Springer Series in Statistics

Springer Series in Statistics

Alho/Spencer: Statistical Demography and Forecasting.
Andersen/Borgan/Gill/Keiding: Statistical Models Based on Counting Processes.
Atkinson/Riani: Robust Diagnostic Regression Analysis.
Atkinson/Riani/Cerioli: Exploring Multivariate Data with the Forward Search.
Berger: Statistical Decision Theory and Bayesian Analysis, 2nd edition.
Borg/Groenen: Modern Multidimensional Scaling: Theory and Applications, 2nd edition.
Brockwell/Davis: Time Series: Theory and Methods, 2nd edition.
Bucklew: Introduction to Rare Event Simulation.
Cappé/Moulines/Rydén: Inference in Hidden Markov Models.
Chan/Tong: Chaos: A Statistical Perspective.
Chen/Shao/Ibrahim: Monte Carlo Methods in Bayesian Computation.
Coles: An Introduction to Statistical Modeling of Extreme Values.
David/Edwards: Annotated Readings in the History of Statistics.
Devroye/Lugosi: Combinatorial Methods in Density Estimation.
Efromovich: Nonparametric Curve Estimation: Methods, Theory, and Applications.
Eggermont/LaRiccia: Maximum Penalized Likelihood Estimation, Volume I: Density Estimation.
Fahrmeir/Tutz: Multivariate Statistical Modelling Based on Generalized Linear Models, 2nd edition.
Fan/Yao: Nonlinear Time Series: Nonparametric and Parametric Methods.
Farebrother: Fitting Linear Relationships: A History of the Calculus of Observations 1750-1900.
Federer: Statistical Design and Analysis for Intercropping Experiments, Volume I: Two Crops.
Federer: Statistical Design and Analysis for Intercropping Experiments, Volume II: Three or More Crops.
Ferraty/Vieu: Nonparametric Functional Data Analysis: Models, Theory, Applications, and Implementation
Ghosh/Ramamoorthi: Bayesian Nonparametrics.
Glaz/Naus/Wallenstein: Scan Statistics.
Good: Permutation Tests: Parametric and Bootstrap Tests of Hypotheses, 3rd edition.
Gouriéroux: ARCH Models and Financial Applications.
Gu: Smoothing Spline ANOVA Models.
Györfi/Kohler/Krzyżak/Walk: A Distribution-Free Theory of Nonparametric Regression.
Haberman: Advanced Statistics, Volume I: Description of Populations.
Hall: The Bootstrap and Edgeworth Expansion.
Härdle: Smoothing Techniques: With Implementation in S.
Harrell: Regression Modeling Strategies: With Applications to Linear Models, Logistic Regression, and Survival Analysis.
Hart: Nonparametric Smoothing and Lack-of-Fit Tests.
Hastie/Tibshirani/Friedman: The Elements of Statistical Learning: Data Mining, Inference, and Prediction.
Hedayat/Sloane/Stufken: Orthogonal Arrays: Theory and Applications.
Heyde: Quasi-Likelihood and its Application: A General Approach to Optimal Parameter Estimation.

(continued after index)

Anastasios A. Tsiatis

Semiparametric Theory and Missing Data

Springer

Anastasios A. Tsiatis
Department of Statistics
North Carolina State University
Raleigh, NC 27695
USA
tsiatis@stat.ncsu.edu

Library of Congress Control Number: 2006921164

ISBN-10: 0-387-32448-8
ISBN-13: 978-0387-32448-7

Printed on acid-free paper.

Printed in the United States of America. (MVY)

9 8 7 6 5 4 3 2 1

springer.com

To
My Mother, Anna
My Wife, Marie
and
My Son, Greg

Preface

Missing data are prevalent in many studies, especially when the studies involve human beings. Not accounting for missing data properly when analyzing data can lead to severe biases. For example, most software packages, by default, delete records for which any data are missing and conduct the so-called "complete-case analysis". In many instances, such an analysis will lead to an incorrect inference. Since the 1980s there has been a serious attempt to understand the underlying issues involved with missing data. In this book, we study the different mechanisms for missing data and some of the different analytic strategies that have been suggested in the literature for dealing with such problems. A special case of missing data includes censored data, which occur frequently in the area of survival analysis. Some discussion of how the missing-data methods that are developed will apply to problems with censored data is also included.

Underlying any missing-data problem is the statistical model for the data if none of the data were missing (i.e., the so-called full-data model). In this book, we take a very general approach to statistical modeling. That is, we consider statistical models where interest focuses on making inference on a finite set of parameters when the statistical model consists of the parameters of interest as well as other nuisance parameters. Unlike most traditional statistical models, where the nuisance parameters are finite-dimensional, we consider the more general problem of infinite-dimensional nuisance parameters. This allows us to develop theory for important statistical methods such as regression models that model the conditional mean of a response variable as a function of covariates without making any additional distributional assumptions on the variables and the proportional hazards regression model for survival data. Models where the parameters of interest are finite-dimensional and the nuisance parameters are infinite-dimensional are called semiparametric models.

The first five chapters of the book consider semiparametric models when there are no missing data. In these chapters, semiparametric models are defined and some of the theoretical developments for estimators of the parame-

ters in these models are reviewed. The semiparametric theory and the properties of the estimators for parameters in semiparametric models are developed from a geometrical perspective. Consequently, in Chapter 2, a quick review of the geometry of Hilbert spaces is given. The geometric ideas are first developed for finite-dimensional parametric models in Chapter 3 and then extended to infinite-dimensional models in Chapters 4 and 5.

A rigorous treatment of semiparametric theory is given in the book *Efficient and Adaptive Estimation for Semiparametric Models* by Bickel et al. (1993). (Johns Hopkins University Press: Baltimore, MD). My experience has been that this book is too advanced for many students in statistics and biostatistics even at the Ph.D. level. The attempt here is to be more expository and heuristic, trying to give an intuition for the basic ideas without going into all the technical details. Although the treatment of this subject is not rigorous, it is not trivial either. Readers should not be frustrated if they don't grasp all the concepts at first reading. This first part of the book that deals only with semiparametric models (absent missing data) and the geometric theory of semiparametrics will be important in its own right. It is a beautiful theory, where the geometric perspective gives a new insight and deeper understanding of statistical models and the properties of estimators for parameters in such models.

The remainder of the book focuses on missing-data methods, building on the semiparametric techniques developed in the earlier chapters. In Chapter 6, a discussion and overview of missing-data mechanisms is given. This includes the definition and motivation for the three most common categories of missingness, namely

- missing completely at random (MCAR)
- missing at random (MAR)
- nonmissing at random (NMAR)

These ideas are extended to the broader class of coarsened data. We show how statistical models for full data can be integrated with missingness or coarsening mechanisms that allow us to derive likelihoods and models for the observed data in the presence of missingness. The geometric ideas for semiparametric full-data models are extended to missing-data models. This treatment will give the reader a deep understanding of the underlying theory for missing and coarsened data. Methods for estimating parameters with missing or coarsened data in as efficient a manner as possible are emphasized. This theory leads naturally to inverse probability weighted complete-case (IPWCC) and augmented inverse probability weighted complete-case (AIPWCC) estimators, which are discussed in great detail in Chapters 7 through 11. As we will see, some of the proposed methods can become computationally challenging if not infeasible. Therefore, in Chapter 12, we give some approximate methods for obtaining more efficient estimators with missing data that are easier to implement. Much of the theory developed in this book is taken from a series of

ground-breaking papers by Robins and Rotnitzky (together with colleagues), who developed this elegant semiparametric theory for missing-data problems.

A short discussion on how missing-data semiparametric methods can be applied to estimate causal treatment effects in a point exposure study is given in Chapter 13 to illustrate the broad applicability of these methods. In Chapter 14, the final chapter, we deviate slightly from semiparametric models to discuss some of the theoretical properties of multiple-imputation estimators for finite-dimensional parametric models. However, even here, the theory developed throughout the book will be useful in understanding the properties of such estimators.

Anastasios (Butch) Tsiatis

Contents

1

Introduction to Semiparametric Models

Statistical problems are described using probability models. That is, data are envisioned as realizations of a vector of random variables $Z_1, \ldots, Z_n$, where Z_i itself is a vector of random variables corresponding to the data collected on the i-th individual in a sample of n individuals chosen from some population of interest. We will assume throughout the book that $Z_1, \ldots, Z_n$ are identically and independently distributed (iid) with density belonging to some probability (or statistical model), where a model consists of a class of densities that we believe might have generated the data. The densities in a model are often identified through a set of parameters; i.e., a real-valued vector used to describe the densities in a statistical model. The problem is usually set up in such a way that the value of the parameters or, at the least, the value of some subset of the parameters that describes the density that generates the data, is of importance to the investigator. Much of statistical inference considers how we can learn about this "true" parameter value from a sample of observed data. Models that are described through a vector of a finite number of real values are referred to as finite-dimensional parametric models. For finite-dimensional parametric models, the class of densities can be described as

$$\mathcal{P} = \{p(z, \theta), \theta \in \Omega \subset \mathbb{R}^p\},$$

where the dimension p is some finite positive integer.

For many problems, we are interested in making inference only on a subset of the parameters. Nonetheless, the entire set of parameters is necessary to properly describe the class of possible distributions that may have generated the data. Suppose, for example, we are interested in estimating the mean response of a variable, which we believe follows a normal distribution. Typically, we conduct an experiment where we sample from that distribution and describe the data that result from that experiment as a realization of the random vector

$$Z_1, \ldots, Z_n \text{ assumed iid } N(\mu, \sigma^2);\ \mu \in \mathbb{R}, \sigma^2 > 0; \theta = (\mu, \sigma^2) \in \mathbb{R} \times \mathbb{R}^+ \subset \mathbb{R}^2.$$

Here we are interested in estimating μ, the mean of the distribution, but σ^2, the variance of the distribution, is necessary to properly describe the possible probability distributions that might have generated the data. It is useful to write the parameter θ as $(\beta^T, \eta^T)^T$, where $\beta^{q\times 1}$ (a q-dimensional vector) is the parameter of interest and $\eta^{r\times 1}$ (an r-dimensional vector) is the nuisance parameter. In the previous example, $\beta = \mu$ and $\eta = \sigma^2$. The entire parameter space Ω has dimension $p = q + r$.

In some cases, we may want to consider models where the class of densities is so large that the parameter θ is infinite-dimensional. Examples of this will be given shortly. For such models, we will consider the problem where we are interested in estimating β, which we still take to be finite-dimensional, say q-dimensional. For some problems, it will be natural to partition the parameter θ as (β, η), where β is the q-dimensional parameter of interest and η is the nuisance parameter, which is infinite-dimensional. In other cases, it is more natural to consider the parameter β as the function $\beta(\theta)$. These models are referred to as *semiparametric models* in the literature because, generally, there is both a parametric component β and a nonparametric component η that describe the model. By allowing the space of parameters to be infinite-dimensional, we are putting less restrictions on the probabilistic constraints that our data might have (compared with finite-dimensional parametric models). Therefore, solutions, if they exist and are reasonable, will have greater applicability and robustness.

Because the notion of an infinite-dimensional parameter space is so important in the subsequent development of this book, we start with a short discussion of infinite-dimensional spaces.

1.1 What Is an Infinite-Dimensional Space?

The parameter spaces that we will consider in this book will always be subsets of linear vector spaces. That is, we will consider a parameter space $\Omega \subset \mathcal{S}$, where $\mathcal{S}$ is a linear space. A space $\mathcal{S}$ is a linear space if, for θ_1 and θ_2 elements of $\mathcal{S}$, $a\theta_1 + b\theta_2$ will also be an element of $\mathcal{S}$ for any scalar constants a and b. Such a linear space is finite-dimensional if it can be spanned by a finite number of elements in the space. That is, $\mathcal{S}$ is a finite-dimensional linear space if elements $\theta_1, \ldots, \theta_m$ exist, where m is some finite positive integer such that any element $\theta \in \mathcal{S}$ is equal to some linear combination of $\theta_1, \ldots, \theta_m$; i.e., $\theta = a_1\theta_1 + \ldots + a_m\theta_m$ for some scalar constants $a_1, \ldots, a_m$. The dimension of a finite-dimensional linear space is defined by the minimum number of elements in the space that span the entire space or, equivalently, the number of linearly independent elements that span the entire space, where a set of elements are linearly independent if no element in the set can be written as a linear combination of the other elements. Parameter spaces that are defined in p-dimensional Euclidean spaces are clearly finite-dimensional spaces. A linear

space $\mathcal{S}$ that cannot be spanned by any finite set of elements is called an infinite-dimensional parameter space.

An example of an infinite-dimensional linear space is the space of continuous functions defined on the real line. Consider the space $\mathcal{S} = \{f(x), x \in \mathbb{R}\}$ for all continuous functions $f(\cdot)$. Clearly $\mathcal{S}$ is a linear space. In order to show that this space is infinite-dimensional, we must demonstrate that it cannot be spanned by any finite set of elements in $\mathcal{S}$. This can be accomplished by noting that the space $\mathcal{S}$ contains the linear subspaces made up of the class of polynomials of order m; that is, the space $\mathcal{S}_m = \{f(x) = \sum_{j=0}^{m} a_j x^j\}$ for all constants $a_0, \ldots, a_m$. Clearly, the space $\mathcal{S}_m$ is finite-dimensional (i.e., spanned by the elements $x^0, x^1, \ldots, x^m$). In fact, this space is exactly an $m+1$-dimensional linear space since the elements $x^0, \ldots, x^m$ are linearly independent.

Linear independence follows because x^j cannot be written as a linear combination of $x^0, \ldots, x^{j-1}$ for any $j = 1, 2, \ldots$. If it could, then

$$x^j = \sum_{\ell=0}^{j-1} a_\ell x^\ell \text{ for all } x \in \mathbb{R}$$

for some constants $a_0, \ldots, a_{j-1}$. If this were the case, then the derivatives of x^j of all orders would have to be equal to the corresponding derivatives of $\sum_{\ell=0}^{j-1} a_\ell x^\ell$. But the j-th derivative of x^j is equal to $j! \neq 0$, whereas the j-th derivative of $\sum_{\ell=0}^{j-1} a_\ell x^\ell$ is zero, leading to a contradiction and implying that $x^0, \ldots, x^m$ are linearly independent.

Consequently, the space $\mathcal{S}$ cannot be spanned by any finite number, say m elements of $\mathcal{S}$, because, if this were possible, then the space of polynomials of order greater than m could also be spanned by the m elements. But this is impossible since such spaces of polynomials have dimension greater than m. Hence, $\mathcal{S}$ is infinite-dimensional.

From the arguments above, we can easily show that the space of arbitrary densities $p_Z(z)$ for a continuous random variable Z defined on the closed finite interval $[0, 1]$ (i.e., the so-called nonparametric model for such a random variable) spans a space that is infinite-dimensional. This follows by noticing that the functions $p_{Z_j}(z) = (j+1)^{-1} z^j,\ 0 \leq z \leq 1,\ j = 1, \ldots$ are densities that are linearly independent.

1.2 Examples of Semiparametric Models

Example 1: Restricted Moment Models

A common statistical problem is to model the relationship of a response variable Y (possibly vector-valued) as a function of a vector of covariates X. Throughout, we will use the convention that a vector of random variables Z that is not indexed will correspond to a single observation, whereas $Z_i, i = 1, \ldots, n$ will denote a sample of n iid random vectors. Consider a

family of probability distributions for $Z = (Y, X)$ that satisfy the regression relationship

$$E(Y|X) = \mu(X, \beta),$$

where $\mu(X, \beta)$ is a known function of X and the unknown q-dimensional parameter β.

The function $\mu(X, \beta)$ may be linear or nonlinear in β, and it is assumed that β is finite-dimensional. For example, we might consider a linear model where $\mu(X, \beta) = \beta^T X$ or a nonlinear model, such as a log-linear model, where $\mu(X, \beta) = \exp(\beta^T X)$. No other assumptions will be made on the class of probability distributions other than the constraint given by the conditional expectation of Y given X stated above. As we will demonstrate shortly, such models are semiparametric, as they will be defined through an infinite-dimensional parameter space. We will refer to such semiparametric models as "restricted moment models." These models were studied by Chamberlain (1987) and Newey (1988) in the econometrics literature. They were also popularized in the statistics literature by Liang and Zeger (1986).

For illustration, we will take Y to be a one-dimensional random variable that is continuous on the real line. This model can also be written as

$$Y = \mu(X, \beta) + \varepsilon,$$

where $E(\varepsilon|X) = 0$. The data are realizations of $(Y_1, X_1), \ldots, (Y_n, X_n)$ that are iid with density for a single observation given by

$$p_{Y,X}\{y, x; \beta, \eta(\cdot)\},$$

where $\eta(\cdot)$ denotes the infinite-dimensional nuisance parameter function characterizing the joint distribution of ε and X, to be defined shortly. Knowledge of β and the joint distribution of (ε, X) will induce the joint distribution of (Y, X). Since

$$\varepsilon = Y - \mu(X, \beta),$$
$$p_{Y,X}(y, x) = p_{\varepsilon,X}\{y - \mu(x, \beta), x\}.$$

The restricted moment model only makes the assumption that

$$E(\varepsilon|X) = 0.$$

That is, we will allow any joint density $p_{\varepsilon,X}(\varepsilon, x) = p_{\varepsilon|X}(\varepsilon|x)p_X(x)$ such that

$$p_{\varepsilon|X}(\varepsilon|x) \geqslant 0 \quad \text{for all } \varepsilon, x,$$
$$\int p_{\varepsilon|X}(\varepsilon|x)d\varepsilon = 1 \quad \text{for all } x,$$
$$\int \varepsilon p_{\varepsilon|X}(\varepsilon|x)d\varepsilon = 0 \quad \text{for all } x,$$
$$p_X(x) \geq 0 \quad \text{for all } x,$$
$$\int p_X(x)d\nu_X(x) = 1.$$

Remark 1. When we refer to the density, joint density, or conditional density of one or more random variables, to avoid confusion, we will often index the variables being used as part of the notation. For example, $p_{Y,X}(y,x)$ is the joint density of the random variables Y and X evaluated at the values (y,x). This notation will be suppressed when the variables are obvious. □

Remark 2. We will use the convention that random variables are denoted by capital letters such as Y and X, whereas realizations of those random variables will be denoted by lowercase letters such as y and x. One exception to this is that the random variable corresponding to the error term $Y - \mu(X,\beta)$ is denoted by the Greek lowercase ε. This is in keeping with the usual notation for such error terms used in statistics. The realization of this error term will also be denoted by the Greek lowercase ε. The distinction between the random variable and the realization of the error term will have to be made in the context it is used and should be obvious in most cases. For example, when we refer to $p_{\varepsilon,X}(\varepsilon,x)$, the subscript ε is a random variable and the argument ε inside the parentheses is the realization. □

Remark 3. $\nu_X(x)$ is a dominating measure for which densities for the random vector X are defined. For the most part, we will consider $\nu(\cdot)$ to be the Lebesgue measure for continuous random variables and the counting measure for discrete random variables. The random variable Y and hence ε will be taken to be continuous random variables dominated by Lebesgue measure dy or $d\varepsilon$, respectively. □

Without going into the measure-theoretical technical details, the class of conditional densities for ε given X, such that $E(\varepsilon|X) = 0$, can be constructed through the following steps.

(a) Choose any arbitrary positive function of ε and x (subject to regularity conditions):
$$h^{(0)}(\varepsilon,x) > 0.$$

(b) Normalize this function to be a conditional density:
$$h^{(1)}(\varepsilon,x) = \frac{h^{(0)}(\varepsilon,x)}{\int h^{(0)}(\varepsilon,x)d\varepsilon};$$
i.e.,
$$\int h^{(1)}(\varepsilon,x)d\varepsilon = 1 \text{ for all } x.$$

(c) Center it:
A random variable ε^* whose conditional density, given $X = x$ is $h^{(1)}(\varepsilon',x) = p(\varepsilon^* = \varepsilon'|X = x)$, has mean

$$\mu(x) = \int \varepsilon' h^{(1)}(\varepsilon', x) d\varepsilon'.$$

In order to construct a random variable ε whose conditional density, given $X = x$, has mean zero, we consider $\varepsilon = \varepsilon^* - \mu(X)$ or $\varepsilon^* = \varepsilon + \mu(X)$. It is clear that $E(\varepsilon|X = x) = E(\varepsilon^*|X = x) - \mu(x) = 0$. Since the transformation from ε to ε^*, given $X = x$, has Jacobian equal to 1, the conditional density of ε given X, defined by $\eta_1(\varepsilon, x)$, is given by

$$\eta_1(\varepsilon, x) = h^{(1)}\left\{\varepsilon + \int \varepsilon h^{(1)}(\varepsilon, x) d\varepsilon, x\right\},$$

which, by construction, satisfies $\int \varepsilon \eta_1(\varepsilon, x) d\varepsilon = 0$ for all x.

Thus, any function $\eta_1(\varepsilon, x)$ constructed as above will satisfy $\eta_1(\varepsilon, x) > 0$,

$$\int \eta_1(\varepsilon, x) d\varepsilon = 1 \quad \text{for all } x,$$

$$\int \varepsilon \eta_1(\varepsilon, x) d\varepsilon = 0 \quad \text{for all } x.$$

Since the class of all such conditional densities $\eta_1(\varepsilon, x)$ was derived from arbitrary positive functions $h^{(0)}(\varepsilon, x)$ (subject to regularity conditions), and since the space of positive functions is infinite-dimensional, then the set of such resulting conditional densities is also infinite-dimensional.

Similarly, we can construct densities for X where $p_X(x) = \eta_2(x)$ such that

$$\eta_2(x) > 0,$$

$$\int \eta_2(x) d\nu_X(x) = 1.$$

The set of all such functions $\eta_2(x)$ will also be infinite-dimensional as long as the support of X is infinite.

Therefore, the restricted moment model is characterized by

$$\{\beta, \eta_1(\varepsilon, x), \eta_2(x)\},$$

where $\beta \in \mathbb{R}^q$ is finite-dimensional and $\eta_1(\varepsilon, x) = p_{\varepsilon|X}(\varepsilon|x)$, $\eta_2(x) = p_X(x)$ are infinite-dimensional. Consequently, the joint density of (Y, X) is given by

$$\begin{aligned} p_{Y,X}\{y, x; \beta, \eta_1(\cdot), \eta_2(\cdot)\} &= p_{Y|X}\{y|x; \beta, \eta_1(\cdot)\} p_X\{x; \eta_2(\cdot)\} \\ &= \eta_1\{y - \mu(x, \beta), x\} \eta_2(x). \end{aligned}$$

This is an example of a semiparametric model because the parametrization is through a finite-dimensional parameter of interest $\beta \in \mathbb{R}^q$ and infinite-dimensional nuisance parameters $\{\eta_1(\cdot), \eta_2(\cdot)\}$.

Contrast this semiparametric model with the more traditional parametric model where

$$Y_i = \mu(X_i, \beta) + \varepsilon_i, i = 1, \ldots, n,$$

where ε_i are iid $N(0, \sigma^2)$. That is,

$$p_{Y|X}(y|x; \beta, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{\frac{1}{2}}} \exp\left[-\frac{1}{2}\frac{\{y - \mu(x, \beta)\}^2}{\sigma^2}\right].$$

This model is much more restrictive than the semiparametric model defined earlier.

