equal to those of the joint parameter $\boldsymbol{\gamma}=(\boldsymbol{\beta}',\boldsymbol{\alpha}')'$.

In (5.11) we can see that $\pi_i^{-1}(\mathbf{1};\boldsymbol{\alpha})\{d(\boldsymbol{X}_i;\boldsymbol{\beta})\epsilon_i-\sum_{\boldsymbol{r}\neq\mathbf{1}}\pi_i(\boldsymbol{r};\boldsymbol{\alpha})\phi_{\boldsymbol{r}}\left(\boldsymbol{L}_{(\boldsymbol{r})i}\right)\}$ is the contribution of a fully observed i-th individual, and each individual with partially observed covariates contributes $\phi_{\boldsymbol{R}_i}(\boldsymbol{L}_{(\boldsymbol{R}_i)})$ in (5.10). On the other hand, if $\boldsymbol{\beta}^*,\boldsymbol{\alpha}^*$ and $\boldsymbol{\gamma}^*$ denote the true values of the parameters $\boldsymbol{\beta},\boldsymbol{\alpha}$ and $\boldsymbol{\gamma}$, it is straightforward to prove, after taking conditional expectations on $\boldsymbol{L}_i$, that $E(\boldsymbol{U}_i(\boldsymbol{\beta}^*,\boldsymbol{\alpha}^*;d,\phi))=0$. Therefore IPWGEE is a class of unbiased estimating equations and they will provide consistent and asymptotically normal estimates for $\boldsymbol{\gamma}^*$ (Newey, 1990). It is important to observe that the augmented term $\boldsymbol{A}_i(\phi)$ in (5.11) it is not necessary to derive unbiased estimating equations, but it is convenient to use the information of partially observed individuals in order to gain efficiency. In a similar way that in Lemma 5.4.1 and Theorems 5.4.1, 5.4.2 and 5.4.3, the next theorem describes the asymptotic properties of the solutions $\left(\widehat{\boldsymbol{\beta}}', \widehat{\boldsymbol{\alpha}}'\right)'$ to (5.10).

Theorem 5.5.1 *(Rotnitzky et al., 1998) Under the mild regularity conditions*

1. *$\boldsymbol{\gamma}$ lies in the interior of a compact set,*

2. *$(\boldsymbol{L}_i, \boldsymbol{R}_i)$, $i = 1, \ldots, n$ are independently and identically distributed,*

3. *for some c, $\pi(\mathbf{1}; \boldsymbol{\alpha}) > c > 0$ for all $\boldsymbol{\alpha}$,*

4. *$E(\boldsymbol{U}_i(\boldsymbol{\gamma}; d, \phi)) \neq \mathbf{0}$ if $\boldsymbol{\gamma} \neq \boldsymbol{\gamma}^*$,*

5. *Var$(\boldsymbol{U}_i(\boldsymbol{\gamma}^*; d, \phi))$ is finite and positive definite,*

6. *$\boldsymbol{\Gamma} = E\left(\partial \boldsymbol{U}_i(\boldsymbol{\gamma}^*; d, \phi)/\partial \boldsymbol{\gamma}\right)$ exists and is invertible,*

7. *there exist a neighborhood N of $\boldsymbol{\gamma}^*$ such that $E(\sup_{\boldsymbol{\gamma} \in N} \|\boldsymbol{U}_i(\boldsymbol{\gamma}; d, \phi)\|)$, $E(\sup_{\boldsymbol{\gamma} \in N} \|\partial \boldsymbol{U}_i(\boldsymbol{\gamma}; d, \phi)/\partial \boldsymbol{\gamma}\|)$, and $E(\sup_{\boldsymbol{\gamma} \in N} \|\boldsymbol{U}_i(\boldsymbol{\gamma}; d, \phi)\boldsymbol{U}_i'(\boldsymbol{\gamma}; d, \phi)\|)$ are all finite, where $\|\boldsymbol{A}\| = (\sum_{ij} a_{ij}^2)^{1/2}$ for any matrix $\boldsymbol{A} = (a_{ij})$,*

8. *$f(\boldsymbol{L}, \boldsymbol{R}; \boldsymbol{\gamma})$ is a regular parametric model where $f(\boldsymbol{L}, \boldsymbol{R}; \boldsymbol{\gamma})$ is a density that differs from the true density $f(\boldsymbol{L}, \boldsymbol{R}) = f(\boldsymbol{L}, \boldsymbol{R}; \boldsymbol{\gamma}^*)$ only in that $\boldsymbol{\gamma}$ replaces $\boldsymbol{\gamma}^*$,*