Example 2: Proportional Hazards Model

In many biomedical applications, we are often interested in modeling the survival time of individuals as functions of covariates. Let the response variable be the survival time of an individual, denoted by T, whose distribution depends on explanatory variables X. A popular model in survival analysis is Cox's proportional hazards model, which was first introduced in the seminal paper by Cox (1972). This model assumes that the conditional hazard rate, as a function of X, is given by

$$\begin{aligned}\lambda(t|X) &= \lim_{h \to 0}\left\{\frac{P(t \leq T < t + h | T \geqslant t, X)}{h}\right\} \\ &= \lambda(t)\exp(\beta^T X).\end{aligned}$$

The proportional hazards model is especially convenient when survival times may be right censored, as we will discuss in greater detail in Chapter 5.

Interest often focuses on the finite-dimensional parameter β, as this describes the magnitude of the effect that the covariates have on the survival time. The underlying hazard function $\lambda(t)$ is left unspecified and is considered a nuisance parameter. Since this function can be any positive function in t, subject to some regularity conditions, it, too, is infinite-dimensional. Using the fact that the density of a positive random variable is related to the hazard function through

$$p_T(t) = \lambda(t)\exp\left\{-\int_0^t \lambda(u)du\right\},$$

then the density of a single observation $Z = (T, X)$ is given by

$$p_{T,X}\{t, x; \beta, \lambda(\cdot), \eta_2(\cdot)\} = p_{T|X}\{t|x; \beta, \lambda(\cdot)\}\eta_2(x),$$

where

$$p_{T|X}\{t|x; \beta, \lambda(\cdot)\} = \lambda(t)\exp(\beta^T x)\exp\left\{-\exp(\beta^T x)\int_0^t \lambda(u)du\right\},$$

and exactly as in Example 1, $\eta_2(x)$ is defined as a function of x such that

$$\eta_2(x) \geqslant 0,$$
$$\int \eta_2(x) d\nu_X(x) = 1,$$

for all x. The proportional hazards model has gained a great deal of popularity because it is more flexible than a finite-dimensional parametric model, that assumes that the hazard function for T has a specific functional form in terms of a few parameters; e.g.,

$$\lambda(t, \eta) = \eta \text{ (constant hazard over time – exponential model),}$$

or

$$\lambda(t, \eta) = \eta_1 t^{\eta_2} \text{ (Weibull model).}$$

Example 3: Nonparametric Model

In the two previous examples, the probability models were written in terms of an infinite-dimensional parameter θ, which was partitioned as $\{\beta^T, \eta(\cdot)\}$, where β was the finite-dimensional parameter of interest and $\eta(\cdot)$ was the infinite-dimensional nuisance parameter. We now consider the problem of estimating the moments of a single random variable Z where we put no restriction on the distribution of Z except that the moments of interest exist. That is, we denote the density of Z by $\theta(z)$, where $\theta(z)$ can be any positive function of z such that $\int \theta(z) d\nu_Z(z) = 1$ and any additional restrictions necessary for the moments of interest to exist. Clearly, the class of all $\theta(\cdot)$ is infinite-dimensional as long as the support of Z is infinite. Suppose we were interested in estimating some functional of $\theta(\cdot)$, say $\beta(\theta)$ (for example, the first or second moment $E(Z)$ or $E(Z^2)$, where $\beta(\theta)$ is equal to $\int z\theta(z) d\nu_Z(z)$ or $\int z^2\theta(z) d\nu_Z(z)$, respectively). For such a problem, it is not convenient to try to partition the parameter space in terms of the parameter β of interest and a nuisance parameter but rather to work directly with the functional $\beta(\theta)$.

1.3 Semiparametric Estimators

In a semiparametric model, a semiparametric estimator for β, say $\hat{\beta}_n$, is one that, loosely speaking, has the property that it is consistent and asymptotically normal in the sense that

$$(\hat{\beta}_n - \beta) \xrightarrow{P\{\beta, \eta(\cdot)\}} 0,$$
$$n^{1/2}(\hat{\beta}_n - \beta) \xrightarrow{\mathcal{D}\{\beta, \eta(\cdot)\}} N(0, \Sigma^{q \times q}\{\beta, \eta(\cdot)\}),$$

for all densities "$p\{z, \beta, \eta(\cdot)\}$" within some semiparametric family, where $\xrightarrow{P\{\beta,\eta(\cdot)\}}$ denotes convergence in probability and $\xrightarrow{\mathcal{D}\{\beta,\eta(\cdot)\}}$ denotes convergence in distribution when the density of the random variable Z is $p\{z, \beta, \eta(\cdot)\}$.

We know, for example, that the solution to the linear estimating equations

$$\sum_{i=1}^{n} A^{q\times 1}(X_i, \hat{\beta}_n)\left\{Y_i - \mu(X_i, \hat{\beta}_n)\right\} = 0^{q\times 1},$$

under suitable regularity conditions, leads to an estimator for β that is consistent and asymptotically normal for the semiparametric restricted moment model of Example 1. In fact, this is the basis for "generalized estimating equations" (GEE) proposed by Liang and Zeger (1986).

The maximum partial likelihood estimator proposed by Cox (1972, 1975) is an example of a semiparametric estimator for β in the proportional hazards model given in Example 2. Also, in Example 3, a semiparametric estimator for the first and second moments is given by $n^{-1}\sum Z_i$ and $n^{-1}\sum Z_i^2$, respectively.

Some issues that arise when studying semiparametric models are:

(i) How do we find semiparametric estimators, or do they even exist?
(ii) How do we find the best estimator among the class of semiparametric estimators?

Both of these problems are difficult. Understanding the geometry of estimators, more specifically the geometry of the *influence function* of estimators, will help us in this regard.

Much of this book will rely heavily on geometric constructions. We will define the influence function of an asymptotically linear estimator and describe the geometry of all possible influence functions for a statistical model. We will start by looking at finite-dimensional parametric models and then generalize the results to the more complicated infinite-dimensional semiparametric models.

Since the geometry that is considered is the geometry of Hilbert spaces, we begin with a quick review of Hilbert spaces, the notion of orthogonality, minimum distance, and how this relates to efficient estimators (i.e., estimators with the smallest asymptotic variance).

2

Hilbert Space for Random Vectors

In this section, we will introduce a Hilbert space without going into much of the technical details. We will focus primarily on the Hilbert space whose elements are random vectors with mean zero and finite variance that will be used throughout the book. For more details about Hilbert spaces, we recommend that the reader study Chapter 3 of Luenberger (1969).

2.1 The Space of Mean-Zero q-dimensional Random Functions

As stated earlier, data are envisioned as realizations of the random vectors $Z_1, Z_2, \ldots, Z_n$, assumed iid. Let Z denote the random vector for a single observation. As always, there is an underlying probability space $(\mathscr{Z}, \mathcal{A}, P)$, where $\mathscr{Z}$ denotes the sample space, $\mathcal{A}$ the corresponding σ-algebra, and P the probability measure. For the time being, we will not consider a statistical model consisting of a family of probability measures, but rather we will assume that P is the true probability measure that generates the realizations of Z.

Consider the space consisting of q-dimensional mean-zero random functions of Z,

$$h : \mathscr{Z} \to \mathbb{R}^q,$$

where $h(Z)$ is measurable and also satisfies

$$\text{(i) } E\{h(Z)\} = 0,$$
$$\text{(ii) } E\{h^T(Z)h(Z)\} < \infty.$$

Since the elements of this space are random functions, when we refer to an element as h, we implicitly mean $h(Z)$. Clearly, the space of all such h that satisfy (i) and (ii) is a linear space. By linear, we mean that if h_1, h_2 are elements of the space, then for any real constants a and b, $ah_1 + bh_2$ also belongs to the space.

In the same way that we consider points in Euclidean space as vectors from the origin, here we will consider the q-dimensional random functions as points in a space. The intuition we have developed in understanding the geometry of two- and three-dimensional Euclidean space will aid us in understanding the geometry of more complex spaces through analogy. The random function

$$h(Z) = 0^{q\times 1}$$

will denote the origin of this space.

The Dimension of the Space of Mean-Zero Random Functions

An element of the linear space defined above is a q-dimensional function of Z. This should not be confused with the dimensionality of the space itself. To illustrate this point more clearly, let us first consider the space of one-dimensional random functions of Z (random variables), where Z is a discrete variable with finite support. Specifically, let Z be allowed to take on one of a finite number of values $z_1, \ldots, z_k$ with positive probabilities $\pi_1, \ldots, \pi_k$, where $\sum_{i=1}^{k} \pi_i = 1$. For such a case, any one-dimensional random function of Z can be defined as $h(Z) = a_1 I(Z = z_1) + \ldots + a_k I(Z = z_k)$ for any real valued constants $a_1, \ldots, a_k$, where $I(\cdot)$ denotes the indicator function. The space of all such random functions is a linear space spanned by the k linearly independent functions $I(Z = z_i), i = 1, \ldots, k$. Hence this space is a k-dimensional linear space. If we put the further constraint that the mean must be zero (i.e., $E\{h(Z)\} = 0$), then this implies that $\sum_{i=1}^{k} a_i \pi_i = 0$, or equivalently that $a_k = -(\sum_{i=1}^{k-1} a_i \pi_i)/\pi_k$. Some simple algebra leads us to conclude that the space of one-dimensional mean-zero random functions of Z is a linear space spanned by the $k-1$ linearly independent functions $\{I(Z = z_i) - \frac{\pi_i}{\pi_k} I(Z = z_k)\}, i = 1, \ldots, k-1$. Hence this space is a $k-1$-dimensional linear space.

Similarly, the space of q-dimensional mean-zero random functions of Z, where Z has finite support at the k values $z_1, \ldots, z_k$, can be shown to be a linear space with dimension $q \times (k-1)$. Clearly, as the number of support points k for the distribution of Z increases, so does the dimension of the linear space of q-dimensional mean-zero random functions of Z.

If the support of the random vector Z is infinite, as would be the case if any element of the random vector Z was a continuous random variable, then the space of measurable functions that make up the Hilbert space will be infinite-dimensional. As we indicated in Section 1.1, the set of one-dimensional continuous functions of Z is infinite-dimensional. Consequently, the set of q-dimensional continuous functions will also be infinite-dimensional. Clearly, the set of q-dimensional measurable functions is a larger class and hence must also be infinite-dimensional.

2.2 Hilbert Space

A Hilbert space, denoted by $\mathcal{H}$, is a complete normed linear vector space equipped with an inner product. As well as being a linear space, a Hilbert space also allows us to consider distance between elements and angles and orthogonality between vectors in the space. This is accomplished by defining an inner product.

Definition 1. Corresponding to each pair of elements h_1, h_2 belonging to a linear vector space $\mathcal{H}$, an inner product, defined by $\langle h_1, h_2 \rangle$, is a function that maps to the real line. That is, $\langle h_1, h_2 \rangle$ is a scalar that satisfies

1. $\langle h_1, h_2 \rangle = \langle h_2, h_1 \rangle$,
2. $\langle h_1 + h_2, h_3 \rangle = \langle h_1, h_3 \rangle + \langle h_2, h_3 \rangle$, where h_1, h_2, h_3 belong to $\mathcal{H}$,
3. $\langle \lambda h_1, h_2 \rangle = \lambda \langle h_1, h_2 \rangle$ for any scalar constant λ,
4. $\langle h_1, h_1 \rangle \geq 0$ with equality if and only if $h_1 = 0$.

Note 1. In some cases, the function $\langle \cdot, \cdot \rangle$ may satisfy conditions 1–3 above and the first part of condition 4, but $\langle h_1, h_1 \rangle = 0$ may not imply that $h_1 = 0$. In that case, we can still define a Hilbert space by identifying equivalence classes where individual elements in our space correspond to different equivalence classes.

Definition 2. For the linear vector space of q-dimensional measurable random functions with mean zero and finite second moments, we can define the inner product

$$\langle h_1, h_2 \rangle \quad \text{by} \quad E(h_1^T h_2).$$

We shall refer to this inner product as the "covariance inner product."

This definition of inner product clearly satisfies the first three conditions of the definition given above. As for condition 4, we can define an equivalence class where h_1 is equivalent to h_2,

$$h_1 \equiv h_2,$$

if $h_1 = h_2$ a.e. or $P(h_1 \neq h_2) = 0$. In this book, we will generally not concern ourselves with such measure-theoretical subtleties.

Once an inner product is defined, we then define the norm or "length" of any vector (i.e., element of $\mathcal{H}$) (distance from any point $h \in \mathcal{H}$ to the origin) as

$$\|h\| = \langle h, h \rangle^{1/2}.$$

Hilbert spaces also allow us to define orthogonality; that is, $h_1, h_2 \in \mathcal{H}$ are orthogonal if $\langle h_1, h_2 \rangle = 0$.

Remark 1. Technically speaking, the definitions above are those for a pre-Hilbert space. In order to be a Hilbert space, we also need the space to be complete (i.e., every Cauchy sequence has a limit point that belongs to the space). That the space of q-dimensional random functions with mean zero and bounded second moments is complete follows from the L_2-completeness theorem (see Loève 1963, p. 161) and hence is a Hilbert space. □

2.3 Linear Subspace of a Hilbert Space and the Projection Theorem

A space $\mathcal{U} \subset \mathcal{H}$ is a linear subspace if $u_1, u_2 \in \mathcal{U}$ implies that $au_1 + bu_2 \in \mathcal{U}$ for all scalar constants a, b. A linear subspace must contain the origin. This is clear by letting the scalars be $a = b = 0$.

A simple example of a linear subspace is obtained by taking $h_1, \ldots, h_k$ to be arbitrary elements of a Hilbert space. Then the space $a_1h_1 + \cdots + a_kh_k$ for all scalars $(a_1, \ldots, a_k) \in \mathbb{R}^k$ is a linear subspace spanned by $\{h_1, \ldots, h_k\}$.

One of the key results for Hilbert spaces, which we will use repeatedly throughout this book, is given by the projection theorem.

Projection Theorem for Hilbert Spaces

Theorem 2.1. Let $\mathcal{H}$ be a Hilbert space and $\mathcal{U}$ a linear subspace that is closed (i.e., contains all its limit points). Corresponding to any $h \in \mathcal{H}$, there exists a unique $u_0 \in \mathcal{U}$ that is closest to h; that is,

$$\|h - u_0\| \leq \|h - u\| \quad \text{for all} \quad u \in \mathcal{U}.$$

Furthermore, $h - u_0$ is orthogonal to $\mathcal{U}$; that is,

$$\langle h - u_0, u \rangle = 0 \quad \text{for all} \quad u \in \mathcal{U}.$$

We refer to u_0 as the projection of h onto the space $\mathcal{U}$, and this is denoted as $\Pi(h|\mathcal{U})$. Moreover, u_0 is the only element $u \in \mathcal{U}$ such that $h - u$ is orthogonal to $\mathcal{U}$ (see Figure 2.1).

The proof of the projection theorem for arbitrary Hilbert spaces is not much different or more difficult than for a finite-dimensional Euclidean space. The condition that a Hilbert space be complete is necessary to guarantee the existence of the projection. A formal proof can be found in Luenberger (1969, Theorem 2, p. 51). The intuition of orthogonality and distance carries over very nicely from simple Euclidean spaces to more complex Hilbert spaces.

A simple consequence of orthogonality is the Pythagorean theorem, which we state for completeness.

Theorem 2.2. *Pythagorean Theorem*
If h_1 and h_2 are orthogonal elements of the Hilbert space $\mathcal{H}$ (i.e., $\langle h_1, h_2 \rangle = 0$), then

$$\|h_1 + h_2\|^2 = \|h_1\|^2 + \|h_2\|^2.$$

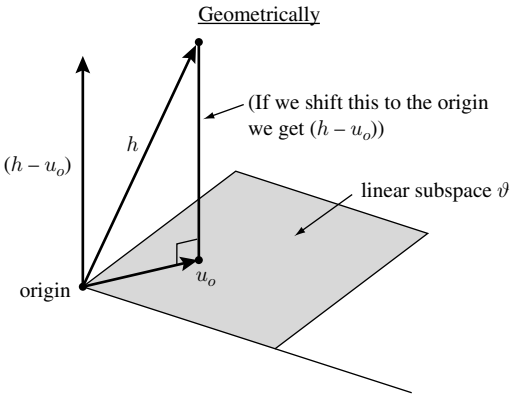

Fig. 2.1. Projection onto a linear subspace

2.4 Some Simple Examples of the Application of the Projection Theorem

Example 1: One-Dimensional Random Functions

Consider the Hilbert space $\mathcal{H}$ of one-dimensional random functions, $h(Z)$, with mean zero and finite variance equipped with the inner product

$$\langle h_1, h_2 \rangle = E(h_1 h_2)$$

for $h_1(Z), h_2(Z) \in \mathcal{H}$. Let $u_1(Z), \ldots, u_k(Z)$ be arbitrary elements of this space and $\mathcal{U}$ be the linear subspace spanned by $\{u_1, \cdots, u_k\}$. That is,

$$\mathcal{U} = \{a^T u; \quad \text{for} \quad a \in \mathbb{R}^k\},$$

where

$$u^{k\times 1} = \begin{pmatrix} u_1 \\ \vdots \\ u_k \end{pmatrix}.$$

The space $\mathcal{U}$ is an example of a finite-dimensional linear subspace since it is spanned by the finite number of elements $u_1(Z), \ldots, u_k(Z)$. This subspace is contained in the infinite-dimensional Hilbert space $\mathcal{H}$. Moreover, if the elements $u_1(Z), \ldots, u_k(Z)$ are linearly independent, then the dimension of $\mathcal{U}$ is identically equal to k.

Let h be an arbitrary element of $\mathcal{H}$. Then the projection of h onto the linear subspace $\mathcal{U}$ is given by the unique element $a_0^T u$ that satisfies

$$\langle h - a_0^T u, a^T u \rangle = 0 \quad \text{for all} \quad a = (a_1, \ldots, a_k)^T \in \mathbb{R}^k,$$

or

$$\sum_{j=1}^{k} a_j \langle h - a_0^T u, u_j \rangle = 0 \quad \text{for all} \;\; a_j, \;\; j = 1, \ldots, k.$$

Equivalently, $\langle h - a_0^T u, u_j \rangle = 0 \quad$ for all $j = 1, \ldots, k$,

or

$$E\{(h - a_0^T u)u^T\} = 0^{(1\times k)},$$

or

$$E(hu^T) - a_0^T E(uu^T) = 0^{(1\times k)}.$$

Any solution of a_0 such that

$$a_0^T E(uu^T) = E(hu^T)$$

would lead to the unique projection $a_0^T u$.

If $E(uu^T)$ is positive definite, and therefore has a unique inverse, then

$$a_0^T = E(hu^T)\{E(uu^T)\}^{-1},$$

in which case the unique projection will be

$$u_0 = a_0^T u = E(hu^T)\{E(uu^T)\}^{-1}u.$$

The norm-squared of this projection is equal to

$$E(hu^T)\{E(uu^T)\}^{-1}E(uh).$$

By the Pythagorean theorem,

$$\begin{aligned} \|h - a_0^T u\|^2 &= E(h - a_0^T u)^2 \\ &= E(h^2) - E(hu^T)\{E(uu^T)\}^{-1}E(uh). \end{aligned}$$

Example 2: *q*-dimensional Random Functions

Let $\mathcal{H}$ be the Hilbert space of mean-zero q-dimensional measurable random functions with finite second moments equipped with the inner product

$$\langle h_1, h_2 \rangle = E(h_1^T h_2).$$

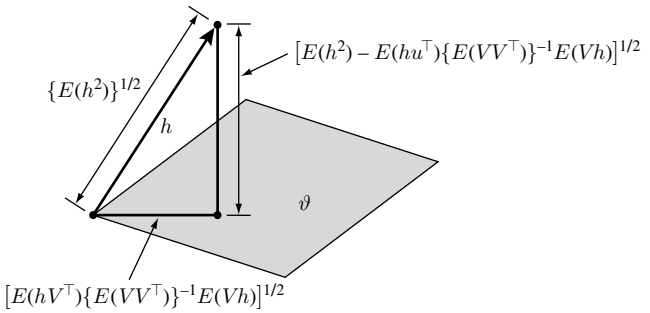

Fig. 2.2. Geometric illustration of the Pythagorean theorem

Let "$v(Z)$" be an r-dimensional random function with mean zero and $E(v^T v) < \infty$. Consider the linear subspace $\mathcal{U}$ spanned by $v(Z)$; that is,

$$\mathcal{U} = \{B^{q\times r}v, \text{ where } B \text{ is any arbitrary } q \times r \text{ matrix of real numbers}\}.$$

The linear subspace $\mathcal{U}$ defined above is a finite-dimensional linear subspace contained in the infinite-dimensional Hilbert space $\mathcal{H}$. If the elements $v_1(Z), \ldots, v_r(Z)$ are linearly independent, then the dimension of $\mathcal{U}$ is $q \times r$. This can easily be seen by noting that $\mathcal{U}$ is spanned by the $q \times r$ linearly independent elements $u_{ij}(Z), i = 1, \ldots, q, j = 1, \ldots, r$, of $\mathcal{H}$, where, for any $i = 1, \ldots, q$, we take the element $u_{ij}^{q\times 1}(Z) \in \mathcal{H}$ to be the q-dimensional function of Z, where all except the i-th element are equal to 0 and the i-th element is equal to $v_j(Z)$ for $j = 1, \ldots, r$.

We now consider the problem of finding the projection of an arbitrary element $h \in \mathcal{H}$ onto $\mathcal{U}$. By the projection theorem, such a projection $B_0 v$ is unique and must satisfy

$$E\{(h - B_0 v)^T Bv\} = 0 \quad \text{for all } B \in \mathbb{R}^{q\times r}. \tag{2.1}$$

The statement above being true for all B is equivalent to

$$E\{(h - B_0 v)v^T\} = 0^{q\times r} \quad \text{(matrix of all zeros)}. \tag{2.2}$$

To establish (2.2), we write

$$E\{(h - B_0 v)^T Bv\} = \sum_i \sum_j B_{ij} E\{(h - B_0 v)_i v_j\}, \tag{2.3}$$

where $(h - B_0 v)_i$ denotes the i-th element of the q-dimensional vector $(h - B_0 v)$, v_j denotes the j-th element of the r-dimensional vector v, and B_{ij} denotes the (i, j)-th element of the matrix B.

If we take $B_{ij} = 1$ for $i = i'$ and $j = j'$, and 0 otherwise, it becomes clear from (2.3) that

$$E\{(h - B_0 v)_{i'} v_{j'}\} = 0 \quad \text{for all} \quad i', j'$$

and hence

$$E\{(h - B_0 v)v^T\} = 0^{q \times r}.$$

Conversely, if (2.2) is true, then for any matrix B we have (2.1) being true. Consequently, by (2.2), we deduce that $E(hv^T) = B_0 E(vv^T)$. Therefore, assuming $E(vv^T)$ is nonsingular (i.e., positive definite),

$$B_0 = E(hv^T)\{E(vv^T)\}^{-1}.$$

Hence, the unique projection is

$$\Pi(h|\mathcal{U}) = E(hv^T)\{E(vv^T)\}^{-1}v. \tag{2.4}$$

Remark 2. Finding the projection of the q-dimensional vector h onto the subspace $\mathcal{U}$ is equivalent to taking each element of h and projecting it individually to the subspace spanned by $\{v_1, \ldots, v_r\}$ for the Hilbert space of one-dimensional random functions considered in Example 1 and then stacking these individual projections into a vector. This can be deduced by noting that minimizing $\|h - Bv\|^2$ corresponds to minimizing

$$\begin{aligned} & E\{(h - Bv)^T(h - Bv)\} \\ &= \sum_{i=1}^{q} E(h - Bv)_i^2 \\ &= \sum_{i=1}^{q} E\left(h_i - \sum_{j=1}^{r} B_{ij} v_j\right)^2. \end{aligned} \tag{2.5}$$

Since the B_{ij} are arbitrary, we can minimize the sum in (2.5) by minimizing each of the elements in the sum separately. □

The norm-squared of this projection is given by

$$E\left[v^T\{E(vv^T)\}^{-1}E(vh^T)E(hv^T)\{E(vv^T)\}^{-1}v\right],$$

and, by the Pythagorean theorem (see Figure 2.2 for illustration), the norm-squared of the residual $(h - B_0 v)$ is

$$\begin{aligned} & E(h^T h) - E\left[v^T\{E(vv^T)\}^{-1}E(vh^T)E(hv^T)\{E(vv^T)\}^{-1}v\right] \\ &\quad = \mathrm{tr}\{[hh^T] - E(hv^T)\{E(vv^T)\}^{-1}E(vh^T)\}. \end{aligned}$$

There are other properties of Hilbert spaces that will be used throughout the book. Rather than giving all the properties in this introductory chapter, we will instead define these as they are needed. There is, however, one very important result that we do wish to highlight. This is the Cauchy-Schwartz inequality given in Theorem 2.3.

Theorem 2.3. *Cauchy-Schwartz inequality*
For any two elements $h_1, h_2 \in \mathcal{H}$,

$$|\langle h_1, h_2 \rangle|^2 \leq \|h_1\|^2 \|h_2\|^2,$$

with equality holding if and only if h_1 and h_2 are linearly related; i.e., $h_1 = ch_2$ for some scalar constant c.

2.5 Exercises for Chapter 2

1. Prove the projection theorem (Theorem 2.1) for Hilbert spaces.
2. Let $Z = (Z_1, \ldots, Z_p)^T$ be a p-dimensional multivariate normally distributed random vector with mean zero and covariance matrix $\Sigma^{p\times p}$. We also write Z as the partitioned vector $Z = (Y_1^T, Y_2^T)^T$, where $Y_1^{q\times 1} = (Z_1, \ldots, Z_q)^T$ and $Y_2^{(p-q)\times 1} = (Z_{q+1}, \ldots, Z_p)^T, q < p$, and

$$\Sigma = \begin{pmatrix} \Sigma_{11}^{q\times q} & \Sigma_{12}^{q\times(p-q)} \\ \Sigma_{21}^{(p-q)\times q} & \Sigma_{22}^{(p-q)\times(p-q)} \end{pmatrix}, \text{where}$$

$$\Sigma_{11} = E(Y_1 Y_1^T),\ \Sigma_{12} = E(Y_1 Y_2^T),\ \Sigma_{21} = E(Y_2 Y_1^T),\ \Sigma_{22} = E(Y_2 Y_2^T).$$

Let $\mathcal{H}$ be the Hilbert space of all q-dimensional measurable functions of Z with mean zero, finite variance, and equipped with the covariance inner product. Let $\mathcal{U}$ be the linear subspace spanned by Y_2; i.e., $\mathcal{U}$ consists of all the elements

$$\left\{ B^{q\times(p-q)} Y_2 : \quad \text{for all} \quad q \times (p-q) \ \text{matrices}\, B \right\}.$$

(a) Find the projection of Y_1 onto $\mathcal{U}$.
(b) Compute the norm of the residual

$$\|Y_1 - \Pi(Y_1|\mathcal{U})\|.$$

3. Let $Z = (Z_1, Z_2)^T$ be a bivariate normally distributed vector with mean zero and covariance matrix $\begin{pmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{pmatrix}$.

Consider the Hilbert space of all one-dimensional measurable functions of Z with mean zero, finite variance, and covariance inner product. Let $\mathcal{U}$ denote the linear subspace spanned by Z_2 and $(Z_1^2 - \sigma_1^2)$; i.e., the space whose elements are

$$a_1(Z_1^2 - \sigma_1^2) + a_2 Z_2 \quad \text{for all} \quad a_1, a_2.$$

(a) Find the projection of Z_1 onto the space $\mathcal{U}$.
(b) Find the variance of the residual (i.e., $\text{var}\{Z_1 - \Pi(Z_1|\mathcal{U})\}$).

3

The Geometry of Influence Functions

As we will describe shortly, most reasonable estimators for the parameter β, in either parametric or semiparametric models, are asymptotically linear and can be uniquely characterized by the influence function of the estimator. The class of influence functions for such estimators belongs to the Hilbert space of all mean-zero q-dimensional random functions with finite variance that was defined in Chapter 2. As such, this construction will allow us to view estimators or, more specifically, the influence function of estimators, from a geometric point of view. This will give us intuitive insight into the construction of such estimators and a geometric way of assessing the relative efficiencies of the various estimators.

As always, consider the statistical model where $Z_1, \ldots, Z_n$ are iid random vectors and the density of a single Z is assumed to belong to the class $\{p_Z(z;\theta), \theta \epsilon \Omega\}$ with respect to some dominating measure ν_Z. The parameter θ can be written as $(\beta^T, \eta^T)^T$, where $\beta^{q\times 1}$ is the parameter of interest and η, the nuisance parameter, may be finite- or infinite-dimensional. The truth will be denoted by $\theta_0 = (\beta_0^T, \eta_0^T)^T$. For the remainder of this chapter, we will only consider parametric models where $\theta = (\beta^T, \eta^T)^T$ and the vector θ is p-dimensional, the parameter of interest β is q-dimensional, and the nuisance parameter η is r-dimensional, with $p = q + r$.

An estimator $\hat{\beta}_n$ of β is a q-dimensional measurable random function of $Z_1, \ldots, Z_n$. Most reasonable estimators for β are *asymptotically linear*; that is, there exists a random vector (i.e., a q-dimensional measurable random function) $\varphi^{q\times 1}(Z)$, such that $E\{\varphi(Z)\} = 0^{q\times 1}$,

$$n^{1/2}(\hat{\beta}_n - \beta_0) = n^{-1/2} \sum_{i=1}^{n} \varphi(Z_i) + o_p(1), \tag{3.1}$$

where $o_p(1)$ is a term that converges in probability to zero as n goes to infinity and $E(\varphi\varphi^T)$ is finite and nonsingular.

Remark 1. The function $\varphi(Z)$ is defined with respect to the true distribution $p(z, \theta_0)$ that generates the data. Consequently, we sometimes may write

$\varphi(Z, \theta)$ to emphasize that this random function will vary according to the value of θ in the model. Unless otherwise stated, it will be assumed that $\varphi(Z)$ is evaluated at the truth and expectations are taken with respect to the truth. Therefore, $E\{\varphi(Z)\}$ is shorthand for

$$E_{\theta_0}\{\varphi(Z, \theta_0)\}.$$

□

The random vector $\varphi(Z_i)$ in (3.1) is referred to as the i-th influence function of the estimator $\hat{\beta}_n$ or the influence function of the i-th observation of the estimator $\hat{\beta}_n$. The term influence function comes from the robustness literature, where, to first order, $\varphi(Z_i)$ is the influence of the i-th observation on $\hat{\beta}_n$; see Hampel (1974).

Example 1. As a simple example, consider the model where $Z_1, \ldots, Z_n$ are iid $N(\mu, \sigma^2)$. The maximum likelihood estimators for μ and σ^2 are given by $\hat{\mu}_n = n^{-1}\sum_{i=1}^n Z_i$ and $\hat{\sigma}_n^2 = n^{-1}\sum_{i=1}^n (Z_i - \hat{\mu}_n)^2$, respectively. That the estimator $\hat{\mu}_n$ for μ is asymptotically linear follows immediately because

$$n^{1/2}(\hat{\mu}_n - \mu_0) = n^{-1/2}\sum_{i=1}^n (Z_i - \mu_0).$$

Therefore, $\hat{\mu}_n$ is an asymptotically linear estimator for μ whose i-th influence function is given by $\varphi(Z_i) = (Z_i - \mu_0)$.

After some straightforward algebra, we can express the estimator $\hat{\sigma}_n^2$ minus the estimand as

$$(\hat{\sigma}_n^2 - \sigma_0^2) = n^{-1}\sum_{i=1}^n \{(Z_i - \mu_0)^2 - \sigma_0^2\} + (\hat{\mu}_n - \mu_0)^2. \tag{3.2}$$

Multiplying (3.2) by $n^{1/2}$, we obtain

$$n^{1/2}(\hat{\sigma}_n^2 - \sigma_0^2) = n^{-1/2}\sum_{i=1}^n \{(Z_i - \mu_0)^2 - \sigma_0^2\} + n^{1/2}(\hat{\mu}_n - \mu_0)^2.$$

Since $n^{1/2}(\hat{\mu}_n - \mu_0)$ converges to a normal distribution and $(\hat{\mu}_n - \mu_0)$ converges in probability to zero, this implies that $n^{1/2}(\hat{\mu}_n - \mu_0)^2$ converges in probability to zero (i.e., is $o_p(1)$). Consequently, we have demonstrated that $\hat{\sigma}_n^2$ is an asymptotically linear estimator for σ^2 whose i-th influence function is given by $\varphi(Z_i) = \{(Z_i - \mu_0)^2 - \sigma_0^2\}$. □

When considering the asymptotic properties of an asymptotically linear estimator (e.g., asymptotic normality and asymptotic variance), it suffices to consider the influence function of the estimator. This follows as a simple consequence of the central limit theorem (CLT). Since, by definition,

$$n^{1/2}(\hat{\beta}_n - \beta_0) = n^{-1/2} \sum_{i=1}^{n} \varphi(Z_i) + o_p(1),$$

then, by the central limit theorem,

$$n^{-1/2} \sum_{i=1}^{n} \varphi(Z_i) \xrightarrow{\mathcal{D}} N\Big(0^{q\times 1}, E(\varphi\varphi^T)\Big),$$

and, by Slutsky's theorem,

$$n^{1/2}(\hat{\beta}_n - \beta_0) \xrightarrow{\mathcal{D}} N\Big(0, E(\varphi\varphi^T)\Big).$$

In an asymptotic sense, an asymptotically linear estimator can be identified through its influence function, as we now demonstrate in the following theorem.

Theorem 3.1. An asymptotically linear estimator has a unique (a.s.) influence function.

Proof. By contradiction
Suppose not. Then there exists another influence function $\varphi^*(Z)$ such that

$$E\{\varphi^*(Z)\} = 0,$$

and

$$n^{1/2}(\hat{\beta}_n - \beta_0) = n^{-1/2} \sum_{i=1}^{n} \varphi^*(Z_i) + o_p(1).$$

Since $n^{1/2}(\hat{\beta}_n - \beta_0)$ is also equal to $n^{-1/2} \sum_{i=1}^{n} \varphi(Z_i) + o_p(1)$, this implies that

$$n^{-1/2} \sum_{i=1}^{n} \{\varphi(Z_i) - \varphi^*(Z_i)\} = o_p(1).$$

However, by the CLT,

$$n^{-1/2} \sum_{i=1}^{n} \{\varphi(Z_i) - \varphi^*(Z_i)\} \xrightarrow{\mathcal{D}} N\Big(0, E\{(\varphi - \varphi^*)(\varphi - \varphi^*)^T\}\Big).$$

In order for this limiting normal distribution to be $o_p(1)$, we would require that the covariance matrix

$$E\{(\varphi - \varphi^*)(\varphi - \varphi^*)^T\} = 0^{q\times q},$$

which implies that $\varphi(Z) = \varphi^*(Z)$ a.s. □

The representation of estimators through their influence function lends itself nicely to geometric interpretations in terms of Hilbert spaces, discussed in Chapter 2. Before describing this geometry, we briefly comment on some regularity conditions that will be imposed on the class of estimators we will consider.

Reminder. We know that the variance of any unbiased estimator must be greater than or equal to the Cràmer-Rao lower bound; see, for example, Casella and Berger (2002, Section 7.3). When considering asymptotic theory, where we let the sample size n go to infinity, most reasonable estimators are asymptotically unbiased. Thus, we might expect the asymptotic variance of such asymptotically unbiased estimators also to be greater than the Cràmer-Rao lower bound. This indeed is the case for the most part, and estimators whose asymptotic variance equals the Cràmer-Rao lower bound are referred to as *asymptotically efficient*. For parametric models, with suitable regularity conditions, the maximum likelihood estimator (MLE) is an example of an efficient estimator. One of the peculiarities of asymptotic theory is that asymptotically unbiased estimators can be constructed that have asymptotic variance equal to the Cràmer-Rao lower bound for most of the parameter values in the model but have smaller variance than the Cràmer-Rao lower bound for the other parameters. Such estimators are referred to as *super-efficient* and for completeness we give the construction of such an estimator (Hodges) as an example.

3.1 Super-Efficiency

Example Due to Hodges

Let $Z_1, \ldots, Z_n$ be iid $N(\mu, 1), \mu \in \mathbb{R}$. For this simple model, we know that the maximum likelihood estimator (MLE) of μ is given by the sample mean $\bar{Z}_n = n^{-1} \sum_{i=1}^{n} Z_i$ and that

$$n^{1/2}(\bar{Z}_n - \mu) \xrightarrow{\mathcal{D}(\mu)} N(0, 1).$$

Now, consider the estimator $\hat{\mu}_n$ given by Hodges in 1951 (see LeCam, 1953):

$$\hat{\mu}_n = \begin{cases} \bar{Z}_n & \text{if} \quad |\bar{Z}_n| > n^{-1/4} \\ 0 & \text{if} \quad |\bar{Z}_n| \leq n^{-1/4}. \end{cases}$$

Some of the properties of this estimator are as follows.

If $\mu \neq 0$, then with increasing probability, the support of $\bar{Z}_n$ moves away from 0 (see Figure 3.1).

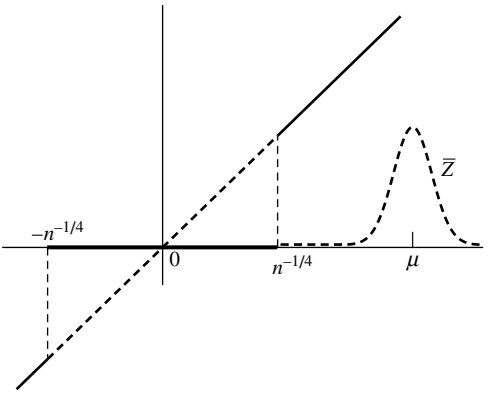

Fig. 3.1. When $\mu \neq 0$, $P_\mu(\bar{Z}_n \neq \hat{\mu}_n) \to 0$

Therefore $n^{1/2}(\bar{Z}_n - \mu) = n^{1/2}(\hat{\mu}_n - \mu) + o_p(1)$ and $n^{1/2}(\hat{\mu}_n - \mu) \xrightarrow{\mathcal{D}(\mu)} N(0,1)$.

If $\mu = 0$, then the support of $\bar{Z}_n$ will be concentrated in an $O(n^{-1/2})$ neighborhood about the origin and hence, with increasing probability, will be within $\pm n^{-1/4}$ (see Figure 3.2).

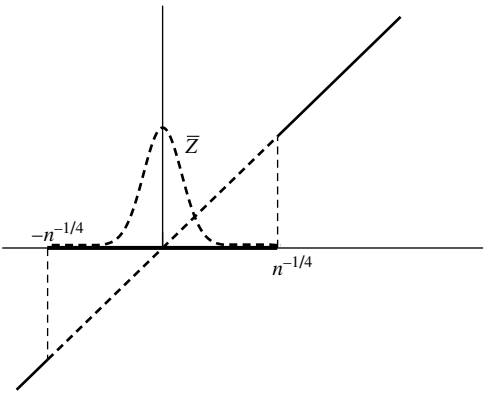

Fig. 3.2. When $\mu = 0$, $P_0\{|\bar{Z}_n| < n^{-1/4}\} \to 1$

Therefore, this implies that $P_0\left(\hat{\mu}_n = 0\right) \to 1$. Hence $P_0\left(n^{1/2}\hat{\mu}_n = 0\right) \to 1$, and $n^{1/2}\left(\hat{\mu}_n - 0\right) \xrightarrow{P_0} 0$ or $\xrightarrow{\mathcal{D}(0)} N(0,0)$. Consequently, the asymptotic variance of $n^{1/2}(\hat{\mu}_n - \mu)$ is equal to 1 for all $\mu \neq 0$, as it is for the MLE $\bar{Z}_n$, but for $\mu = 0$, the asymptotic variance of $n^{1/2}(\hat{\mu}_n - \mu)$ equals 0 and thus is super-efficient.

Although super-efficiency, at the surface, may seem like a good property for an estimator to possess, upon further study we find that super-efficiency is gained at the expense of poor estimation in a neighborhood of zero. To

illustrate this point, consider the sequence $\mu_n = n^{-1/3}$, which converges to zero, the value at which the estimator $\hat{\mu}_n$ is super-efficient. The MLE has the property that

$$n^{1/2}(\bar{Z}_n - \mu_n) \xrightarrow{\mathcal{D}(\mu_n)} N(0,1).$$

However, because $\bar{Z}_n$ concentrates its mass in an $O(n^{-1/2})$ neighborhood about $\mu_n = n^{-1/3}$, which eventually, as n increases, will be completely contained within the range $\pm n^{-1/4}$ with probability converging to one (see Figure 3.3).

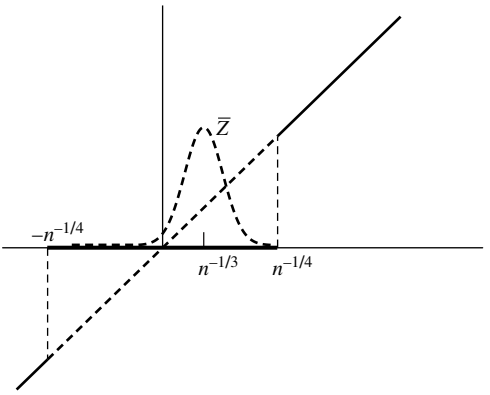

Fig. 3.3. When $\mu_n = n^{-1/3}$, $P_{\mu_n}(\hat{\mu}_n = 0) \to 1$

Therefore,

$$P_{\mu_n}\left[n^{1/2}(\hat{\mu}_n - \mu_n) = -n^{1/2}\mu_n\right] \to 1.$$

Consequently, if $\mu_n = n^{-1/3}$, then

$$-n^{1/2}\mu_n \to -\infty.$$

Therefore, $n^{1/2}(\hat{\mu}_n - \mu_n)$ diverges to $-\infty$.

Although super-efficient estimators exist, they are unnatural and have undesirable local properties associated with them. Therefore, in order to avoid problems associated with super-efficient estimators, we will impose some additional regularity conditions on the class of estimators that will exclude such estimators. Specifically, we will require that an estimator be *regular*, as we now define.

Definition 1. Consider a local data generating process (LDGP), where, for each n, the data are distributed according to "θ_n," where $n^{1/2}(\theta_n - \theta^*)$ converges to a constant (i.e., θ_n is close to some fixed parameter θ^*). That is,

$$Z_{1n}, Z_{2n}, \ldots, Z_{nn} \quad \text{are} \quad \text{iid } p(z, \theta_n),$$

where

$$\theta_n = (\beta_n^T, \eta_n^T)^T, \quad \theta^* = (\beta^{*^T}, \eta^{*^T})^T.$$

An estimator $\hat{\beta}_n$, more specifically $\hat{\beta}_n(Z_{1n}, \ldots, Z_{nn})$, is said to be regular if, for each $\theta^*, n^{1/2}(\hat{\beta}_n - \beta_n)$ has a limiting distribution that does not depend on the LDGP. □

For our purposes, this will ordinarily mean that if

$$n^{1/2}\left\{\hat{\beta}_n\left(Z_{1n}, \ldots, Z_{nn}\right) - \beta^*\right\} \xrightarrow{\mathcal{D}(\theta^*)} N\left(0, \Sigma^*\right),$$

where

$$Z_{1n}, \ldots, Z_{nn} \quad \text{are} \quad \text{iid } p(z, \theta^*), \quad \text{for all } n,$$

then

$$n^{1/2}\left\{\hat{\beta}_n(Z_{1n}, \ldots, Z_{nn}) - \beta_n\right\} \xrightarrow{\mathcal{D}(\theta_n)} N(0, \Sigma^*),$$

where

$$Z_{1n}, \ldots, Z_{nn} \quad \text{are} \quad \text{iid } p(z, \theta_n),$$

and $n^{1/2}(\theta_n - \theta^*) \to \tau^{p\times 1}$, where τ is any arbitrary constant vector.

It is easy to see that, in our previous example, the MLE $\bar{Z}_n$ is a regular estimator, whereas the super-efficient estimator $\hat{\mu}_n$, given by Hodges, is not.

From now on, we will restrict ourselves to regular estimators; in fact, we will only consider estimators that are regular and asymptotically linear (RAL). Although most reasonable estimators are RAL, regular estimators do exist that are not asymptotically linear. However, as a consequence of Hájek's (1970) representation theorem, it can be shown that the most efficient regular estimator is asymptotically linear; hence, it is reasonable to restrict attention to RAL estimators.

In Theorem 3.2 and its subsequent corollary, given below, we present a very powerful result that allows us to describe the geometry of influence functions for regular asymptotically linear (RAL) estimators. This will aid us in defining and visualizing efficiency and will also help us generalize ideas to semiparametric models.

First, we define the score vector for a single observation Z in a parametric model, where $Z \sim p_Z(z, \theta),\ \theta = (\beta^T, \eta^T)^T$, by $S_\theta(Z, \theta_0)$, where

$$S_\theta(z, \theta_0) = \left.\frac{\partial \log p_Z(z, \theta)}{\partial \theta}\right|_{\theta=\theta_0} \tag{3.3}$$

is the p-dimensional vector of derivatives of the log-likelihood with respect to the elements of the parameter θ and θ_0 denotes the true value of θ that generates the data.

This vector can be partitioned according to β (the parameters of interest) and η (the nuisance parameters) as

$$S_\theta(Z,\theta_0) = \{S_\beta^T(Z,\theta_0),\ S_\eta^T(Z,\theta_0)\}^T,$$

where

$$S_\beta(z,\theta_0) = \frac{\partial \log p_Z(z,\theta)}{\partial \beta}\bigg|_{\theta=\theta_0}^{q\times 1}$$

and

$$S_\eta(z,\theta_0) = \frac{\partial \log p_Z(z,\theta)}{\partial \eta}\bigg|_{\theta=\theta_0}^{r\times 1}.$$

Although, in many applications, we can naturally partition the parameter space θ as $(\beta^T,\eta^T)^T$, we will first give results for the more general representation where we define the q-dimensional parameter of interest as a smooth q-dimensional function of the p-dimensional parameter θ; namely, $\beta(\theta)$. As we will show later, especially when we consider infinite-dimensional semiparametric models, in some applications this will be a more natural representation. For parametric models, this is really not a great distinction, as we can always reparametrize the problem so that there is a one-to-one relationship between $\{\beta^T(\theta),\eta^T(\theta)\}^T$ and θ for some r-dimensional nuisance function $\eta(\theta)$.