9. *for all $\bar{\boldsymbol{\gamma}}$ in a neighborhood N of $\boldsymbol{\gamma}^*$, $E_{\bar{\boldsymbol{\gamma}}}(\boldsymbol{U}_i(\boldsymbol{\gamma}; d, \phi))$ and $E_{\bar{\boldsymbol{\gamma}}}(\sup_{\boldsymbol{\gamma} \in N} \|\boldsymbol{U}_i(\boldsymbol{\gamma}; d, \phi)\boldsymbol{U}_i'(\boldsymbol{\gamma}; d, \phi)\|)$ are bounded, where $E_{\bar{\boldsymbol{\gamma}}}$ refers to expectation with respect to the density $f(\boldsymbol{L}, \boldsymbol{R}; \bar{\boldsymbol{\gamma}})$,*

if the function g and the model for $\pi(\boldsymbol{r})$ are correctly specified then

a) *with probability approaching 1, there is a unique solution $\widehat{\boldsymbol{\gamma}}$ to (5.10),*

b) *$\sqrt{n}\,(\widehat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}^*)$ asymptotically follows a $\mathcal{N}(\mathbf{0}, \boldsymbol{\Upsilon})$ distribution, with $\boldsymbol{\Upsilon} = \boldsymbol{\Gamma}^{-1}\boldsymbol{\Omega}\boldsymbol{\Gamma}^{-1'}$ with $\boldsymbol{\Omega} =$ Var$(\boldsymbol{U}_i(\boldsymbol{\gamma}^*; d, \phi))$*

c) the asymptotic variance-covariance matrix $\boldsymbol{\Upsilon}$ *can be consistently estimated by* $\widehat{\boldsymbol{\Upsilon}} = \widehat{\boldsymbol{\Gamma}}^{-1}\widehat{\boldsymbol{\Omega}}\widehat{\boldsymbol{\Gamma}}^{-1'}$ *where* $\widehat{\boldsymbol{\Gamma}} = n^{-1}\sum \partial \boldsymbol{U}_i(\boldsymbol{\gamma}; d, \phi)/\partial\boldsymbol{\gamma}$ *and* $\widehat{\boldsymbol{\Omega}} = n^{-1}\sum \boldsymbol{U}_i(\boldsymbol{\gamma}; d, \phi)\boldsymbol{U}_i'(\boldsymbol{\gamma}; d, \phi)$ *are evaluated in* $\boldsymbol{\gamma} = \widehat{\boldsymbol{\gamma}}$.

Part a) holds by applying Theorem 5.4.2. Due to the nonsingularity of $\boldsymbol{\Gamma}$, part b) follows by Slutzky's theorem and the central limit theorem. The consistency of the variance estimators in c) follows from the law of large numbers.

One of the advantages of the IPWGEE class is that the solutions to (5.10) essentially comprise all regular and asymptotically linear estimators of $\boldsymbol{\beta}^*$ as is stated in the next theorem (Lemma 1 in Rotnitzky *et al.* (1998)).

Theorem 5.5.2 *If* $\bar{\boldsymbol{\beta}}$ *is any regular and asymptotically linear estimator of* $\boldsymbol{\beta}^*$ *in the semiparametric model of the conditional expectation, with* $\pi_i(\mathbf{1})$ *bounded away from 0 with probability 1, and the model for* $\pi_i(\boldsymbol{r}; \boldsymbol{\alpha})$ *is correctly specified, then there exist functions* $d(\boldsymbol{X}_i; \boldsymbol{\beta})$ *and* $\phi_{\boldsymbol{r}}(L_{(\boldsymbol{r})i})$, $\boldsymbol{r} \neq \mathbf{1}$, *such that for* $\widehat{\boldsymbol{\beta}}$ *solving (5.10) using these functions, then* $\sqrt{n}(\bar{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}})$ *converges to 0 in probability, and thus* $\sqrt{n}(\bar{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)$ *and* $\sqrt{n}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)$ *have the same asymptotic distribution.*

5.6 Efficiency

The most interesting point of the IPWGEE class of estimators is that it includes the efficient semiparametric estimator for the specified conditional expectation semiparametric model. This result was established in Theorem 2 in Rotnitzky *et al.* (1998).