Theorem 3.2. Let the parameter of interest $\beta(\theta)$ be a q-dimensional function of the p-dimensional parameter θ, $q<p$, such that $\Gamma^{q\times p}(\theta) = \partial\beta(\theta)/\partial\theta^T$, the $q\times p$-dimensional matrix of partial derivatives, exists, has rank q, and is continuous in θ in a neighborhood of the truth θ_0. Also let $\hat\beta_n$ be an asymptotically linear estimator with influence function $\varphi(Z)$ such that $E_\theta(\varphi^T\varphi)$ exists and is continuous in θ in a neighborhood of θ_0. Then, if $\hat\beta_n$ is regular, this will imply that

$$E\{\varphi(Z)S_\theta^T(Z,\theta_0)\} = \Gamma(\theta_0). \tag{3.4}$$

In the special case where θ can be partitioned as $(\beta^T,\eta^T)^T$, we obtain the following corollary.

Corollary 1.

(i)
$$E\{\varphi(Z)S_\beta^T(Z,\theta_0)\} = I^{q\times q}$$

and

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

where $I^{q\times q}$ denotes the $q\times q$ identity matrix and $0^{q\times r}$ denotes the $q\times r$ matrix of zeros.

Theorem 3.2 follows from the definition of regularity together with sufficient smoothness conditions that makes a local data generating process contiguous (to be defined shortly) to the sequence of distributions at the truth. For completeness, we will give an outline of the proof. Before giving the general proof of Theorem 3.2, which is complicated and can be skipped by the reader not interested in all the technical details, we can gain some insight by first showing how Corollary 1 could be proved for the special (and important) case of the class of m-estimators.

3.2 m-Estimators (Quick Review)

In order to define an m-estimator, we consider a $p \times 1$-dimensional function of Z and $\theta, m(Z, \theta)$, such that

$$E_\theta\{m(Z, \theta)\} = 0^{p\times 1},$$

$E_\theta\{m^T(Z, \theta)m(Z, \theta)\} < \infty$, and $E_\theta\{m(Z, \theta)m^T(Z, \theta)\}$ is positive definite for all $\theta \in \Omega$. Additional regularity conditions are also necessary and will be defined as we need them.

The m-estimator $\hat{\theta}_n$ is defined as the solution (assuming it exists) of

$$\sum_{i=1}^{n} m(Z_i, \hat{\theta}_n) = 0$$

from a sample

$$Z_1, \ldots, Z_n \text{ iid } \ p_Z(z, \theta)$$
$$\theta \in \Omega \subset \mathbb{R}^p.$$

Under suitable regularity conditions, the maximum likelihood estimator (MLE) of θ is an m-estimator. The MLE is defined as the value θ that maximizes the likelihood

$$\prod_{i=1}^{n} p_Z(Z_i, \theta),$$

or, equivalently, the value of θ that maximizes the log-likelihood

$$\sum_{i=1}^{n} \log p_Z(Z_i, \theta).$$

Under suitable regularity conditions, the maximum is found by taking the derivative of the log-likelihood with respect to θ and setting it equal to zero. That is, solving the score equation in θ,

$$\sum_{i=1}^{n} S_\theta(Z_i, \theta) = 0, \tag{3.5}$$

where $S_\theta(z,\theta)$ is the score vector (i.e., the derivative of the log-density) defined in (3.3). Since the score vector $S_\theta(Z,\theta)$, under suitable regularity conditions, has the property that $E_\theta\{S_\theta(Z,\theta)\} = 0$ – see, for example, equation (7.3.8) of Casella and Berger (2002) – , this implies that the MLE is an example of an m-estimator.

In order to prove the consistency and asymptotic normality of m-estimators, we need to assume certain regularity conditions. Some of the conditions that are discussed in Chapter 36 of the *Handbook of Econometrics* by Newey and McFadden (1994) include that $E\left\{\dfrac{\partial m(Z,\theta_0)}{\partial\theta^T}\right\}$ be nonsingular, where $\dfrac{\partial m(Z_i,\theta)}{\partial\theta^T}$ is defined as the $p\times p$ matrix of all partial derivatives of the elements of $m(\cdot)$ with respect to the elements of θ, and that

$$n^{-1}\sum_{i=1}^{n}\frac{\partial m(Z_i,\theta)}{\partial\theta^T}\xrightarrow{P}E_{\theta_0}\left\{\frac{\partial m(Z,\theta)}{\partial\theta^T}\right\}$$

uniformly in θ in a neighborhood of θ_0. For example, uniform convergence would be satisfied if the sample paths of $\dfrac{\partial m(Z,\theta)}{\partial\theta^T}$ are continuous in θ about θ_0 almost surely and

$$\sup_{\theta\in\mathcal{N}(\theta_0)}\left|\frac{\partial m(Z,\theta)}{\partial\theta^T}\right|\le g(Z),\quad E\{g(Z)\}<\infty,$$

where $\mathcal{N}(\theta_0)$ denotes a neighborhood in θ about θ_0. In fact, these regularity conditions would suffice to prove that the estimator $\hat\theta_n$ is consistent; that is, $\hat\theta_n\xrightarrow{P}\theta_0$.

Therefore, assuming that these regularity conditions hold, the influence function for $\hat\theta_n$ is found by using the expansion

$$0=\sum_{i=1}^{n}m(Z_i,\hat\theta_n)=\sum_{i=1}^{n}m(Z_i,\theta_0)+\left\{\sum_{i=1}^{n}\frac{\partial m(Z_i,\theta_n^*)}{\partial\theta^T}\right\}^{p\times p}(\hat\theta_n-\theta_0),$$

where θ_n^* is an intermediate value between $\hat\theta_n$ and θ_0.

Because we have assumed sufficient regularity conditions to guarantee the consistency of $\hat\theta_n$,

$$\left\{n^{-1}\sum_{i=1}^{n}\frac{\partial m(Z_i,\theta_n^*)}{\partial\theta^T}\right\}\xrightarrow{P}E\left\{\frac{\partial m(Z,\theta_0)}{\partial\theta^T}\right\},$$

and by the nonsingularity assumption

$$\left\{n^{-1}\sum_{i=1}^{n}\frac{\partial m(Z_i,\theta_n^*)}{\partial\theta^T}\right\}^{-1}\xrightarrow{P}\left[E\left\{\frac{\partial m(Z,\theta_0)}{\partial\theta^T}\right\}\right]^{-1}.$$

Therefore,

$$n^{1/2}(\hat{\theta}_n - \theta_0) = -\left[n^{-1}\sum_{i=1}^{n}\frac{\partial m(Z_i,\theta_n^*)}{\partial\theta^T}\right]^{-1}\left\{n^{-1/2}\sum_{i=1}^{n}m(Z_i,\theta_0)\right\}$$
$$= -\left[E\left\{\frac{\partial m(Z_i,\theta_0)}{\partial\theta^T}\right\}\right]^{-1}\left\{n^{-1/2}\sum_{i=1}^{n}m(Z_i,\theta_0)\right\} + o_p(1).$$

Since, by definition, $E\{m(Z,\theta_0)\} = 0$, we immediately deduce that the influence function of $\hat{\theta}_n$ is given by

$$-\left[E\left\{\frac{\partial m(Z,\theta_0)}{\partial\theta^T}\right\}\right]^{-1} m(Z_i,\theta_0) \tag{3.6}$$

and

$$n^{1/2}(\hat{\theta}_n - \theta_0) \quad \xrightarrow{D} \tag{3.7}$$
$$N\left(0, \left[E\left\{\frac{\partial m(Z,\theta_0)}{\partial\theta^T}\right\}\right]^{-1}\text{var}\left\{m(Z,\theta_0)\right\}\left[E\left\{\frac{\partial m(Z,\theta_0)}{\partial\theta^T}\right\}\right]^{-1^T}\right),$$

where

$$\text{var}\ \{m(Z,\theta_0)\} = E\left\{m(Z,\theta_0)m^T(Z,\theta_0)\right\}.$$

Estimating the Asymptotic Variance of an m-Estimator

In order to use an m-estimator for the parameter θ for practical applications, such as constructing confidence intervals for the parameter θ or a subset of the parameter, we must be able to derive a consistent estimator for the asymptotic variance of $\hat{\theta}_n$. Under suitable regularity conditions, a consistent estimator for the asymptotic variance of $\hat{\theta}_n$ can be derived intuitively using what is referred to as the "sandwich" variance estimator. This estimator is motivated by considering the asymptotic variance derived in (3.7). The following heuristic argument is used.

If θ_0 (the truth) is known, then a simple application of the weak law of large numbers can be used to obtain a consistent estimator for $E\left\{\frac{\partial m(Z,\theta_0)}{\partial\theta^T}\right\}$, namely

$$\hat{E}\left\{\frac{\partial m(Z,\theta_0)}{\partial\theta^T}\right\} = n^{-1}\sum_{i=1}^{n}\frac{\partial m(Z_i,\theta_0)}{\partial\theta^T}, \tag{3.8}$$

and a consistent estimator for $\text{var}\{m(Z,\theta_0)\}$ can be obtained by using

$$\hat{\text{var}}\{m(Z,\theta_0)\} = n^{-1}\sum_{i=1}^{n}m(Z_i,\theta_0)m^T(Z_i,\theta_0). \tag{3.9}$$

Since θ_0 is not known, we instead substitute $\hat{\theta}_n$ for θ_0 in equations (3.8) and (3.9) to obtain the sandwich estimator for the asymptotic variance, (3.7), of $\hat{\theta}_n$, given by

$$\left[\hat{E}\left\{\frac{\partial m(Z,\hat{\theta}_n)}{\partial \theta^T}\right\}\right]^{-1} \hat{\text{var}}\left\{m(Z,\hat{\theta}_n)\right\}\left[\hat{E}\left\{\frac{\partial m(Z,\hat{\theta}_n)}{\partial \theta^T}\right\}\right]^{-1^T}. \tag{3.10}$$

The estimator (3.10) is referred to as the sandwich variance estimator as we see the term $\hat{\text{var}}(\cdot)$ sandwiched between two terms involving $\hat{E}(\cdot)$. The sandwich variance will be discussed in greater detail in Chapter 4 when we introduce estimators that solve generalized estimating equations (i.e., the so-called GEE estimators). For more details on m-estimators, we refer the reader to the excellent expository article by Stefanski and Boos (2002).

When we consider the special case where the m-estimator is the MLE of θ (i.e., where $m(Z,\theta) = S_\theta(Z,\theta)$; see (3.5)), we note that $-\frac{\partial m(Z,\theta)}{\partial \theta^T} = -\frac{\partial S_\theta(Z,\theta)}{\partial \theta^T}$ corresponds to minus the $p \times p$ matrix of second partial derivatives of the log-likelihood with respect to θ, which we denote by $-S_{\theta\theta}(Z,\theta)$. Under suitable regularity conditions (see Section 7.3 of Casella and Berger, 2002), the information matrix, which we denote by $I(\theta_0)$, is given by

$$I(\theta_0) = E_{\theta_0}\{-S_{\theta\theta}(Z,\theta_0)\} = E_{\theta_0}\{S_\theta(Z,\theta_0)S_\theta^T(Z,\theta_0)\}. \tag{3.11}$$

As a consequence of (3.6) and (3.7), we obtain the well-known results that the i-th influence function of the MLE is given by $\{I(\theta_0)\}^{-1}S_\theta(Z_i,\theta_0)$ and the asymptotic distribution is normal with mean zero and variance matrix equal to $I^{-1}(\theta_0)$ (i.e., the inverse of the information matrix).

Returning to the general m-estimator, since

$$\theta = (\beta^T, \eta^T)^T$$

and

$$\hat{\theta}_n = (\hat{\beta}_n^T, \hat{\eta}_n^T)^T,$$

the influence function of $\hat{\beta}_n$ is made up of the first q elements of the p-dimensional influence function for $\hat{\theta}_n$ given above.

We will now illustrate why Corollary 1 applies to m-estimators. By definition,

$$E_\theta\{m(Z,\theta)\} = 0^{p\times 1}.$$

That is,

$$\int m(z,\theta)p(z,\theta)d\nu(z) = 0 \text{ for all } \theta.$$

Therefore,

$$\frac{\partial}{\partial \theta^T}\int m(z,\theta)p(z,\theta)d\nu(z) = 0.$$

Assuming suitable regularity conditions that allow us to interchange integration and differentiation, we obtain

$$\int \left\{ \frac{\partial}{\partial \theta^T} m(z,\theta) \right\} p(z,\theta) d\nu(z) + \int m(z,\theta) \left\{ \frac{\frac{\partial p(z,\theta)}{\partial \theta^T}}{\underbrace{p(z,\theta)}} \right\} p(z,\theta) d\nu(z) = 0. \tag{3.12}$$

$$\| \quad S_\theta^T(z,\theta) \text{ or the transpose of the score vector}$$

At $\theta = \theta_0$, we deduce from equation (3.12) that

$$E\left\{ \frac{\partial m(Z,\theta_0)}{\partial \theta^T} \right\} = -E\left\{ m(Z,\theta_0) S_\theta^T(Z,\theta_0) \right\},$$

which can also be written as

$$I^{p\times p} = -\left[E\left\{ \frac{\partial m(Z,\theta_0)}{\partial \theta^T} \right\} \right]^{-1} E\left\{ m(Z,\theta_0) S_\theta^T(Z,\theta_0) \right\}, \tag{3.13}$$

where $I^{p\times p}$ denotes the $p\times p$ identity matrix. Recall that the influence function for $\hat{\theta}_n$, given by (3.6), is

$$\varphi_{\hat{\theta}_n}(Z_i) = -\left[E\left\{ \frac{\partial m(Z,\theta_0)}{\partial \theta^T} \right\} \right]^{-1} m(Z_i,\theta_0)$$

and can be partitioned as $\left\{ \varphi_{\hat{\beta}_n}^T(Z_i), \varphi_{\hat{\eta}_n}^T(Z_i) \right\}^T$.

The covariance of the influence function $\varphi_{\hat{\theta}_n}(Z_i)$ and the score vector $S_\theta(Z_i,\theta_0)$ is

$$\begin{aligned} & E\left\{ \varphi_{\hat{\theta}_n}(Z_i) S_\theta^T(Z_i,\theta_0) \right\} \\ &= -\left[E\left\{ \frac{\partial m(Z,\theta_0)}{\partial \theta^T} \right\} \right]^{-1} E\left\{ m(Z,\theta_0) S_\theta^T(Z,\theta_0) \right\}, \end{aligned} \tag{3.14}$$

which by (3.13) is equal to $I^{(q+r)\times(q+r)}$, the identity matrix. This covariance matrix (3.14) can be partitioned as

$$E\begin{bmatrix} \varphi_{\hat{\beta}_n}(Z_i) S_\beta^T(Z_i,\theta_0) & \varphi_{\hat{\beta}_n}(Z_i) S_\eta^T(Z_i,\theta_0) \\ \varphi_{\hat{\eta}_n}(Z_i) S_\beta^T(Z_i,\theta_0) & \varphi_{\hat{\eta}_n}(Z_i) S_\eta^T(Z_i,\theta_0) \end{bmatrix}.$$

Consequently,

(i) $E\left\{ \varphi_{\hat{\beta}_n}(Z_i) S_\beta^T(Z_i,\theta_0) \right\} = I^{q\times q}$ (the $q\times q$ identity matrix)

and

(ii) $E\left\{\varphi_{\hat{\beta}_n}(Z_i)S_\eta^T(Z_i,\theta_0)\right\} = 0^{q\times r}$.

Thus, we have verified that the two conditions of Corollary 1 hold for influence functions of m-estimators.

Proof of Theorem 3.2

In order to prove Theorem 3.2, we must introduce the theory of contiguity, which we now review briefly. An excellent overview of contiguity theory can be found in the Appendix of Hájek and Sidak (1967). Those readers not interested in the theoretical details can skip the remainder of this section.

Definition 2. Let V_n be a sequence of random vectors and let P_{1n} and P_{0n} be sequences of probability measures with densities $p_{1n}(v_n)$ and $p_{0n}(v_n)$, respectively. The sequence of probability measures P_{1n} is contiguous to the sequence of probability measures P_{0n} if, for any sequence of events A_n defined with respect to V_n, $P_{0n}(A_n) \to 0$ as $n \to \infty$ implies that $P_{1n}(A_n) \to 0$ as $n \to \infty$. □

In our applications, we let $V_n = (Z_{1n}, \ldots, Z_{nn})$, where $Z_{1n}, \ldots, Z_{nn}$ are iid random vectors and

$$p_{0n}(v_n) = \prod_{i=1}^{n} p(z_{in}, \theta_0),$$

$$p_{1n}(v_n) = \prod_{i=1}^{n} p(z_{in}, \theta_n),$$

where $n^{1/2}(\theta_n - \theta_0) \to \tau$, τ being a p-dimensional vector of constants.

Letting the parameter θ_0 denote the true value of the parameter that generates the data, then $p_{1n}(\cdot)$ is an example of a local data generating process (LDGP) as given by Definition 1. If we could show that the sequence P_{1n} is contiguous to the sequence P_{0n}, then a sequence of random variables $T_n(V_n)$ that converges in probability to zero under the truth (i.e., for every $\epsilon > 0$, $P_{0n}(|T_n| > \epsilon) \to 0$) would also satisfy that $P_{1n}(|T_n| > \epsilon) \to 0$; hence, $T_n(V_n)$ would converge in probability to zero for the LDGP. This fact can be very useful because in some problems it may be relatively easy to show that a sequence of random variables converges in probability to zero under the truth, in which case convergence in probability to zero under the LDGP follows immediately from contiguity.

LeCam, in a series of lemmas (see Hájek and Sidak, 1967), proved some important results regarding contiguity. One of LeCam's results that is of particular use to us is as follows.

Lemma 3.1. *LeCam*
If

$$\log\left\{\frac{p_{1n}(V_n)}{p_{0n}(V_n)}\right\} \xrightarrow{\mathcal{D}(P_{0n})} N(-\sigma^2/2, \sigma^2), \tag{3.15}$$

then the sequence P_{1n} is contiguous to the sequence P_{0n}.

Heuristic justification of contiguity for LDGP

To illustrate that (3.15) holds for LDGPs under sufficient smoothness and regularity conditions, we sketch out the following heuristic argument. Define

$$L_n(V_n) = \frac{p_{1n}(V_n)}{p_{0n}(V_n)} = \prod_{i=1}^{n} \frac{p(Z_{in}, \theta_n)}{p(Z_{in}, \theta_0)}.$$

By a simple Taylor series expansion, we obtain

$$\begin{aligned}\log\{L_n(V_n)\} &= \sum_{i=1}^{n}\{\log p(Z_{in}, \theta_n) - \log p(Z_{in}, \theta_0)\} \\ &= (\theta_n - \theta_0)^T \left\{\sum_{i=1}^{n} S_\theta(Z_{in}, \theta_0)\right\} \\ &\quad + \frac{(\theta_n - \theta_0)^T\{\sum_{i=1}^{n} S_{\theta\theta}(Z_{in}, \theta_n^*)\}(\theta_n - \theta_0)}{2},\end{aligned} \tag{3.16}$$

where $S_\theta(z, \theta_0)$ is the p-dimensional score vector defined as $\partial \log p(z, \theta_0)/\partial\theta$, $S_{\theta\theta}(z, \theta_n^*)$ is the $p \times p$ matrix $\partial^2 \log p(z, \theta_n^*)/\partial\theta\partial\theta^T$, and θ_n^* is some intermediate value between θ_n and θ_0.

The expression (3.16) can be written as

$$\begin{aligned}&n^{1/2}(\theta_n - \theta_0)^T \left\{n^{-1/2}\sum_{i=1}^{n} S_\theta(Z_{in}, \theta_0)\right\} \\ &+ \frac{n^{1/2}(\theta_n - \theta_0)^T\{n^{-1}\sum_{i=1}^{n} S_{\theta\theta}(Z_{in}, \theta_n^*)\}n^{1/2}(\theta_n - \theta_0)}{2}.\end{aligned}$$

Under P_{0n}:

(i) $S_\theta(Z_{in}, \theta_0), i = 1, \ldots, n$ are iid mean zero random vectors with variance matrix equal to the information matrix $I(\theta_0)$ defined by (3.11). Consequently, by the CLT,

$$n^{-1/2}\sum_{i=1}^{n} S_\theta(Z_{in}, \theta_0) \xrightarrow{\mathcal{D}(P_{0n})} N\Big(0, I(\theta_0)\Big).$$

(ii) Since $\theta_n^* \to \theta_0$ and $S_{\theta\theta}(Z_{in}, \theta_0), i = 1, \ldots, n$ are iid random matrices with mean $-I(\theta_0)$, then, under sufficient smoothness conditions,

$$n^{-1}\sum_{i=1}^{n}\{S_{\theta\theta}(Z_{in},\theta_n^*)-S_{\theta\theta}(Z_{in},\theta_0)\}\xrightarrow{P}0,$$

and by the weak law of large numbers

$$n^{-1}\sum_{i=1}^{n}S_{\theta\theta}(Z_{in},\theta_0)\xrightarrow{P}-I(\theta_0),$$

hence

$$n^{-1}\sum_{i=1}^{n}S_{\theta\theta}(Z_{in},\theta_n^*)\xrightarrow{P}-I(\theta_0).$$

By assumption, $n^{1/2}(\theta_n-\theta_0)\to\tau$. Therefore, (i), (ii), and Slutsky's theorem imply that

$$\log\{L_n(V_n)\}\xrightarrow{\mathcal{D}(P_{0n})}N\left(-\frac{\tau^T I(\theta_0)\tau}{2},\tau^T I(\theta_0)\tau\right).$$

Consequently, by LeCam's lemma, the sequence P_{1n} is contiguous to the sequence P_{0n}.

Now we are in a position to prove Theorem 3.2.

Proof. Theorem 3.2
Consider the sequence of densities $p_{0n}(v_n)=\prod p(z_{in},\theta_0)$ and the LDGP $p_{1n}(v_n)=\prod p(z_{in},\theta_n)$, where $n^{1/2}(\theta_n-\theta_0)\to\tau$. By the definition of asymptotic linearity,

$$n^{1/2}\{\hat{\beta}_n-\beta(\theta_0)\}=n^{-1/2}\sum_{i=1}^{n}\varphi(Z_{in})+o_{P_{0n}}(1),\tag{3.17}$$

where $o_{P_{0n}}(1)$ is a sequence of random vectors that converge in probability to zero with respect to the sequence of probability measures P_{0n}. Consider the LDGP defined by θ_n. By contiguity, terms that are $o_{P_{0n}}(1)$ are also $o_{P_{1n}}(1)$. Consequently, by (3.17),

$$n^{1/2}\{\hat{\beta}_n-\beta(\theta_0)\}=n^{-1/2}\sum_{i=1}^{n}\varphi(Z_{in})+o_{P_{1n}}(1).$$

By adding and subtracting common terms, we obtain

$$\begin{aligned}n^{1/2}\{\hat{\beta}_n-\beta(\theta_n)\}=n^{-1/2}&\sum_{i=1}^{n}[\varphi(Z_{in})-E_{\theta_n}\{\varphi(Z)\}]\\&+n^{1/2}E_{\theta_n}\{\varphi(Z)\}-n^{1/2}\{\beta(\theta_n)-\beta(\theta_0)\}\\&+o_{P_{1n}}(1).\end{aligned}\tag{3.18}$$

By assumption, the estimator $\hat{\beta}_n$ is regular; that is,

$$n^{1/2}\{\hat{\beta}_n - \beta(\theta_n)\} \xrightarrow{\mathcal{D}(P_{1n})} N\Big(0, E_{\theta_0}(\varphi\varphi^T)\Big). \tag{3.19}$$

Also, under P_{1n}, $[\varphi(Z_{in}) - E_{\theta_n}\{\varphi(Z)\}], i = 1, \ldots, n$ are iid mean-zero random vectors with variance matrix $E_{\theta_n}(\varphi\varphi^T) - E_{\theta_n}(\varphi)E_{\theta_n}(\varphi^T)$. By the smoothness assumption, $E_{\theta_n}(\varphi\varphi^T) \to E_{\theta_0}(\varphi\varphi^T)$ and $E_{\theta_n}(\varphi) \to 0$ as $n \to \infty$. Hence, by the CLT, we obtain

$$n^{-1/2}\sum_{i=1}^{n}[\varphi(Z_{in}) - E_{\theta_n}\{\varphi(Z)\}] \xrightarrow{\mathcal{D}(P_{1n})} N\Big(0, E_{\theta_0}(\varphi\varphi^T)\Big). \tag{3.20}$$

By a simple Taylor series expansion, we deduce that $\beta(\theta_n) \approx \beta(\theta_0) + \Gamma(\theta_0)(\theta_n - \theta_0)$, where $\Gamma(\theta_0) = \partial\beta(\theta_0)/\partial\theta^T$. Hence,

$$n^{1/2}\{\beta(\theta_n) - \beta(\theta_0)\} \to \Gamma(\theta_0)\tau. \tag{3.21}$$

Finally,

$$\begin{aligned} n^{1/2}E_{\theta_n}\{\varphi(Z)\} &= n^{1/2}\int \varphi(z)p(z,\theta_n)d\nu(z) \\ = n^{1/2}\int \varphi(z)p(z,\theta_0)d\nu(z) &+ n^{1/2}\int \varphi(z)\left\{\frac{\partial p(z,\theta_n^*)}{\partial\theta}\right\}^T(\theta_n - \theta_0)d\nu(z) \\ &\xrightarrow{n\to\infty} 0 + \int \varphi(z)\left\{\frac{\partial p(z,\theta_0)}{\partial\theta}/p(z,\theta_0)\right\}^T p(z,\theta_0)d\nu(z)\tau \\ &= E_{\theta_0}\{\varphi(Z)S_\theta^T(Z,\theta_0)\}\tau, \end{aligned} \tag{3.22}$$

where θ_n^* is some intermediate value between θ_n and θ_0. The only way that (3.19) and (3.20) can hold is if the limit of (3.18), as $n \to \infty$, is identically equal to zero. By (3.21) and (3.22), this implies that

$$\left[E_{\theta_0}\{\varphi(Z)S_\theta^T(Z,\theta_0)\} - \Gamma(\theta_0)\right]\tau = 0^{q\times 1}.$$

Since τ is arbitrary, this implies that

$$E_{\theta_0}\{\varphi(Z)S_\theta^T(Z,\theta_0)\} = \Gamma(\theta_0),$$

which proves Theorem 3.2. □

We now show how the results of Theorem 3.2 lend themselves to a geometric interpretation that allows us to compare the efficiency of different RAL estimators using our intuition of minimum distance and orthogonality.

3.3 Geometry of Influence Functions for Parametric Models

Consider the Hilbert space $\mathcal{H}$ of all q-dimensional measurable functions of Z with mean zero and finite variance equipped with the inner product $\langle h_1, h_2 \rangle = E(h_1^T h_2)$. We first note that the score vector $S_\theta(Z, \theta_0)$, under suitable regularity conditions, has mean zero (i.e., $E\{S_\theta(Z, \theta_0)\} = 0^{p \times 1}$). Similar to Example 2 of Chapter 2, we can define the finite-dimensional linear subspace $\mathscr{T} \subset \mathcal{H}$ spanned by the p-dimensional score vector $S_\theta(Z, \theta_0)$ as the set of all q-dimensional mean-zero random vectors consisting of

$$B^{q \times p} S_\theta(Z, \theta_0)$$

for all $q \times p$ matrices B. The linear subspace $\mathscr{T}$ is referred to as the tangent space.

In the case where θ can be partitioned as $(\beta^T, \eta^T)^T$, consider the linear subspace spanned by the nuisance score vector $S_\eta(Z, \theta_0)$,

$$B^{q \times r} S_\eta(Z, \theta_0), \tag{3.23}$$

for all $q \times r$ matrices B. This space is referred to as the nuisance tangent space and will be denoted by Λ. We note that condition (ii) of Corollary 1 is equivalent to saying that the q-dimensional influence function $\varphi_{\hat{\beta}_n}(Z)$ for $\hat{\beta}_n$ is orthogonal to the nuisance tangent space Λ.

In addition to being orthogonal to the nuisance tangent space, the influence function of $\hat{\beta}_n$ must also satisfy condition (i) of Corollary 1; namely,

$$E\left\{\varphi_{\hat{\beta}_n}(Z) S_\beta^T(Z, \theta_0)\right\} = I^{q \times q}.$$

Although influence functions of RAL estimators for β must satisfy conditions (i) and (ii) of Corollary 1, a natural question is whether the converse is true; that is, for any element of the Hilbert space satisfying conditions (i) and (ii) of Corollary 1, does there exist an RAL estimator for β with that influence function?

Remark 2. To prove this in full generality, especially later when we consider infinite-dimensional nuisance parameters, is difficult and requires that some careful technical regularity conditions hold. Nonetheless, it may be instructive to see how one may, heuristically, construct estimators that have influence functions corresponding to elements in the subspace of the Hilbert space satisfying conditions (i) and (ii). □

Constructing Estimators

Let $\varphi(Z)$ be a q-dimensional measurable function with zero mean and finite variance that satisfies conditions (i) and (ii) of Corollary 1. Define

$$m(Z, \beta, \eta) = \varphi(Z) - E_{\beta,\eta}\{\varphi(Z)\}.$$

Assume that we can find a root-n consistent estimator for the nuisance parameter $\hat{\eta}_n$ (i.e., where $n^{1/2}(\hat{\eta}_n - \eta_0)$ is bounded in probability). In many cases the estimator $\hat{\eta}_n$ will be β-dependent (i.e., $\hat{\eta}_n(\beta)$). For example, we might use the MLE for η, or the restricted MLE for η, fixing the value of β.

We will now argue that the solution to the equation

$$\sum_{i=1}^{n} m\{Z_i, \beta, \hat{\eta}_n(\beta)\} = 0, \tag{3.24}$$

which we denote by $\hat{\beta}_n$, will be an asymptotically linear estimator with influence function $\varphi(Z)$.

By construction, we have

$$E_{\beta_0,\eta}\{m(Z, \beta_0, \eta)\} = 0,$$

or

$$\int m(z, \beta_0, \eta) p(z, \beta_0, \eta) d\nu(z) = 0.$$

Consequently,

$$\frac{\partial}{\partial \eta^T}\bigg|_{\eta=\eta_0} \int m(z, \beta_0, \eta) p(z, \beta_0, \eta) d\nu(z) = 0,$$

or

$$\int \frac{\partial m(z, \beta_0, \eta_0)}{\partial \eta^T} p(z, \beta_0, \eta_0) d\nu(z) + \int m(z, \beta_0, \eta_0) \times S_\eta^T(z, \beta_0, \eta_0) p(z, \beta_0, \eta_0) d\nu(z) = 0. \tag{3.25}$$

By definition, $\varphi(Z) = m(Z, \beta_0, \eta_0)$ must satisfy

$$E\left\{\varphi(Z) S_\eta^T(Z, \theta_0)\right\} = 0.$$

(This is condition (ii) of Corollary 1.) Consequently, by (3.25), we obtain

$$E\left\{\frac{\partial}{\partial \eta^T} m(Z, \beta_0, \eta_0)\right\} = 0. \tag{3.26}$$

Similarly, we can show that

$$E\left\{\frac{\partial}{\partial \beta^T} m(Z, \beta_0, \eta_0)\right\} = -I^{q \times q}. \tag{3.27}$$

A standard expansion yields

$$
\begin{aligned}
0 &= \sum_{i=1}^{n} m\{Z_i, \hat{\beta}_n, \hat{\eta}_n(\hat{\beta}_n)\} \\
&= \sum_{i=1}^{n} m\{Z_i, \beta_0, \hat{\eta}_n(\hat{\beta}_n)\} \\
&\quad + \left[\sum_{i=1}^{n} \frac{\partial m}{\partial \beta^T}\{Z_i, \beta_n^*, \underbrace{\hat{\eta}_n(\hat{\beta}_n)}_{\Downarrow}\}\right](\hat{\beta}_n - \beta_0), \qquad (3.28)
\end{aligned}
$$

Notice that this term is held fixed

where β_n^* is an intermediate value between $\hat{\beta}_n$ and β_0. Therefore,

$$
\begin{aligned}
& n^{1/2}(\hat{\beta}_n - \beta_0) \\
&= -\underbrace{\left[n^{-1}\sum_{i=1}^{n} \frac{\partial}{\partial \beta^T} m\{Z_i, \beta_n^*, \hat{\eta}_n(\hat{\beta}_n)\}\right]^{-1}}_{\Downarrow p} \left[n^{-1/2}\sum_{i=1}^{n} m\{Z_i, \beta_0, \hat{\eta}_n(\hat{\beta}_n)\}\right]. \\
& \qquad \left[E\left\{\frac{\partial}{\partial \beta^T} m(Z, \beta_0, \eta_0)\right\}\right]^{-1} = -I^{q\times q} \text{ by (3.27)} \qquad (3.29)
\end{aligned}
$$

Let us consider the second term of (3.29); namely, $n^{-1/2}\sum_{i=1}^{n} m\{Z_i, \beta_0, \hat{\eta}_n(\hat{\beta}_n)\}$. By expansion, this equals

$$
\begin{aligned}
& n^{-1/2}\sum_{i=1}^{n} m(Z_i, \beta_0, \eta_0) \\
& + \underbrace{\left\{n^{-1}\sum_{i=1}^{n} \frac{\partial m(Z_i, \beta_0, \eta_n^*)}{\partial \eta^T}\right\}}_{\substack{\Downarrow p \\ E\left\{\frac{\partial}{\partial \eta^T} m(Z, \beta_0, \eta_0)\right\} \\ = 0 \text{ by (3.26)}}} \quad \underbrace{\left[n^{1/2}\{\hat{\eta}_n(\hat{\beta}_n) - \eta_0\}\right]}_{\substack{\Downarrow \\ \text{bounded in probability}}}, \qquad (3.30)
\end{aligned}
$$

where η_n^* is an intermediate value between $\hat{\eta}_n(\hat{\beta}_n)$ and η_0.

Combining (3.29) and (3.30), we obtain

$$
\begin{aligned}
n^{1/2}(\hat{\beta}_n - \beta_0) &= n^{-1/2}\sum_{i=1}^{n} m(Z_i, \beta_0, \eta_0) + o_p(1), \\
&= n^{-1/2}\sum_{i=1}^{n} \varphi(Z_i) + o_p(1),
\end{aligned}
$$

which illustrates that $\varphi(Z_i)$ is the influence function for the i-th observation of the estimator $\hat{\beta}_n$ above.

Remark 3. This argument was independent of the choice of the root-n consistent estimator for the nuisance parameter η. □

Remark 4. In the derivation above, the asymptotic distribution of the estimator obtained by solving the estimating equation, which uses the estimating function $m(Z, \beta, \hat{\eta}_n)$, is the same as the asymptotic distribution of the estimator solving the estimating equation using the estimating function $m(Z, \beta, \eta_0)$ had the true value of the nuisance parameter η_0 been known to us. This fact follows from the orthogonality of the estimating function (evaluated at the truth) to the nuisance tangent space. This type of robustness, where the asymptotic distribution of an estimator is independent of whether the true value of the nuisance parameter is known or whether (and how) the nuisance parameter is estimated in an estimating equation, is one of the bonuses of working with estimating equations with estimating functions that are orthogonal to the nuisance tangent space. □

Remark 5. We want to make it clear that the estimator we just presented is for theoretical purposes only and not of practical use. The starting point was the choice of a function satisfying the conditions of Lemma 3.1. To find such a function necessitates knowledge of the truth, which, of course, we don't have. Nonetheless, starting with some truth, say θ_0, and some function $\varphi(Z)$ satisfying the conditions of Corollary 1 (under the assumed true model), we constructed an estimator whose influence function is $\varphi(Z)$ when θ_0 is the truth. If, however, the data were generated, in truth, by some other value of the parameter, say θ^*, then the estimator constructed by solving (3.24) would have some other influence function $\varphi^*(Z)$ satisfying the conditions of Lemma 3.1 at θ^*. □

Thus, by Corollary 1, all RAL estimators have influence functions that belong to the subspace of our Hilbert space satisfying

(i) $E\{\varphi(Z)S_\beta^T(Z, \theta_0)\} = I^{q\times q}$

and

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

and, conversely, any element in the subspace above is the influence function of some RAL estimator.

Why Is this Important?

RAL estimators are asymptotically normally distributed; i.e.,

$$n^{1/2}(\hat{\beta}_n - \beta_0) \xrightarrow{\mathcal{D}} N\Big(0, E(\varphi\varphi^T)\Big).$$

Because of this, we can compare competing RAL estimators for β by looking at the asymptotic variance, where clearly the better estimator is the one with smaller asymptotic variance. We argued earlier, however, that the asymptotic variance of an RAL estimator is the variance of its influence function. Therefore, it suffices to consider the variance of influence functions. We already illustrated that influence functions can be viewed as elements in a subspace of a Hilbert space. Moreover, in this Hilbert space the distance to the origin (squared) of any element (random function) is the variance of the element. Consequently, the search for the best estimator (i.e., the one with the smallest asymptotic variance) is equivalent to the search for the element in the subspace of influence functions that has the shortest distance to the origin.

Remark 6. We want to emphasize again that Hilbert spaces are characterized by both the elements that make up the space (random functions in our case) and the inner product, $\langle h_1, h_2 \rangle = E(h_1^T h_2)$, where expectation is always taken with respect to the truth (θ_0). Therefore, for different θ_0, we have different Hilbert spaces. This also means that the subspace that defines the class of influence functions is θ_0-dependent. □

3.4 Efficient Influence Function

We will show how the geometry of Hilbert spaces will allow us to identify the most efficient influence function (i.e., the influence function with the smallest variance). First, however, we give some additional notation and definitions regarding operations on linear subspaces that will be needed shortly.

Definition 3. We say that $M \oplus N$ is a direct sum of two linear subspaces $M \subset \mathcal{H}$ and $N \subset \mathcal{H}$ if $M \oplus N$ is a linear subspace in $\mathcal{H}$ and if every element $x \in M \oplus N$ has a unique representation of the form $x = m + n$, where $m \in M$ and $n \in N$. □

Definition 4. The set of elements of a Hilbert space that are orthogonal to a linear subspace M is denoted by $M^\perp$. The space $M^\perp$ is also a linear subspace (referred to as the orthogonal complement of M) and the entire Hilbert space

$$\mathcal{H} = M \oplus M^\perp. \square$$

Condition (ii) of Corollary 1 can now be stated as follows: If $\varphi(Z)$ is an influence function of an RAL estimator, then $\varphi \in \Lambda^\perp$, where Λ denotes the nuisance tangent space defined by (3.23).

Definition 5. If we consider any arbitrary element $h(Z) \in \mathcal{H}$, then by the projection theorem, there exists a unique element $a_0(Z) \in \Lambda$ such that $\|h - a_0\|$ has the minimum norm and a_0 must uniquely satisfy the relationship

$$\langle h - a_0, a \rangle = 0 \text{ for all } a \in \Lambda.$$

The element a_0 is referred to as the projection of h onto the space Λ and is denoted by $\Pi(h|\Lambda)$. The element with the minimum norm, $h-a_0$, is sometimes referred to as the residual of h after projecting onto Λ, and it is easy to show that $h - a_0 = \Pi(h|\Lambda^{\perp})$. □

As we discussed earlier, condition (ii) of Corollary 1 is equivalent to an element $h(Z)$ in our Hilbert space $\mathcal{H}$ being orthogonal to the nuisance tangent space; i.e., the linear subspace generated by the nuisance score vector, namely

$$\Lambda = \{B^{q\times r} S_\eta(Z, \theta_0) \text{ for all } B^{q\times r}\}.$$

If we want to identify all elements orthogonal to the nuisance tangent space, we can consider the set of elements $h - \Pi(h|\Lambda)$ for all $h \in \mathcal{H}$, where using the results in Example 2 of Chapter 2,

$$\Pi(h|\Lambda) = E(hS_\eta^T)\{E(S_\eta S_\eta^T)\}^{-1} S_\eta(Z, \theta_0).$$

It is also straightforward to show that the tangent space

$$\mathscr{T} = \{B^{q\times p} S_\theta(Z, \theta_0) \text{ for all } B^{q\times p}\}$$

can be written as the direct sum of the nuisance tangent space and the tangent space generated by the score vector with respect to the parameter of interest "β". That is, if we define $\mathscr{T}_\beta$ as the space $\{B^{q\times q} S_\beta(Z, \theta_0) \text{ for all } B^{q\times q}\}$, then $\mathscr{T} = \mathscr{T}_\beta \oplus \Lambda$.

Asymptotic Variance when Dimension Is Greater than One

When the parameter of interest β has dimension ≥ 2, we must be careful about what we mean by smaller asymptotic variance for an estimator or its influence function. Consider two RAL estimators for β with influence function $\varphi^{(1)}(Z)$ and $\varphi^{(2)}(Z)$, respectively. We say that

$$\text{var}\,\{\varphi^{(1)}(Z)\} \leq \text{var}\,\{\varphi^{(2)}(Z)\}$$

if and only if

$$\text{var}\,\{a^T\varphi^{(1)}(Z)\} \leq \text{var}\,\{a^T\varphi^{(2)}(Z)\}$$

for all $q \times 1$ constant vectors a. Equivalently,

$$a^T E\{\varphi^{(1)}(Z)\varphi^{(1)^T}(Z)\}a \leq a^T E\{\varphi^{(2)}(Z)\varphi^{(2)^T}(Z)\}a.$$

This means that

$$a^T[E\{\varphi^{(2)}(Z)\varphi^{(2)^T}(Z)\} - E\{\varphi^{(1)}(Z)\varphi^{(1)^T}(Z)\}]\ a \geq 0,$$

or $E(\varphi^{(2)}\varphi^{(2)^T}) - E(\varphi^{(1)}\varphi^{(1)^T})$ is nonnegative definite.

If $\mathcal{H}^{(1)}$ is the Hilbert space of one-dimensional mean-zero random functions of Z, where we use the superscript (1) to emphasize one-dimensional random functions, and if h_1 and h_2 are elements of $\mathcal{H}^{(1)}$ that are orthogonal to each other, then, by the Pythagorean theorem, we know that $\text{var}(h_1 + h_2) = \text{var}(h_1) + \text{var}(h_2)$, making it clear that $\text{var}(h_1 + h_2)$ is greater than or equal to $\text{var}(h_1)$ or $\text{var}(h_2)$. Unfortunately, when $\mathcal{H}$ consists of q-dimensional mean-zero random functions, there is no such general relationship with regard to the variance matrices. However, there is an important special case when this does occur, which we now discuss.

Definition 6. *q-replicating linear space*
A linear subspace $\mathcal{U} \subset \mathcal{H}$ is a q-replicating linear space if $\mathcal{U}$ is of the form $\mathcal{U}^{(1)} \times \ldots \times \mathcal{U}^{(1)}$ or $\{\mathcal{U}^{(1)}\}^q$, where $\mathcal{U}^{(1)}$ denotes a linear subspace in $\mathcal{H}^{(1)}$ and $\{\mathcal{U}^{(1)}\}^q \subset \mathcal{H}$ represents the linear subspace in $\mathcal{H}$ that consists of elements $h = (h^{(1)}, \ldots, h^{(q)})^T$ such that $h^{(j)} \in \mathcal{U}^{(1)}$ for all $j = 1, \ldots, q$; i.e., $\{\mathcal{U}^{(1)}\}^q$ consists of q-dimensional random functions, where each element in the vector is an element of $\mathcal{U}^{(1)}$, or the space $\mathcal{U}^{(1)}$ stacked up on itself q times. □

The linear subspace spanned by an r-dimensional vector of mean zero finite variance random functions $v^{r\times 1}(Z)$, namely the subspace

$$\mathcal{S} = \{B^{q\times r} v(Z) : \text{ for all constant matrices } B^{q\times r}\},$$

is such a subspace. This is easily seen by defining $\mathcal{U}^{(1)}$ to be the space

$$\{b^T v(Z) : \text{ for all constant } r\text{-dimensional constant vectors } b^{r\times 1}\},$$

in which case $\mathcal{S} = \{\mathcal{U}^{(1)}\}^q$. Since tangent spaces and nuisance tangent spaces are linear subspaces spanned by score vectors, these are examples of q-replicating linear spaces.

Theorem 3.3. *Multivariate Pythagorean theorem*
If $h \in \mathcal{H}$ and is an element of a q-replicating linear space $\mathcal{U}$, and $\ell \in \mathcal{H}$ is orthogonal to $\mathcal{U}$, then

$$\text{var}(\ell + h) = \text{var}(\ell) + \text{var}(h), \tag{3.31}$$

where $\text{var}(h) = E(hh^T)$. As a consequence of (3.31), we obtain a multivariate version of the Pythagorean theorem; namely, for any $h^* \in \mathcal{H}$,

$$\text{var}(h^*) = \text{var}\left(\Pi[h^*|\mathcal{U}]\right) + \text{var}\left(h^* - \Pi[h^*|\mathcal{U}]\right). \tag{3.32}$$

Proof. It is easy to show that an element $\ell = (\ell^{(1)}, \ldots, \ell^{(q)})^T \in \mathcal{H}$ is orthogonal to $\mathcal{U} = \{\mathcal{U}^{(1)}\}^q$ if and only if each element $\ell^{(j)}, j = 1, \ldots, q$ is orthogonal to $\mathcal{U}^{(1)}$. Consequently, such an element ℓ is not only orthogonal to $h \in \{\mathcal{U}^{(1)}\}^q$ in the sense that $E(\ell^T h) = 0$ but also in that $E(\ell h^T) = E(h\ell^T) = 0^{q\times q}$. This is important because for such an ℓ and h, we obtain

$$\text{var}(\ell + h) = \text{var}(\ell) + \text{var}(h),$$

where $\text{var}(h) = E(hh^T)$. □

This means that, for such cases, the variance matrix of $\ell+h$, for q-dimensional ℓ and h, is larger (in the multidimensional sense defined above) than either the variance matrix of ℓ or the variance matrix of h.