Theorem 5.6.1 *Under the same hypothesis in Theorem 5.5.1, there exist functions* $d_{opt}(\boldsymbol{X}_i; \boldsymbol{\beta}^*)$ *and* $\phi_{\boldsymbol{r}\,opt}(\boldsymbol{L}_{(\boldsymbol{r})i})$ *such that the asymptotic variance of the solution* $\left(\widehat{\boldsymbol{\beta}}', \widehat{\boldsymbol{\alpha}}'\right)'$ *of (5.10) equals* $\boldsymbol{\Omega}_{opt}^{-1} = (\,Var(\boldsymbol{U}_i(\boldsymbol{\beta}^*, \boldsymbol{\alpha}^*; d_{opt}, \phi_{opt})))^{-1}$. *Furthermore it attains the semiparametric variance bound for regular estimators of* $(\boldsymbol{\beta}^{*\prime}, \boldsymbol{\alpha}^{*\prime})'$. *In particular, the upper left* $p \times p$ *submatrix of* $\boldsymbol{\Omega}_{opt}^{-1}$ *is the semiparametric variance bound for* $\boldsymbol{\beta}^*$.

The optimal functions $d_{\text{opt}}(\boldsymbol{X}_i; \boldsymbol{\beta}^*)$ and $\phi_{\boldsymbol{r}\text{opt}}(\boldsymbol{L}_{(\boldsymbol{r})i})$ are not available for data analysis since they depend on the unknown true distribution generating the data.

For the missing covariates case there is no closed-form for $d_{\text{opt}}(\boldsymbol{X}_i;\boldsymbol{\beta}^*)$ and $\phi_{\boldsymbol{r}\text{opt}}(\boldsymbol{L}_{(\boldsymbol{r})i})$, but they are solutions of functional integral equations without analytic solution. As a consequence, the adaptive procedure has to add an extra iterative step to solve these functional equations. This means that, in practice, the methodology to obtain efficient semiparametric estimators will be computationally intensive. Rotnitzky and Robins (1997) gave these equations in detail for the case in which the subvector of $\boldsymbol{X}$ with possibly non-observed components is fully observed or not (*e.g.,* if $p = 1$), and in Appendix IV they described how to obtain locally efficient semiparametric estimators $\widehat{\boldsymbol{\beta}}(\widehat{d_{\text{opt}}},\widehat{\phi_{\text{opt}}})$ of $\boldsymbol{\beta}^*$.

5.7 Sensitivity analysis

In a non-ignorable non-response pattern setting and in order to have a better interpretation of the probabilities $\pi(\boldsymbol{r})$, for each realization $\boldsymbol{r}$ of the random variable $\boldsymbol{R}$ we will model the log-ratio between individuals who have only observed the covariates indicated by $\boldsymbol{r}$ and those with $\boldsymbol{X}$ completely observed, in the form

$$\log\frac{P(\boldsymbol{R}=\boldsymbol{r}|Y,\boldsymbol{X},\boldsymbol{V})}{P(\boldsymbol{R}=\mathbf{1}|Y,\boldsymbol{X},\boldsymbol{V})}=m_{\boldsymbol{r}}(Y,\boldsymbol{V};\boldsymbol{\alpha})+q_{\boldsymbol{r}}(Y,\boldsymbol{X},\boldsymbol{V};\boldsymbol{\tau}). \tag{5.13}$$

This characterization allows us to separate the non-response pattern in two parts: $m_{\boldsymbol{r}}(Y,\boldsymbol{V};\boldsymbol{\alpha})$ the ignorable part, depending on the parameter $\boldsymbol{\alpha}$, and $q_{\boldsymbol{r}}(Y,\boldsymbol{X},\boldsymbol{V};\boldsymbol{\tau})$ the non-ignorable part, depending on $\boldsymbol{\tau}$. We will consider that $q_{\boldsymbol{r}}(.)$ only contains the contributions of Y and $\boldsymbol{V}$ that are inseparably related with $\boldsymbol{X}$ and they can not be reduced to a separate form (*e.g.,* $q_{\boldsymbol{r}}(.)$ includes interactions between Y or $\boldsymbol{V}$ and some components of $\boldsymbol{X}$). In the following, we will refer to $\boldsymbol{\tau}$ as the *non-ignorability parameter* or *selection bias parameter.* This model allows us to include the MCAR and MAR mechanisms as particular cases for a pre-determined setup of parameters $\boldsymbol{\alpha}$ and $\boldsymbol{\tau}$ (*e.g.,* when $\boldsymbol{\tau} = 0$ we could derive the MAR case).