In many of the arguments that follow, we will be decomposing elements of the Hilbert space as the projection to a tangent space or a nuisance tangent space plus the residual after the projection. For such problems, because the tangent space or nuisance tangent space is a q-replicating linear space, we now know that we can immediately apply the multivariate version of the Pythagorean theorem where the variance matrix of any element is always larger than the variance matrix of the projection or the variance matrix of the residual after projection. Consequently, we don't have to distinguish between the Hilbert space of one-dimensional random functions and q-dimensional random functions.

Geometry of Influence Functions

Before describing the geometry of influence functions, we first give the definition of a linear variety (sometimes also called an affine space).

Definition 7. A *linear variety* is the translation of a linear subspace away from the origin; i.e., a linear variety V can be written as $V = x_0 + M$, where $x_0 \in \mathcal{H}$ and $x_0 \notin M, \|x_0\| \neq 0$, and M is a linear subspace (see Figure 3.4). □

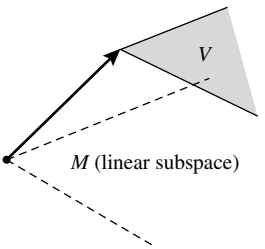

Fig. 3.4. Depiction of a linear variety

Theorem 3.4. The set of all influence functions, namely the elements of $\mathcal{H}$ that satisfy condition (3.4) of Theorem 3.2, is the linear variety $\varphi^*(Z) + \mathscr{T}^\perp$, where $\varphi^*(Z)$ is any influence function and $\mathscr{T}^\perp$ is the space perpendicular to the tangent space.

Proof. Any element $l(Z) \in \mathscr{T}^\perp$ must satisfy

$$E\{l(Z)S_\theta^T(Z,\theta_0)\} = 0^{q\times p}. \tag{3.33}$$

Therefore, if we take

$$\varphi(Z) = \varphi^*(Z) + l(Z),$$

then

$$\begin{aligned} E\{\varphi(Z)S_\theta^T(Z,\theta_0)\} &= E\left[\{\varphi^*(Z) + l(Z)\}\, S_\theta^T(Z,\theta_0)\right] \\ &= E\left[\varphi^*(Z)S_\theta^T(Z,\theta_0)\right] + E\left[l(Z)S_\theta^T(Z,\theta_0)\right] \\ &= \Gamma(\theta_0) + 0^{q\times p} = \Gamma(\theta_0). \end{aligned}$$

Hence, $\varphi(Z)$ is an influence function satisfying condition (3.4) of Theorem 3.2.

Conversely, if $\varphi(Z)$ is an influence function satisfying (3.4) of Theorem 3.2, then

$$\varphi(Z) = \varphi^*(Z) + \{\varphi(Z) - \varphi^*(Z)\}.$$

It is a simple exercise to verify that $\{\varphi(Z) - \varphi^*(Z)\} \in \mathscr{T}^\perp$. □

Deriving the Efficient Influence Function

The efficient influence function $\varphi_{\text{eff}}(Z)$, if it exists, is the influence function with the smallest variance matrix; that is, for any influence function $\varphi(Z) \neq \varphi_{\text{eff}}(Z)$, $\text{var}\{\varphi_{\text{eff}}(Z)\} - \text{var}\{\varphi(Z)\}$ is negative definite. That an efficient influence function exists and is unique is now easy to see from the geometry of the problem.

Theorem 3.5. The efficient influence function is given by

$$\varphi_{\text{eff}}(Z) = \varphi^*(Z) - \Pi(\varphi^*(Z)|\mathscr{T}^\perp) = \Pi(\varphi^*(Z)|\mathscr{T}), \tag{3.34}$$

where $\varphi^*(Z)$ is an arbitrary influence function and $\mathscr{T}$ is the tangent space, and can explicitly be written as

$$\varphi_{\text{eff}}(Z) = \Gamma(\theta_0)I^{-1}(\theta_0)S_\theta(Z,\theta_0). \tag{3.35}$$

Proof. By Theorem 3.4, the class of influence functions is a linear variety, $\varphi^*(Z) + \mathscr{T}^\perp$. Let $\varphi_{\text{eff}} = \varphi^* - \Pi(\varphi^*|\mathscr{T}^\perp) = \Pi(\varphi^*|\mathscr{T})$. Because $\Pi(\varphi^*|\mathscr{T}^\perp) \in \mathscr{T}^\perp$, this implies that φ_{eff} is an influence function and, moreover, is orthogonal to $\mathscr{T}^\perp$. Consequently, any other influence function can be written as $\varphi = \varphi_{\text{eff}} + l$, with $l \in \mathscr{T}^\perp$. The tangent space $\mathscr{T}$ and its orthogonal complement $\mathscr{T}^\perp$ are examples of q-replicating linear spaces as defined by Definition 6. Therefore, because of Theorem 3.3, equation (3.31), we obtain $\text{var}(\varphi) = \text{var}(\varphi_{\text{eff}}) + \text{var}(l)$, which demonstrates that φ_{eff}, constructed as above, is the efficient influence function.

We deduce from the argument above that the efficient influence function $\varphi_{\text{eff}} = \Pi(\varphi^*|\mathscr{T})$ is an element of the tangent space $\mathscr{T}$ and hence can be expressed as $\varphi_{\text{eff}}(Z) = B_{\text{eff}}^{q\times p} S_\theta(Z,\theta_0)$ for some constant matrix $B_{\text{eff}}^{q\times p}$. Since $\varphi_{\text{eff}}(Z)$ is an influence function, it must also satisfy relationship (3.4) of Theorem 3.2; i.e.,

$$E\{\varphi_{\text{eff}}(Z)S_\theta^T(Z,\theta_0)\} = \Gamma(\theta_0),$$

or

$$B_{\text{eff}} E\{S_\theta(Z,\theta_0) S_\theta^T(Z,\theta_0)\} = \Gamma(\theta_0),$$

which implies

$$B_{\text{eff}} = \Gamma(\theta_0) I^{-1}(\theta_0),$$

where $I(\theta_0) = E\{S_\theta(Z,\theta_0) S_\theta^T(Z,\theta_0)\}$ is the information matrix. Consequently, the efficient influence function is given by

$$\varphi_{\text{eff}}(Z) = \Gamma(\theta_0) I^{-1}(\theta_0) S_\theta(Z,\theta_0). \square$$

It is instructive to consider the special case $\theta = (\beta^T, \eta^T)^T$. We first define the important notion of an efficient score vector and then show the relationship of the efficient score to the efficient influence function.

Definition 8. The *efficient score* is the residual of the score vector with respect to the parameter of interest after projecting it onto the nuisance tangent space; i.e.,

$$S_{\text{eff}}(Z,\theta_0) = S_\beta(Z,\theta_0) - \Pi(S_\beta(Z,\theta_0)|\Lambda).$$

Recall that

$$\Pi(S_\beta(Z,\theta_0)|\Lambda) = E(S_\beta S_\eta^T)\{E(S_\eta S_\eta^T)\}^{-1} S_\eta(Z,\theta_0). \square$$

Corollary 2. When the parameter θ can be partitioned as $(\beta^T, \eta^T)^T$, where β is the parameter of interest and η is the nuisance parameter, then the efficient influence function can be written as

$$\varphi_{\text{eff}}(Z,\theta_0) = \{E(S_{\text{eff}} S_{\text{eff}}^T)\}^{-1}\{S_{\text{eff}}(Z,\theta_0)\}. \tag{3.36}$$

Proof. By construction, the efficient score vector is orthogonal to the nuisance tangent space; i.e., it satisfies condition (ii) of being an influence function.

By appropriately scaling the efficient score, we can construct an influence function, which we will show is the efficient influence function. We first note that $E\{S_{\text{eff}}(Z,\theta_0) S_\beta^T(Z,\theta_0)\} = E\{S_{\text{eff}}(Z,\theta_0) S_{\text{eff}}^T(Z,\theta_0)\}$. This follows because

$$E\{S_{\text{eff}}(Z,\theta_0) S_\beta^T(Z,\theta_0)\} = E\{S_{\text{eff}}(Z,\theta_0) S_{\text{eff}}^T(Z,\theta_0)\} + \underbrace{E\{S_{\text{eff}}(Z,\theta_0)\Pi(S_\beta|\Lambda)^T\}}_{\substack{\text{This equals zero since} \\ S_{\text{eff}}(Z,\theta_0\} \perp \Lambda}}.$$

Therefore, if we define

$$\varphi_{\text{eff}}(Z,\theta_0) = \{E(S_{\text{eff}} S_{\text{eff}}^T)\}^{-1} S_{\text{eff}}(Z,\theta_0),$$

then

$$\text{(i) } E[\varphi_{\text{eff}}(Z,\theta_0) S_\beta^T(Z,\theta_0)] = I^{q\times q}$$

and

$$\text{(ii) } E[\varphi_{\text{eff}}(Z,\theta_0)S_\eta^T(Z,\theta_0)] = 0^{q\times r};$$

i.e., $\varphi_{\text{eff}}(Z,\theta_0)$ satisfies conditions (i) and (ii) of Corollary 1 and thus is an influence function.

As argued above, the efficient influence function is the unique influence function belonging to the tangent space $\mathscr{T}$. Since both $S_\beta(Z,\theta_0)$ and $\Pi(S_\beta(Z,\theta_0)|\Lambda)$ are elements of $\mathscr{T}$, so is

$$\varphi_{\text{eff}}(Z,\theta_0) = \{E(S_{\text{eff}}S_{\text{eff}}^T)\}^{-1}\{S_\beta(Z,\theta_0) - \Pi(S_\beta|\Lambda)\},$$

thus demonstrating that (3.36) is the efficient influence function for RAL estimators of β. □

Remark 7. When the parameter θ can be partitioned as $(\beta^T,\eta^T)^T$, then $\Gamma(\theta_0)$ can be partitioned as $[I^{q\times q} : 0^{q\times r}]$, and it is a straightforward exercise to show that (3.35) leads to (3.36). □

Remark 8. If we denote by $(\hat{\beta}_n^{MLE},\hat{\eta}_n^{MLE})$ the values of β and η that maximize the likelihood

$$\prod_{i=1}^n p(Z_i,\beta,\eta),$$

then under suitable regularity conditions, the estimator $\hat{\beta}_n^{MLE}$ of β is an RAL estimator whose influence function is the efficient influence function given by (3.36). See Exercise 3.2 below. □

Remark 9. If the parameter of interest is given by $\beta(\theta)$ and we define by $\hat{\theta}_n^{MLE}$ the value of θ that maximizes the likelihood

$$\prod_{i=1}^n p(Z_i,\theta),$$

then, under suitable regularity conditions, the estimator $\beta(\hat{\theta}_n^{MLE})$ of β is an RAL estimator with efficient influence function (3.35). □

Remark 10. By definition,

$$\varphi_{\text{eff}}(Z,\theta_0) = \{E(S_{\text{eff}}S_{\text{eff}}^T)\}^{-1}S_{\text{eff}}(Z,\theta_0)$$

has variance equal to

$$\{E(S_{\text{eff}}S_{\text{eff}}^T)\}^{-1},$$

the inverse of the variance matrix of the efficient score. If we define $I_{\beta\beta} = E(S_\beta S_\beta^T)$, $I_{\eta\eta} = E(S_\eta S_\eta^T)$, and $I_{\beta\eta} = E(S_\beta S_\eta^T)$, then we obtain the well-known result that the minimum variance for the most efficient RAL estimator is

$$\{I_{\beta\beta} - I_{\beta\eta}I_{\eta\eta}^{-1}I_{\beta\eta}^T\}^{-1},$$

where $I_{\beta\beta}, I_{\beta\eta}, I_{\eta\eta}$ are elements of the information matrix used in likelihood theory. □

3.5 Review of Notation for Parametric Models

We now give a quick review of some of the notation and ideas developed in Chapter 3 as a useful reference.

- $Z_1, \ldots, Z_n$ iid $p(z, \beta, \eta)$,

$$\begin{aligned} &\beta \in \mathbb{R}^q, \\ &\eta \in \mathbb{R}^r, \\ &\theta = (\beta^T, \eta^T)^T,\ \theta \in \mathbb{R}^p,\ p = q + r. \end{aligned}$$

- Truth is denoted as $\theta_0 = (\beta_0^T, \eta_0^T)^T$.
- $n^{1/2}(\hat{\beta}_n - \beta_0) = n^{-1/2}\Sigma\varphi(Z_i) + o_p(1)$, where $\varphi(Z_i)$ is the influence function for the i-th observation of $\hat{\beta}_n$.
- Hilbert space: q-dimensional measurable functions of Z with mean zero and finite variance equipped with the covariance inner product $E\{h_1^T(Z)h_2(Z)\}$.
- Score vector: For $\theta = (\beta^T, \eta^T)^T$,

$$\begin{aligned} S_\beta(z) &= \left.\frac{\partial \log p(z,\theta)}{\partial \beta}\right|_{\theta_0}, \\ S_\eta(z) &= \left.\frac{\partial \log p(z,\theta)}{\partial \eta}\right|_{\theta_0}, \\ S_\theta(z) &= \left.\frac{\partial \log p(z,\theta)}{\partial \theta}\right|_{\theta_0} = \{S_\beta^T(z), S_\eta^T(z)\}^T. \end{aligned}$$

Linear subspaces

Nuisance tangent space:

$$\Lambda = \{B^{q\times r} S_\eta : \text{for all } B^{q\times r}\}.$$

Tangent space:

$$\begin{aligned} &\mathscr{T} = \{B^{q\times p} S_\theta : \text{for all } B^{q\times p}\}, \\ &\mathscr{T} = \mathscr{T}_\beta \oplus \Lambda, \text{ where } \mathscr{T}_\beta = \{B^{q\times q} S_\beta : \text{for all } B^{q\times q}\}, \end{aligned}$$

and $\oplus$ denotes the direct sum of linear subspaces.

Influence functions φ must satisfy

(i) $E\{\varphi S_\beta^T\} = I^{q\times q}$

and

(ii) $E\{\varphi S_\eta^T\} = 0^{q\times r}; \varphi \perp \Lambda, \varphi \in \Lambda^\perp$.

- *Efficient score*

$$S_{\text{eff}}(Z,\theta_0) = S_\beta(Z,\theta_0) - \Pi(S_\beta|\Lambda);$$
$$\Pi(S_\beta|\Lambda) = E(S_\beta S_\eta^T)\{E(S_\eta S_\eta^T)\}^{-1} S_\eta(Z,\theta_0).$$

- *Efficient influence function*

$$\varphi_{\text{eff}}(Z) = \{E(S_{\text{eff}} S_{\text{eff}}^T)\}^{-1} S_{\text{eff}}(Z,\theta_0).$$

- Any influence function is equal to

$$\varphi(Z) = \varphi_{\text{eff}}(Z) + l(Z),\ l(Z) \in \mathscr{T}^\perp.$$

That is, influence functions lie on a linear variety and

- $$E(\varphi\varphi^T) = E(\varphi_{\text{eff}}\varphi_{\text{eff}}^T) + E(ll^T).$$

3.6 Exercises for Chapter 3

1. Prove that the Hodges super-efficient estimator $\hat{\mu}_n$, given in Section 3.1, is **not** asymptotically regular.
2. Let $Z_1, \ldots, Z_n$ be iid $p(z,\beta,\eta)$, where $\beta \in \mathbb{R}^q$ and $\eta \in \mathbb{R}^r$. Assume all the usual regularity conditions that allow the maximum likelihood estimator to be a solution to the score equation,

$$\sum_{i=1}^n \begin{pmatrix} S_\beta\ (Z_i,\beta,\eta) \\ S_\eta\ (Z_i,\beta,\eta) \end{pmatrix} = 0^{(q+r)\times 1},$$

and be consistent and asymptotically normal.

a) Show that the influence function for $\hat{\beta}_n$ is the efficient influence function.

b) Sketch out an argument that shows that the solution to the estimating equation

$$\sum_{i=1}^n S_{\text{eff}}^{q\times 1}\{Z_i,\beta,\hat{\eta}_n^*(\beta)\} = 0^{q\times 1},$$

for any root-n consistent estimator $\hat{\eta}_n^*(\beta)$, yields an estimator that is asymptotically linear with the efficient influence function.

3. Assume $Y_1, \ldots, Y_n$ are iid with distribution function $F(y) = P(Y \le y)$, which is differentiable everywhere with density $f(y) = \frac{dF(y)}{dy}$. The median is defined as $\beta = F^{-1}\left(\frac{1}{2}\right)$. The sample median is defined as

$$\hat{\beta}_n \approx \hat{F}_n^{-1}\left(\frac{1}{2}\right),$$

where $\hat{F}_n(y) = n^{-1} \sum_{i=1}^{n} I(Y_i \leq y)$ is the empirical distribution function. Equivalently, $\hat{\beta}_n$ is the solution to the m-estimating equation

$$\sum_{i=1}^{n} \left\{ I(Y_i \leq \beta) - \frac{1}{2} \right\} \approx 0.$$

Remark 11. We use "$\approx$" to denote approximately because the estimating equation is not continuous in β and therefore will not always yield a solution. However, for large n, you can get very close to zero, the difference being asymptotically negligible. □

(a) Find the influence function for the sample median $\hat{\beta}_n$.
Hint: You may assume the following to get your answer.
(i) $\hat{\beta}_n$ is consistent; i.e., $\hat{\beta}_n \to \beta_0 = F^{-1}\left(\frac{1}{2}\right)$.
(ii) Stochastic equicontinuity:

$$\left[n^{1/2} \left\{ \hat{F}_n(\hat{\beta}_n) - F(\hat{\beta}_n) \right\} - n^{1/2} \left\{ \hat{F}_n(\beta_0) - F(\beta_0) \right\} \right] \xrightarrow{P} 0.$$

(b) Let $Y_1, \ldots, Y_n$ be iid $N(\mu, \sigma^2), \mu \in \mathbb{R}, \ \sigma^2 > 0$. Clearly, for this model, the median β is equal to μ. Verify, by direct calculation, that the influence function for the sample median satisfies the two conditions of Corollary 1.

4

Semiparametric Models

In Chapter 3, we developed theoretical results for estimators of parameters in finite-dimensional parametric models where $Z_1, \ldots, Z_n$ are iid $\{p(z, \theta), \theta \in \Omega \subset \mathbb{R}^p\}$, p finite, and where θ can be partitioned as

$$\theta = (\beta^T, \eta^T)^T \quad \beta \in \mathbb{R}^q, \eta \in \mathbb{R}^r \quad p = q + r,$$

β being the parameter of interest and η the nuisance parameter. In this chapter, we will extend this theory to semiparametric models, where the parameter space for θ is infinite-dimensional.

For most of the exposition in this book, as well as most of the examples used throughout, we will consider semiparametric models that can be represented using the class of densities $p(z, \beta, \eta)$, where β, the parameter of interest, is finite-dimensional (q-dimensional); η, the nuisance parameter, is infinite-dimensional; and β and η are variationally independent – that is, any choice of β and η in a neighborhood about the true β_0 and η_0 would result in a density $p(z, \beta, \eta)$ in the semiparametric model. This will allow us, for example, to explicitly define partial derivatives

$$\left.\frac{\partial p(z, \beta, \eta_0)}{\partial \beta}\right|_{\beta=\beta_0} = \frac{\partial p(z, \beta_0, \eta_0)}{\partial \beta}.$$

Keep in mind, however, that some problems lend themselves more naturally to models represented by the class of densities $p(z, \theta)$, where θ is infinite-dimensional and the parameter of interest, $\beta^{q\times 1}(\theta)$, is a smooth q-dimensional function of θ. When the second representation is easier to work with, we will make the distinction explicit.

In Chapter 1, we gave two examples of semiparametric models:

(i) *Restricted moment model*

$$Y_i = \mu(X_i, \beta) + \varepsilon_i,$$
$$E(\varepsilon_i | X_i) = 0,$$

or equivalently

$$E(Y_i|X_i) = \mu(X_i, \beta).$$

(ii) *Proportional hazards model*
The hazard of failing at time t is

$$\lambda(t|X_i) = \lambda(t)\exp(\beta^T X_i).$$

The major aim is to find "good" *semiparametric estimators* for β, where "loosely speaking" a semiparametric estimator has the property that

$$n^{1/2}(\hat{\beta}_n - \beta) \xrightarrow{\mathcal{D}\{p(\cdot,\beta,\eta)\}} N\left(0, \Sigma^{q\times q}(\beta,\eta)\right)$$

for all densities "$p(\cdot, \beta, \eta)$" within some semiparametric model and "good" refers to estimators with small asymptotic variance. All of these ideas will be made precise shortly.

4.1 GEE Estimators for the Restricted Moment Model

The restricted moment model was introduced briefly in Section 1.2. In this model, we are primarily interested in studying the relationship of a response variable Y, possibly vector-valued, as a function of covariates X. Specifically, the restricted moment model considers the conditional expectation of Y given X; that is,

$$E(Y^{d\times 1}|X) = \mu^{d\times 1}(X, \beta)$$

through the function $\mu(X, \beta)$ of X and the q-dimensional parameter β. Here, "d" denotes the dimension of the response variable Y. Therefore, the restricted moment model allows the modeling of multivariate and longitudinal response data as a function of covariates (i.e., $d > 1$) as well as more traditional regression models for a univariate response variable (i.e., $d = 1$).

An example of a semiparametric estimator for the restricted moment model is the solution to the linear estimating equation

$$\sum_{i=1}^{n} A^{q\times d}(X_i, \hat{\beta}_n)\left\{Y_i^{d\times 1} - \mu^{d\times 1}(X_i, \hat{\beta}_n)\right\} = 0^{q\times 1}, \tag{4.1}$$

where $A(X_i, \beta)$ is an arbitrary $(q \times d)$ matrix of functions of the covariate X_i and the parameter β.

Subject to suitable regularity conditions, $\hat{\beta}_n$ is a consistent, asymptotically normal estimator for β and thus is an example of a semiparametric estimator for the restricted moment model. Such an estimator is an example of a solution to a generalized estimating equation, or GEE, as defined by Liang and Zeger (1986). It is also an example of an m-estimator as defined in Chapter 3. For completeness, we sketch out a heuristic argument for the asymptotic normality of $\hat{\beta}_n$ and describe how to estimate its asymptotic variance.

Asymptotic Properties for GEE Estimators

The asymptotic properties of the GEE estimator follow from the expansion

$$\begin{aligned} 0 &= \sum_{i=1}^{n} A(X_i, \hat{\beta}_n)\{Y_i - \mu(X_i, \hat{\beta}_n)\} \\ &= \sum_{i=1}^{n} A(X_i, \beta_0)\{Y_i - \mu(X_i, \beta_0)\} \\ &\quad + \left\{ \sum_{i=1}^{n} \mathcal{Q}(Y_i, X_i, \beta_n^*) - \sum_{i=1}^{n} A(X_i, \beta_n^*)\ D(X_i, \beta_n^*) \right\} (\hat{\beta}_n - \beta_0), \end{aligned} \tag{4.2}$$

where

$$D(X, \beta) = \frac{\partial \mu(X, \beta)}{\partial \beta^T} \tag{4.3}$$

is the gradient matrix $(d \times q)$, made up of all partial derivatives of the d-elements of $\mu(X, \beta)$ with respect to the q-elements of β, and β_n^* denotes some intermediate value between $\hat{\beta}_n$ and β_0.

If we denote the rows of $A(X_i, \beta)$ by $\{A_1(X_i, \beta), \ldots, A_q(X_i, \beta)\}$, then $\mathcal{Q}^{q \times q}(Y_i, X_i, \beta)$ is the $q \times q$ matrix defined by

$$\mathcal{Q}^{q \times q}(Y_i, X_i, \beta) = \begin{pmatrix} \{Y_i - \mu(X_i, \beta)\}^T \frac{\partial A_1^T(X_i, \beta)}{\partial \beta^T} \\ \vdots \\ \{Y_i - \mu(X_i, \beta)\}^T \frac{\partial A_q^T(X_i, \beta)}{\partial \beta^T} \end{pmatrix}.$$

This matrix, although complicated, is made up of a linear combination of functions of X_i multiplied by elements of $\{Y_i - \mu(X_i, \beta)\}$, which, as we will demonstrate shortly, drops out of consideration for the asymptotic theory.

Using (4.2), we obtain

$$\begin{aligned} n^{1/2}(\hat{\beta}_n - \beta_0) = \Bigg\{ & -n^{-1} \sum_{i=1}^{n} \mathcal{Q}(Y_i, X_i, \beta_n^*) \\ & + n^{-1} \sum_{i=1}^{n} A(X_i, \beta_n^*) D(X_i, \beta_n^*) \Bigg\}^{-1} n^{-1/2} \sum_{i=1}^{n} A(X_i, \beta_0)\{Y_i - \mu(X, \beta_0)\}. \end{aligned}$$

Because

$$n^{-1} \sum_{i=1}^{n} \mathcal{Q}(Y_i, X_i, \beta_n^*) \xrightarrow{P} E\left\{\mathcal{Q}(Y, X, \beta_0)\right\} = 0$$

and

$$n^{-1} \sum_{i=1}^{n} A(X_i, \beta_n^*) D(X_i, \beta_n^*) \xrightarrow{P} E\{A(X, \beta_0) D(X, \beta_0)\},$$

we obtain

$$n^{1/2}(\hat{\beta}_n - \beta_0) = n^{-1/2} \sum_{i=1}^{n} \Bigg(\left[E\{A(X, \beta_0) D(X, \beta_0)\} \right]^{-1} A(X_i, \beta_0) \times \{Y_i - \mu(X_i, \beta_0)\} \Bigg) + o_p(1).$$

Consequently, the influence function for the i-th observation of $\hat{\beta}_n$ is

$$\{E(AD)\}^{-1} A(X_i, \beta_0)\{Y_i - \mu(X_i, \beta_0)\}. \tag{4.4}$$

As we demonstrated in Chapter 3, the asymptotic variance of an RAL estimator for β is the variance of its influence function. Therefore, the asymptotic variance of the GEE estimator $\hat{\beta}_n$ is the variance of (4.4). We first compute the variance of $A(X_i, \beta_0)\{Y_i - \mu(X_i, \beta_0)\}$, which equals

$$\begin{aligned} \text{var}\Big[A(X_i, \beta_0)\{Y_i - \mu(X_i, \beta_0)\} \Big] &= E\Big(\text{var}\left[A(X_i, \beta_0)\{Y_i - \mu(X_i, \beta_0)\} | X_i \right] \Big) \\ &\quad + \text{var}\Bigg(\underbrace{E\left[A(X_i, \beta_0)\{Y_i - \mu(X_i, \beta_0)\} | X_i \right]}_{\| \\ 0} \Bigg) \\ &= E\{A(X_i, \beta_0) V(X_i) A^T(X_i, \beta_0)\}, \end{aligned} \tag{4.5}$$

where $V(X_i) = \text{var}(Y_i | X_i)$ is the $d \times d$ conditional variance matrix of Y_i given X_i. Consequently, the asymptotic variance of $\hat{\beta}_n$ is

$$\{E(AD)\}^{-1} \; E\{AV(X)A^T\} \; \{E(AD)\}^{-1^T}. \tag{4.6}$$

In order to use the results above for data analytic applications, such as constructing confidence intervals for β or for some components of β, we must also be able to derive consistent estimators for the asymptotic variance of $\hat{\beta}_n$ given by (4.6). Without going into all the technical details, we now outline the arguments for constructing such an estimator. These arguments are similar to those that resulted in the sandwich variance estimator for the asymptotic variance of an m-estimator given by (3.10) in Chapter 3.

Suppose, for the time being, that we assumed the true-valued β_0 were known to us. Then, by the law of large numbers, a consistent estimator for $E(AD)$ is given by

$$\hat{E}_0(AD) = n^{-1} \sum_{i=1}^{n} A(X_i, \beta_0) D(X_i, \beta_0), \tag{4.7}$$

where the subscript "0" is used to emphasize that this statistic is computed with β_0 known. As we showed in (4.5), the variance of $A(X_i, \beta_0)\{Y_i -$

$\mu(X_i, \beta_0)\}$ is given by $E\{A(X_i, \beta_0)V(X_i)A^T(X_i, \beta_0)\}$, which, by the law of large numbers, can be consistently estimated by

$$\hat{E}_0(AVA^T) = n^{-1}\sum_{i=1}^{n} A(X_i, \beta_0)\{Y_i - \mu(X_i, \beta_0)\}\{Y_i - \mu(X_i, \beta_0)\}^T A^T(X_i, \beta_0). \tag{4.8}$$

Of course, the value β_0 is not known to us. But since $\hat{\beta}_n$ is a consistent estimator for β_0, a natural estimator for the asymptotic variance of $\hat{\beta}_n$ is given by

$$\{\hat{E}(AD)\}^{-1}\hat{E}(AVA^T)\{\hat{E}(AD)\}^{-1^T}, \tag{4.9}$$

where $\hat{E}(AD)$ and $\hat{E}(AVA^T)$ are computed as in equations (4.7) and (4.8), respectively, with $\hat{\beta}_n$ substituted for β_0. The estimator (4.9) is referred to as the sandwich estimator for the asymptotic variance. More details about this methodology can be found in Liang and Zeger (1986).

The results above did not depend on any specific parametric assumptions beyond the moment restriction and regularity conditions. Consequently, the estimator, given as the solution to equation (4.1), is a semiparametric estimator for the restricted moment model.

Example: Log-linear Model

Consider the problem where we want to model the relationship of a response variable Y, which is positive, as a function of covariates X. For example, in the study of HIV disease, CD4 count is a measure of the degree of destruction that HIV disease has on the immune system. Therefore, it may be of interest to model CD4 count as a function of covariates such as treatment, age, race, etc. Let us denote by $X = (X_1, \ldots, X_{q-1})^T$ the $q-1$ vector of covariates that we are considering. Because CD4 count is a positive random variable, a popular model is the log-linear model where it is assumed that

$$\log\{E(Y|X)\} = \alpha + \delta_1 X_1 + \ldots + \delta_{q-1}X_{q-1}.$$

Here, the parameter of interest is given by $\beta^{q\times 1} = (\alpha, \delta_1, \ldots, \delta_{q-1})^T$.

This is an example of a restricted moment model where

$$E(Y|X) = \mu(X, \beta) = \exp(\alpha + \delta_1 X_1 + \ldots + \delta_{q-1}X_{q-1}). \tag{4.10}$$

The log transformation guarantees that the conditional mean response, given the covariates, is always positive. Consequently, this model puts no restrictions on the possible values that β can take.

With a sample of iid data (Y_i, X_i), $i = 1, \ldots, n$, a semiparametric estimator for β can be obtained as the solution to a generalized estimating equation given by (4.1). In this example, the response variable Y is a single random variable; hence $d = 1$. If we take $A^{q\times 1}(X, \beta)$ in (4.1) to equal $(1, X_1, \ldots, X_{q-1})^T$,

then the corresponding GEE estimator $\hat{\beta}_n = (\hat{\alpha}_n, \ldots, \hat{\delta}_{(q-1)n})^T$ is the solution to

$$\sum_{i=1}^{n}(1, X_i^T)^T\{Y_i - \exp(\alpha + \delta_1 X_{1i} + \ldots + \delta_{q-1}X_{(q-1)i})\} = 0^{q\times 1}. \tag{4.11}$$

The estimator $\hat{\beta}_n$ is consistent and asymptotically normal with a variance matrix that can be estimated using (4.9), where the derivative matrix $D^{1\times q}(X, \beta)$, defined by (4.3), is equal to $\mu(X, \beta)(1, X^T)$, and $\hat{E}(AD)$ and $\hat{E}(AVA^T)$ are given by equations (4.7) and (4.8), respectively, with $\hat{\beta}_n$ substituted for β_0. Specifically,

$$\hat{E}(AD) = n^{-1}\sum_{i=1}^{n}(1, X_i^T)^T\mu(X_i, \hat{\beta}_n)(1, X_i^T), \tag{4.12}$$

$$\hat{E}(AVA^T) = n^{-1}\sum_{i=1}^{n}(1, X_i^T)^T\{Y_i - \mu(X_i, \hat{\beta}_n)\}^2(1, X_i^T). \tag{4.13}$$

Remark 1. The asymptotic variance is the variance matrix of the limiting normal distribution to which $n^{1/2}(\hat{\beta}_n - \beta_0)$ converges. That is, the asymptotic variance is equal to the $(q \times q)$ matrix Σ, where $n^{1/2}(\hat{\beta}_n - \beta_0) \xrightarrow{\mathcal{D}} N(0, \Sigma)$. The estimator for the asymptotic variance is denoted by $\hat{\Sigma}_n$ and is given by

$$\hat{\Sigma}_n = \{\hat{E}(AD)\}^{-1}\hat{E}(AVA^T)\{\hat{E}(AD)\}^{-1}, \tag{4.14}$$

where $\hat{E}(AD)$ and $\hat{E}(AVA^T)$ are defined by (4.12) and (4.13), respectively.

For practical applications, say, when we are constructing confidence intervals for δ_j, the regression coefficient for the j-th covariate X_j, $j = 1, \ldots, q-1$, we must be careful to use the appropriate scaling factor when computing the estimated standard error for $\hat{\delta}_{jn}$. That is, the 95% confidence interval for δ_j is given by

$$\hat{\delta}_{jn} \pm 1.96\text{se}(\hat{\delta}_{jn}),$$

and $\text{se}(\hat{\delta}_{jn}) = n^{-1}(\hat{\Sigma}_n)_{(j+1)(j+1)}$, where $(\cdot)^{q\times q}_{(j+1)(j+1)}$ denotes the $(j+1)$-th diagonal element of the $q \times q$ matrix $(\cdot)^{q\times q}$. □

The GEE estimator for β in the log-linear model, given as the solution to equation (4.11), is just one example of many possible semiparametric estimators. Clearly, we would be interested in finding a semiparametric estimator that is as efficient as possible; i.e., with as small an asymptotic variance as possible. We will address this issue later in this chapter.

Some natural questions that arise for semiparametric models are:

(i) How do we find semiparametric estimators, or do they even exist?
(ii) How do we find the best semiparametric estimator?

Although both of these questions are difficult to resolve in general, understanding the geometry of influence functions for semiparametric estimators is often helpful in constructing estimators and assessing efficiency. The ideas and geometry we developed in Chapter 3 for finite-dimensional parametric models will now be generalized to semiparametric models.

4.2 Parametric Submodels

As is often the case in mathematics, infinite-dimensional problems are tackled by first working with a finite-dimensional problem as an approximation and then taking limits to infinity. Therefore, the first step in dealing with a semiparametric model is to consider a simpler finite-dimensional parametric model contained within the semiparametric model and use the theory and methods developed in Chapter 3. Toward that end, we define a *parametric submodel*.

Recall: In a semiparametric model, the data $Z_1, \ldots, Z_n$ are iid random vectors with a density that belongs to the class

$$\mathscr{P} = \Big[p\{z, \beta, \eta(\cdot)\}, \text{ where } \beta \text{ is } q\text{-dimensional and } \eta(\cdot) \text{ is infinite-dimensional} \Big]$$

with respect to some dominating measure ν_Z. As we illustrated in some of the examples of semiparametric models in Chapter 1, the infinite-dimensional nuisance parameter η is itself often a function and hence denoted as $\eta(\cdot)$. □

We will denote the "truth" (i.e., the density that generates the data) by $p_0(z) \in \mathscr{P}$, namely

$$p_0(z) = p\{z, \beta_0, \eta_0(\cdot)\}.$$

A parametric submodel, which we will denote by $\mathscr{P}_{\beta,\gamma} = \{p(z, \beta, \gamma)\}$, is a class of densities characterized by the finite-dimensional parameter $(\beta^T, \gamma^T)^T$ such that

(i) $\mathscr{P}_{\beta,\gamma} \subset \mathscr{P}$ (i.e., every density in $\mathscr{P}_{\beta,\gamma}$ belongs to the semiparametric model $\mathscr{P}$) and

(ii) $p_0(z) \in \mathscr{P}_{\beta,\gamma}$ (i.e., the parametric submodel contains the truth). Another way of saying this is that there exists a density identified by the parameter (β_0, γ_0) within the parametric submodel such that

$$p_0(z) = p(z, \beta_0, \gamma_0).$$

In keeping with the notation of Chapter 3, we will denote the dimension of γ by "r," although, in this case, the value r depends on the choice of parametric submodel.

Remark 2. When we developed the geometry of influence functions for parametric models in Chapter 3, certain regularity conditions were implicitly assumed. For example, the parametric model had to have sufficient regularity

conditions to allow the interchange of differentiation and integration of the density with respect to the parameters. This is necessary, for example, when we want to prove that the score vector has mean zero. Consequently, the parametric model has to satisfy certain smoothness conditions. Similarly, the parametric submodels that we will consider must also satisfy certain smoothness conditions. Appropriate smoothness and regularity conditions on the likelihoods are given in Definition A.1 of the appendix in Newey (1990). Thus, from here on, when we refer to parametric submodels, we implicitly are assuming smooth and regular parametric submodels. □

Remark 3. The terms *parametric submodel* and *parametric model* can be confusing. A parametric model is a model whose probability densities are characterized through a finite number of parameters that the data analyst believes will suffice in identifying the probability distribution that generates the data. For example, we may be willing to assume that our data follow the model

$$Y_i = \mu(X_i, \beta) + \varepsilon_i, \tag{4.15}$$

where ε_i are iid $N(0, \sigma^2)$, independent of X_i. This model is contained within the semiparametric restricted moment model discussed previously.

In contrast, a *parametric submodel* is a conceptual idea that is used to help us develop theory for semiparametric models. The reason we say it is conceptual is that we require a *parametric submodel* to contain the truth. But since we don't know what the truth is, we can only describe such submodels generically and hence such models are not useful for data analysis. The parametric model given by (4.15) is not a parametric submodel if, in truth, the data are not normally distributed. □

We now illustrate how a parametric submodel can be defined using the proportional hazards model as an example. In the proportional hazards model, we assume

$$\lambda(t|X) = \lambda(t) \exp(\beta^T X),$$

where $X = (X_1, \ldots, X_q)^T$ denotes a q-dimensional vector of covariates, $\lambda(t)$ is some arbitrary hazard function of time that is left unspecified and hence is infinite-dimensional, and β is the q-dimensional parameter of interest. We denote (conceptually) the truth by $\lambda_0(t);\ t \geq 0$ and β_0.

An example of a parametric submodel is as follows. Let $h_1(t), \ldots, h_r(t)$ be r different functions of time that are specified by the data analyst. (Any smooth functions will do.) Consider the model

$$\begin{aligned}\mathscr{P}_{\beta,\gamma} = \{&\text{class of densities with hazard function} \\ &\lambda(t|X) = \lambda_0(t) \exp\{\gamma_1 h_1(t) + \cdots + \gamma_r h_r(t)\} \exp(\beta^T X)\},\end{aligned}$$

where $\gamma = (\gamma_1, \ldots, \gamma_r)^T \in \mathbb{R}^r$ and $\beta \in \mathbb{R}^q$.

We note that:

- In this model, the $(q+r)$ parameters $(\beta^T, \gamma^T)^T$ are left unspecified. Hence, this model is indeed a finite-dimensional model.
- For any choice of β and γ, the resulting density follows a proportional hazards model and is therefore contained in the semiparametric model; i.e.,
$$\mathscr{P}_{\beta,\gamma} \subset \mathscr{P}.$$
- The truth is obtained by setting $\beta = \beta_0$ and $\gamma = 0$.
- This parametric submodel is defined using $\lambda_0(t)$, "the truth," which is not known to us; consequently, such a model is not useful for data analysis.

Contrast this with the case where we are willing to consider the parametric model; namely
$$\lambda(t|X) = \lambda \exp(\beta^T X), \ \lambda, \beta \text{ unknown}.$$
That is, we assume that the underlying baseline hazard function is constant over time; i.e., conditional on X, the survival distribution follows an exponential distribution. If we are willing to assume that our data are generated from some distribution within this parametric model, then we only need to estimate the parameters λ and β and use this for any subsequent data analysis. Of course, the disadvantage of such a parametric model is that if the data are not generated from any density within this class, then the estimates we obtain may be meaningless.

4.3 Influence Functions for Semiparametric RAL Estimators

In Chapter 3, we studied RAL estimators for β for finite-dimensional parametric models and derived their asymptotic properties through their influence function. We also described the geometry of the class of influence functions for RAL estimators. Consequently, from this development, we know that influence functions of RAL estimators for β for a parametric submodel:

(i) Belong to the subspace of the Hilbert space $\mathcal{H}$ of q-dimensional mean-zero finite-variance measurable functions (equipped with the covariance inner product) that are orthogonal to the parametric submodel nuisance tangent space
$$\Lambda_\gamma = \{B^{q\times r} S_\gamma(Z, \beta_0, \gamma_0), \text{ for all } B^{q\times r}\},$$
where
$$S_\gamma^{r\times 1} = \frac{\partial \log p(z, \beta_0, \gamma_0)}{\partial \gamma}.$$

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

$$\varphi^{\text{eff}}_{\beta,\gamma}(Z) = \{E(S^{\text{eff}}_{\beta,\gamma} S^{\text{eff}^T}_{\beta,\gamma})\}^{-1} S^{\text{eff}}_{\beta,\gamma}(Z, \beta_0, \gamma_0),$$

where $S^{\text{eff}}_{\beta,\gamma}(Z, \beta_0, \gamma_0)$, the parametric submodel efficient score, is

$$S_\beta(Z, \beta_0, \eta_0) - \Pi(S_\beta(Z, \beta_0, \eta_0)|\Lambda_\gamma),$$

and

$$S^{q\times 1}_\beta(Z, \beta_0, \eta_0) = \frac{\partial \log p(z, \beta_0, \eta_0)}{\partial \beta}.$$

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

$$[E\{S^{\text{eff}}_{\beta,\gamma}(Z) S^{\text{eff}^T}_{\beta,\gamma}(Z)\}]^{-1}.$$

An estimator for β is an RAL estimator for a semiparametric model if it is an RAL estimator for every parametric submodel. Therefore, any influence function of an RAL estimator in a semiparametric model must be an influence function of an RAL estimator within a parametric submodel; i.e.,

$$\begin{bmatrix}\text{class of influence functions}\\ \text{of RAL estimators for } \beta \text{ for } \mathscr{P}\end{bmatrix} \subset \begin{bmatrix}\text{class of influence functions}\\ \text{of RAL estimators for } \mathscr{P}_{\beta,\gamma}\end{bmatrix}.$$

A heuristic way of looking at this is as follows. If $\hat{\beta}_n$ is a semiparametric estimator, then we want

$$n^{1/2}(\hat{\beta}_n - \beta) \xrightarrow{D(\beta,\eta)} N\Big(0, \Sigma(\beta, \eta)\Big)$$

for all $p(z, \beta, \eta) \in \mathscr{P}$. Such an estimator would necessarily satisfy

$$n^{1/2}(\hat{\beta}_n - \beta) \xrightarrow{D(\beta,\gamma)} N\Big(0, \Sigma(\beta, \gamma)\Big)$$

for all $p(z; \beta, \gamma) \in \mathscr{P}_{\beta,\gamma} \subset \mathscr{P}$. However, the converse may not be true. Consequently, the class of semiparametric estimators must be contained within the class of estimators for a parametric submodel. Therefore:

(i) Any influence function of an RAL semiparametric estimator for β must be orthogonal to all parametric submodel nuisance tangent spaces.
(ii) The variance of any RAL semiparametric influence function must be greater than or equal to

$$[E\{S^{\text{eff}}_{\beta,\gamma}(Z) S^{\text{eff}^T}_{\beta,\gamma}(Z)\}]^{-1}$$

for all parametric submodels $\mathscr{P}_{\beta,\gamma}$.

Hence, the variance of the influence function for any semiparametric estimator for β must be greater than or equal to

$$\sup_{\{\text{all parametric submodels}\}} \left\{ E\left(S^{\text{eff}}_{\beta,\gamma} S^{\text{eff}^T}_{\beta,\gamma} \right) \right\}^{-1}. \tag{4.16}$$

This *supremum* is defined to be the *semiparametric efficiency bound*. Any semiparametric RAL estimator $\hat{\beta}_n$ with asymptotic variance achieving this bound for $p_0(z) = p(z, \beta_0, \eta_0)$ is said to be *locally efficient* at $p_0(\cdot)$. If the same estimator $\hat{\beta}_n$ is semiparametric efficient regardless of $p_0(\cdot) \in \mathscr{P}$, then we say that such an estimator is *globally semiparametric efficient*.

Geometrically, the parametric submodel efficient score is the residual of $S_\beta(Z, \beta_0, \eta_0)$ after projecting it onto the parametric submodel nuisance tangent space. For a one-dimensional parameter β, the inverse of the norm-squared of this residual is the smallest variance of all influence functions for RAL estimators for the parametric submodel. This analogy can be extended to q-dimensional β as well by considering the inverse of the variance matrix of the residual q-dimensional vector.

As we increase the complexity of the parametric submodel, or consider the linear space spanned by the nuisance tangent spaces of all the parametric submodels, the corresponding space becomes larger and therefore the norm of the residual becomes smaller. Hence, the inverse of the variance of the residual grows larger. This gives a geometric perspective to the observation that the efficient semiparametric estimator has a variance larger than the efficient estimator for any parametric submodel.

4.4 Semiparametric Nuisance Tangent Space

Let us be a bit more formal.