When we have all the data, or the distributions are known, the probabilities on the left-hand side of (5.13) are well defined, but quantities $m_{\boldsymbol{r}}(.)$ and $q_{\boldsymbol{r}}(.)$ on the right-hand side are identified, up to translation (*i.e.,* if $m_{\boldsymbol{r}}(.)$ and $q_{\boldsymbol{r}}(.)$ verify (5.13), then $m'_{\boldsymbol{r}}(.) = m_{\boldsymbol{r}}(.) - k$ and $q'_{\boldsymbol{r}}(.) = q_{\boldsymbol{r}}(.) + k$ so do, for each value of k). Then, in order to unequivocally identify $m_{\boldsymbol{r}}(.)$ and $q_{\boldsymbol{r}}(.)$ we have to fix the value of the function $q_{\boldsymbol{r}}(.)$ on some pre-determined value of the variables that appear in $q_{\boldsymbol{r}}(.)$ and not in $m_{\boldsymbol{r}}(.)$ (*i.e.,* of $\boldsymbol{X}$). Without loss of generality, we can choose $\boldsymbol{X} = \boldsymbol{0}$ and fix $q_{\boldsymbol{r}}(Y, \boldsymbol{0}, \boldsymbol{V}) = 0$. Then, if we denote by $\ell_{\boldsymbol{r}}(Y, \boldsymbol{X}, \boldsymbol{V}; \boldsymbol{\alpha}, \boldsymbol{\tau})$ the log-ratio probabilities in the equation (5.13), $m_{\boldsymbol{r}}(.)$ and $q_{\boldsymbol{r}}(.)$ are univocally identified as

$$\begin{aligned} m_{\boldsymbol{r}}(Y, \boldsymbol{V}; \boldsymbol{\alpha}) &= \ell_{\boldsymbol{r}}(Y, \boldsymbol{0}, \boldsymbol{V}; \boldsymbol{\alpha}, \boldsymbol{\tau}) \\ q_{\boldsymbol{r}}(Y, \boldsymbol{X}, \boldsymbol{V}; \boldsymbol{\tau}) &= \ell_{\boldsymbol{r}}(Y, \boldsymbol{X}, \boldsymbol{V}; \boldsymbol{\alpha}, \boldsymbol{\tau}) - \ell_{\boldsymbol{r}}(Y, \boldsymbol{0}, \boldsymbol{V}; \boldsymbol{\alpha}, \boldsymbol{\tau}). \end{aligned}$$

We can see that in the univariate case (*i.e.,* $p = 1$), (5.13) is equivalent to model $P(\boldsymbol{R} = \boldsymbol{r}|\boldsymbol{L})$ according to a logistic model in the form

$$\text{logit } (P(R = 1|\boldsymbol{L})) = m(Y, \boldsymbol{V}; \boldsymbol{\alpha}) + q(Y, \boldsymbol{X}, \boldsymbol{V}; \boldsymbol{\tau}). \tag{5.14}$$

Assume that the non-response model in (5.13) or (5.14) is correctly specified and denote by $\boldsymbol{\alpha}^*$ and $\boldsymbol{\tau}^*$ the true non-response parameters. When we have missing data in the covariates vector $\boldsymbol{X}$, we can not estimate parameter $\boldsymbol{\tau}^*$ directly from the sample because for those individuals with partially observed covariates we can not derive their contribution to the likelihood (without making extra parametric assumptions on the distribution of $\boldsymbol{X}$, and the conditional on $\boldsymbol{X}$ distributions of Y and V). That is $\boldsymbol{\tau}^*$ is not identified. So, in order to conduct a sensitivity analysis, for each value of a plausible range of values for the non-ignorability parameter, $\boldsymbol{\tau}$, we will estimate consistently $\widehat{\boldsymbol{\alpha}}(\boldsymbol{\tau})$ and our parameters of interest. Finally, a graphical description of the estimates for the parameter of interest can be done in order to illustrate the robustness of the inferences and their sensitivity to the non-validable assumptions. In the context of conducting this type of sensitivity analysis, Rotnitzky *et al.* (1998) gave in Section 7.4 some practical recommendations to choose d and ϕ in a computable "easy" way. In particular, they suggested how to setup d and ϕ in order to derive quite efficient estimators of $\boldsymbol{\beta}^*$ under moderate departures from the MAR hypothesis. These suggestions will be taken into account in Chapter 6.

If we are analyzing survival data, the outcome random variable Y will be replaced by the observed survival time and the censoring indicator, Y and δ.