Definition 1. The *nuisance tangent space for a semiparametric model*, denoted by Λ, is defined as the mean-square closure of parametric submodel nuisance tangent spaces, where a parametric submodel nuisance tangent space is the set of elements

$$\{B^{q\times r} S_\gamma^{r\times 1}(Z, \beta_0, \eta_0)\},$$

$S_\gamma(Z, \beta_0, \eta_0)$ is the score vector for the nuisance parameter γ for some parametric submodel, and $B^{q\times r}$ is a conformable matrix with q-rows. Specifically, the mean-square closure of the spaces above is defined as the space $\Lambda \subset \mathcal{H}$, where $\Lambda = [h^{q\times 1}(Z) \in \mathcal{H}$ such that $E\{h^T(Z)h(Z)\} < \infty$ and there exists a sequence $B_j S_{\gamma j}(Z)$ such that

$$\|h(Z) - B_j S_{\gamma j}(Z)\|^2 \xrightarrow{j\to\infty} 0$$

for a sequence of parametric submodels indexed by $j]$, where $\|h(Z)\|^2 = E\{h^T(Z)h(Z)\}$. □

Remark 4. The Hilbert space $\mathcal{H}$ is also a metric space (i.e., a set of elements where a notion of distance between elements of the set is defined). For any two elements $h_1, h_2 \in \mathcal{H}$, we can define the distance between two elements $h_1, h_2 \in \mathcal{H}$ as $\|h_2 - h_1\| = [E\{(h_2 - h_1)^T(h_2 - h_1)\}]^{1/2}$. The closure of a set $\mathcal{S}$, where, in this setting, a set consists of q-dimensional random functions with mean zero and finite variance, is defined as the smallest closed set that contains $\mathcal{S}$, or equivalently, as the set of all elements in $\mathcal{S}$ together with all the limit points of $\mathcal{S}$. The closure of $\mathcal{S}$ is denoted by $\bar{\mathcal{S}}$. Thus the closure of a set is itself a closed set. Limits must be defined in terms of a distance between elements. The word mean-square is used because limits are taken with respect to the distance, which in this case is the square root of the expected sum of squared differences between the q-components of the two elements (i.e., between the two q-dimensional random functions). Therefore, the mean-square closure is larger and contains the union of all parametric submodel nuisance tangent spaces. Therefore, if we denote by $\mathcal{S}$ the union of all parametric submodel nuisance tangent spaces, then $\Lambda = \bar{\mathcal{S}}$ is the semiparametric nuisance tangent space. □

Remark 5. Although the space Λ is closed, it may not necessarily be a linear space. However, in most applications it is a linear space, and certainly is in any of the examples used in this book. Therefore, from now on, we will always assume the space Λ is a linear space and, by construction, is also a closed space. This is important because in order for the projection theorem to be guaranteed to apply (i.e., that a unique projection of an element to a linear subspace exists), that linear subspace must be a closed linear subspace of the Hilbert space. □

Before deriving the semiparametric efficient influence function, we first define the semiparametric efficient score vector and give some results regarding the semiparametric efficiency bound.

Definition 2. The semiparametric efficient score for β is defined as

$$S_{\text{eff}}(Z, \beta_0, \eta_0) = S_\beta(Z, \beta_0, \eta_0) - \Pi\{S_\beta(Z, \beta_0, \eta_0)|\Lambda\}.$$

There is no difficulty with this definition since the nuisance tangent space Λ is a closed linear subspace and therefore the projection $\Pi\{S_\beta(Z, \beta_0, \eta_0)|\Lambda\}$ exists and is unique. □

Theorem 4.1. The semiparametric efficiency bound, defined by (4.16), is equal to the inverse of the variance matrix of the semiparametric efficient score; i.e.,

$$[E\{S_{\text{eff}}(Z)S_{\text{eff}}^T(Z)\}]^{-1}.$$

Proof. For simplicity, we take β to be a scalar (i.e., $q = 1$), although this can be extended to $q > 1$ using arguments in Section 3.4, where a generalization of

the Pythagorean theorem to dimension $q > 1$ was derived (see (3.31)). Denote by $\mathcal{V}$ the semiparametric efficiency bound, which, when $q = 1$, is defined by

$$\sup_{\substack{\{\text{all parametric submodels}\} \\ \mathscr{P}_{\beta,\gamma}}} \|S^{\text{eff}}_{\beta,\gamma}(Z)\|^{-2} = \mathcal{V},$$

where

$$S^{\text{eff}}_{\beta,\gamma}(Z) = S_\beta(Z) - \Pi(S_\beta(Z)|\Lambda_\gamma).$$

Since $\Lambda_\gamma \subset \Lambda$, this implies that $\|S_{\text{eff}}(Z)\| \leq \|S^{\text{eff}}_{\beta,\gamma}(Z)\|$ for all parametric submodels $\mathscr{P}_{\beta,\gamma}$. Hence

$$\|S_{\text{eff}}(Z)\|^{-2} \geq \sup_{\{\text{all } \mathscr{P}_{\beta,\gamma}\}} \|S^{\text{eff}}_{\beta,\gamma}(Z)\|^{-2} = \mathcal{V}. \tag{4.17}$$

To complete the proof of the theorem, we need to show that $\|S_{\text{eff}}(Z)\|^{-2}$ is also less than or equal to $\mathcal{V}$. But because $\Pi(S_\beta(Z)|\Lambda) \in \Lambda$, this means that there exists a sequence of parametric submodels $\mathscr{P}_{\beta,\gamma_j}$ with nuisance score vectors $S_{\gamma_j}(Z)$ such that

$$\|\Pi(S_\beta(Z)|\Lambda) - B_j S_{\gamma_j}(Z)\|^2 \xrightarrow{j\to\infty} 0$$

for conformable matrices B_j.

By the definition of $\mathcal{V}$, for any $\mathscr{P}_{\beta,\gamma_j}$, we obtain $\mathcal{V}^{-1} \leq \|S^{\text{eff}}_{\beta,\gamma_j}(Z)\|^2$. Therefore

$$\begin{aligned}\mathcal{V}^{-1} &\leq \|S^{\text{eff}}_{\beta,\gamma_j}(Z)\|^2 = \|S_\beta(Z) - \Pi(S_\beta(Z)|\Lambda_{\gamma_j})\|^2 \leq \|S_\beta(Z) - B_j S_{\gamma_j}(Z)\|^2 \\ &= \|S_\beta(Z) - \Pi(S_\beta(Z)|\Lambda)\|^2 + \|\Pi(S_\beta(Z)|\Lambda) - B_j S_{\gamma_j}(Z)\|^2. \end{aligned} \tag{4.18}$$

Because $S_\beta(Z) - \Pi(S_\beta(Z)|\Lambda)$ is orthogonal to Λ and $\Pi(S_\beta(Z)|\Lambda) - B_j S_{\gamma_j}(Z)$ is an element of Λ, the last equality in (4.18) follows from the Pythagorean theorem. Taking $j \to \infty$ implies

$$\|S_\beta(Z) - \Pi(S_\beta(Z)|\Lambda)\|^2 = \|S_{\text{eff}}(Z)\|^2 \geq \mathcal{V}^{-1}$$

or

$$\|S_{\text{eff}}(Z)\|^{-2} \leq \mathcal{V}.$$

Together with (4.17), we conclude that $\|S_{\text{eff}}(Z)\|^{-2} = \mathcal{V}$. □

Definition 3. The efficient influence function is defined as the influence function of a semiparametric RAL estimator, if it exists (see remark below), that achieves the semiparametric efficiency bound. □

Remark 6. In the development that follows, we will construct a unique element of $\mathcal{H}$ that always exists, which will be defined as the efficient influence function. We will prove that if a semiparametric RAL estimator exists that has an influence function whose variance is the semiparametric efficiency bound, then this influence function must be the efficient influence function. There is no guarantee, however, that such an RAL estimator can be derived. □

Theorem 4.2. Any semiparametric RAL estimator for β must have an influence function $\varphi(Z)$ that satisfies

(i) $E\{\varphi(Z)S_\beta^T(Z,\beta_0,\eta_0)\} = E\left\{\varphi(Z)S_{\text{eff}}^T(Z,\beta_0,\eta_0)\right\} = I^{q\times q}$

and

(ii) $\Pi\{\varphi(Z)|\Lambda\} = 0$; i.e., $\varphi(Z)$ is orthogonal to the nuisance tangent space.

The efficient influence function is now defined as the unique element satisfying conditions (i) and (ii) whose variance matrix equals the efficiency bound and is equal to

$$\varphi_{\text{eff}}(Z,\beta_0,\eta_0) = \left\{E\left(S_{\text{eff}}S_{\text{eff}}^T\right)\right\}^{-1} S_{\text{eff}}(Z,\beta_0,\eta_0).$$

Proof. We first prove condition (ii). To show that $\varphi(Z)$ is orthogonal to Λ, we must prove that $\langle\varphi,h\rangle = 0$ for all $h \in \Lambda$. By the definition of Λ, there exists a sequence $B_jS_{\gamma j}(Z)$ such that

$$\|h(Z) - B_jS_{\gamma j}(Z)\| \xrightarrow{j\to\infty} 0$$

for a sequence of parametric submodels indexed by j. Hence

$$\langle\varphi,h\rangle = \langle\varphi,B_jS_{\gamma j}\rangle + \langle\varphi,h - B_jS_{\gamma j}\rangle.$$

Because any influence function of a semiparametric RAL estimator for β must be an influence function for an RAL estimator in a parametric submodel, then condition (ii) of Corollary 1 of Theorem 3.2 implies that φ is orthogonal to Λ_{γ_j} and hence the first term in the sum above is equal to zero. By the Cauchy-Schwartz inequality, we obtain

$$|\langle\varphi,h\rangle| \le \|\varphi\|\|h - B_jS_{\gamma j}\|.$$

Taking limits as $j\to\infty$ gives us the desired result.

To prove condition (i) above, we note that by condition (i) of Corollary 1 of Theorem 3.2, $\varphi(Z)$ must satisfy $E\{\varphi(Z)S_\beta^T(Z,\beta_0,\eta_0)\} = I^{q\times q}$. Since $E\{\varphi(Z)S_{\text{eff}}^T(Z,\beta_0,\eta_0)\} = E\{\varphi(Z)S_\beta^T(Z,\beta_0,\eta_0)\} - E\{\varphi(Z)\Pi(S_\beta^T(Z,\beta_0,\eta_0)|\Lambda)\}$, the result follows because $\Pi(S_\beta^T(Z,\beta_0,\eta_0)|\Lambda) \in \Lambda$ and since, by condition (ii), $\varphi(Z)$ is orthogonal to Λ, this implies that $E\{\varphi(Z)\Pi(S_\beta^T(Z,\beta_0,\eta_0)|\Lambda)\} = 0$.

It is now easy to show that the efficient influence function is given by

$$\varphi_{\text{eff}}(Z,\beta_0,\eta_0) = \left\{E\left(S_{\text{eff}}S_{\text{eff}}^T\right)\right\}^{-1} S_{\text{eff}}(Z,\beta_0,\eta_0).$$

Clearly, $\varphi_{\text{eff}}(Z, \beta_0, \eta_0)$ satisfies conditions (i) and (ii) above and moreover has variance matrix $E\{\varphi_{\text{eff}}(Z)\varphi_{\text{eff}}^T(Z)\} = \mathcal{V}$, where $\mathcal{V}$ is the semiparametric efficiency bound. □

The results of Theorem 4.2 apply specifically to semiparametric models that can be parametrized through (β, η), where β is the parameter of interest, η is the infinite-dimensional nuisance parameter, and for which the nuisance tangent space Λ can be readily computed. For some semiparametric models, it may be more convenient to parametrize the model through an infinite-dimensional parameter θ for which the tangent space can be readily computed and define the parameter of interest as $\beta(\theta)$; i.e., a smooth q-dimensional function of θ.

For parametric models, we showed in Chapter 3, Theorem 3.4, that the space of influence functions for RAL estimators for β lies on a linear variety defined as $\varphi(Z) + \mathcal{T}_\theta^\perp$, where $\varphi(Z)$ is the influence function of any RAL estimator for β and $\mathcal{T}_\theta$ is the parametric model tangent space, the space spanned by the score vector $S_\theta(Z)$. Similarly, for semiparametric models, we could show that the influence functions of semiparametric RAL estimators for β must lie on the linear variety $\varphi(Z) + \mathcal{T}^\perp$, where $\varphi(Z)$ is the influence function of any semiparametric RAL estimator for β and $\mathcal{T}$ is the semiparametric tangent space, defined as the mean-square closure of all parametric submodel tangent spaces. The element in this linear variety with the smallest norm is the unique element $\varphi(Z) - \Pi\{\varphi(Z)|\mathcal{T}^\perp\} = \Pi\{\varphi(Z)|\mathcal{T}\}$, which can be shown to be the efficient influence function $\varphi_{\text{eff}}(Z)$.

We give these results in the following theorem, whose proof is analogous to that in Theorems 3.4 and 3.5 and is therefore omitted.

Theorem 4.3. If a semiparametric RAL estimator for β exists, then the influence function of this estimator must belong to the space of influence functions, the linear variety $\left\{\varphi(Z) + \mathcal{T}^\perp\right\}$, where $\varphi(Z)$ is the influence function of any semiparametric RAL estimator for β and $\mathcal{T}$ is the semiparametric tangent space, and if an RAL estimator for β exists that achieves the semiparametric efficiency bound (i.e., a semiparametric efficient estimator), then the influence function of this estimator must be the unique and well-defined element

$$\varphi_{\text{eff}}(Z) = \varphi(Z) - \Pi\{\varphi(Z)|\mathcal{T}^\perp\} = \Pi\{\varphi(Z)|\mathcal{T}\}.$$

What is not clear is whether there exist semiparametric estimators that will have influence functions corresponding to the elements of the Hilbert space satisfying conditions (i) and (ii) of Theorem 4.2 or Theorem 4.3 (although we might expect that arguments similar to those in Section 3.3, used to construct estimators for finite-dimensional parametric models, will extend to semiparametric models as well).

In many cases, deriving the space of influence functions, or even the space orthogonal to the nuisance tangent space, for semiparametric models, will

suggest how semiparametric estimators may be constructed and even how to find locally or globally efficient semiparametric estimators. We will illustrate this for the semiparametric restricted moment model. But before doing so, it will be instructive to develop some methods and tools for finding infinite-dimensional tangent spaces. We start with the nonparametric model, where we put no restrictions on the class of densities and show that the corresponding tangent space is the entire Hilbert space $\mathcal{H}$.

Tangent Space for Nonparametric Models

Suppose we are interested in estimating some q-dimensional parameter β for a nonparametric model. That is, let $Z_1, \ldots, Z_n$ be iid random vectors with arbitrary density $p(z)$ with respect to a dominating measure ν_Z, where the only restriction on $p(z)$ is that $p(z) \geq 0$ and

$$\int p(z) d\nu(z) = 1.$$

Theorem 4.4. The tangent space (i.e., the mean-square closure of all parametric submodel tangent spaces) is the entire Hilbert space $\mathcal{H}$.

Proof. Consider any parametric submodel $\mathscr{P}_\theta = \{p(z,\theta), \theta, \text{say } s\text{-dimensional}\}$. The parametric submodel tangent space is

$$\Lambda_\theta = \left\{ B^{q\times s} S_\theta(Z), \text{ for all constant matrices } B^{q\times s} \right\},$$

where

$$S_\theta(z) = \frac{\partial \log p(z,\theta_0)}{\partial \theta}.$$

Denote the truth as $p_0(z) = p(z,\theta_0)$. From the usual properties of score vectors, we know that

$$E\{S_\theta(Z)\} = 0^{s\times 1}.$$

Consequently, the linear subspace $\Lambda_\theta \subset \mathcal{H}$.

Reminder: When we write $E\{S_\theta(Z)\}$, we implicitly mean that the expectation is computed with respect to the truth; i.e.,

$$E_0\{S_\theta(Z,\theta_0)\} \quad \text{or} \quad E_{\theta_0}\{S_\theta(Z,\theta_0)\}. \ \square$$

To complete the proof, we need to show that any element of $\mathcal{H}$ can be written as an element of Λ_θ for some parametric submodel or a limit of such elements. Choose an arbitrary element of $\mathcal{H}$, say $h(Z)$ that is a bounded mean-zero q-dimensional measurable function with finite variance. Consider the parametric submodel $p(z,\theta) = p_0(z)\{1 + \theta^T h(z)\}$, where θ is a q-dimensional vector sufficiently small so that

$$\{1 + \theta^T h(z)\} \geq 0 \quad \text{for all } z. \tag{4.19}$$

Condition (4.19) is necessary to guarantee that $p(z, \theta)$ is a proper density. Because $h(\cdot)$ is a bounded function, the set of θ satisfying (4.19) contains an open set in $\mathbb{R}^q$. This must be the case in order to ensure that the partial derivatives of $p(z, \theta)$ with respect to θ exist. Moreover, every element $p(z, \theta)$ in the parametric submodel satisfies

$$\begin{aligned}\int p(z,\theta)d\nu(z) &= \int p_0(z)\left\{1+\theta^T h(z)\right\}d\nu(z)\\ &= \int \underbrace{p_0(z)d\nu(z)}_{\|\ 1} + \int \underbrace{\theta^T h(z)p_0(z)d\nu(z)}_{\|\ 0} = 1.\end{aligned}$$

This guarantees that $p(z, \theta)$, for θ in some neighborhood of the truth, is a proper density function. For this parametric submodel, the score vector is

$$\begin{aligned}S_\theta(z) &= \left.\frac{\partial \log[p_0(z)\{1+\theta^T h(z)\}]}{\partial\theta}\right|_{\theta=0}\\ &= h(z).\end{aligned}$$

If we choose $B^{q\times q}$ to be $I^{q\times q}$ (i.e., the $q \times q$ identity matrix), then $h(Z)$, which also equals $I^{q\times q}h(Z)$, is an element of this parametric submodel tangent space.

Thus we have shown that the tangent space contains all bounded mean-zero random vectors. The proof is completed by noting that any element of $\mathcal{H}$ can be approximated by a sequence of bounded h. □

Remark 7. On the dimension of θ
When we defined an arbitrary parametric submodel through the parameter θ, the dimension of θ was taken to be s-dimensional, where s was arbitrary. The corresponding score vector, which also is s-dimensional, had to be pre-multiplied by some arbitrary constant matrix $B^{q\times s}$ to obtain an element in the Hilbert space. However, when we were constructing a parametric submodel that led to the score vector $h(Z)$, we chose θ to be q-dimensional to conform to the dimension of $h(Z)$; i.e., for this particular parametric submodel, we took s to equal q. □

Partitioning the Hilbert Space

Suppose Z is an m-dimensional random vector, say $Z = Z^{(1)}, \ldots, Z^{(m)}$. Then the density of Z can be expressed as the product of conditional densities, namely

$$p_Z(z) = p_{Z^{(1)}}(z^{(1)}) \times p_{Z^{(2)}|Z^{(1)}}(z^{(2)}|z^{(1)}) \\ \times \ldots \times p_{Z^{(m)}|Z^{(1)},\ldots,Z^{(m-1)}}(z^{(m)}|z^{(1)},\ldots,z^{(m-1)}),$$

where

$$p_{Z^{(j)}|Z^{(1)},\ldots,Z^{(j-1)}}(z^{(j)}|z^{(1)},\ldots,z^{(j-1)}) \tag{4.20}$$

is the conditional density of $Z^{(j)}$ given $Z^{(1)},\ldots,Z^{(j-1)}$, defined with respect to the dominating measure ν_j. If we put no restrictions on the density of Z (i.e., the nonparametric model) or, equivalently, put no restrictions on the conditional densities above, then the j-th conditional density (4.20) is any positive function $\eta_j(z^{(1)},\ldots,z^{(j)})$ such that

$$\int \eta_j(z^{(1)},\ldots,z^{(j)}) d\nu_j(z^{(j)}) = 1$$

for all $z^{(1)},\ldots,z^{(j-1)}$, $j = 1,\ldots,m$. With this representation, we note that the nonparametric model can be represented by the m variationally independent infinite-dimensional nuisance parameters $\eta_1(\cdot),\ldots,\eta_m(\cdot)$. By variationally independent, we mean that the product of any combination of arbitrary $\eta_1(\cdot),\ldots,\eta_m(\cdot)$ can be used to construct a valid density for $Z^{(1)},\ldots,Z^{(m)}$ in the nonparametric model.

The tangent space $\mathscr{T}$ is the mean-square closure of all parametric submodel tangent spaces, where a parametric submodel is given by the class of densities

$$p(z^{(1)},\gamma_1)\prod_{j=2}^{m} p(z^{(j)}|z^{(1)},\ldots,z^{(j-1)},\gamma_j),$$

γ_j, $j = 1,\ldots,m$ are s_j-dimensional parameters that are variationally independent, and $p(z^{(j)}|z^{(1)},\ldots,z^{(j-1)},\gamma_{0j})$ denotes the true conditional density of $Z^{(j)}$ given $Z^{(1)},\ldots,Z^{(j-1)}$. The parametric submodel tangent space is defined as the space spanned by the score vectors $S_{\gamma_j}(Z^{(1)},\ldots,Z^{(m)})$, $j = 1,\ldots,m$, where $S_{\gamma_j}(z^{(1)},\ldots,z^{(m)}) = \partial \log p(z^{(1)},\ldots,z^{(m)},\gamma_1,\ldots,\gamma_m)/\partial\gamma_j$. Because the density of $Z^{(1)},\ldots,Z^{(m)}$ is a product of conditional densities, each parametrized through parameters γ_j that are variationally independent, this implies that the log-density is a sum of log-conditional densities with respect to variationally independent parameters γ_j. Hence, $S_{\gamma_j}(z^{(1)},\ldots,z^{(m)})$, which is defined as $\partial \log p(z^{(j)}|z^{(1)},\ldots,z^{(j-1)},\gamma_j)/\partial\gamma_j$, is a function of $z^{(1)},\ldots,z^{(j)}$ only. Moreover, the parametric submodel tangent space can be written as the space

$$\mathscr{T}_\gamma = B_1^{q\times s_1} S_{\gamma_1}(Z^{(1)}) + \ldots + B_m^{q\times s_m} S_{\gamma_m}(Z^{(1)},\ldots,Z^{(m)})$$

for all constant matrices $B_1^{q\times s_1},\ldots,B_m^{q\times s_m}$. Consequently, the parametric submodel tangent space is equal to

$$\mathscr{T}_\gamma = \mathscr{T}_{\gamma_1} \oplus \ldots \oplus \mathscr{T}_{\gamma_m},$$

where

$$\mathscr{T}_{\gamma_j} = \{B^{q\times s_j} S_{\gamma_j}(Z^{(1)}, \ldots, Z^{(j)}), \text{ for all constant matrices } B^{q\times s_j}\}.$$

It is now easy to verify that the tangent space $\mathscr{T}$, the mean-square closure of all parametric submodel tangent spaces, is equal to

$$\mathscr{T} = \mathscr{T}_1 \oplus \ldots \oplus \mathscr{T}_m,$$

where $\mathscr{T}_j$, $j = 1, \ldots, m$, is the mean-square closure of parametric submodel tangent spaces for $\eta_j(\cdot)$, where a parametric submodel for $\eta_j(\cdot)$ is given by the class of conditional densities

$$\mathscr{P}_{\gamma_j} = \{p(z^{(j)}|z^{(1)}, \ldots, z^{(j-1)}, \gamma_j),\ \gamma_j - \text{say } s_j - \text{dimensional}\}$$

and the parametric submodel tangent space $\mathscr{T}_{\gamma_j}$ is the linear space spanned by the score vector $S_{\gamma_j}(Z^{(1)}, \ldots, Z^{(j)})$.

We now are in a position to derive the following results regarding the partition of the Hilbert space $\mathcal{H}$ into a direct sum of orthogonal subspaces.

Theorem 4.5. The tangent space $\mathscr{T}$ for the nonparametric model, which we showed in Theorem 4.4 is the entire Hilbert space $\mathcal{H}$, is equal to

$$\mathscr{T} = \mathcal{H} = \mathscr{T}_1 \oplus \ldots \oplus \mathscr{T}_m,$$

where

$$\mathscr{T}_1 = \left\{\alpha_1^{q\times 1}(Z^{(1)}) \in \mathcal{H} :\ E\{\alpha_1^{q\times 1}(Z^{(1)})\} = 0^{q\times 1}\right\}$$

and

$$\mathscr{T}_j = \Big\{\alpha_j^{q\times 1}(Z^{(1)}, \ldots, Z^{(j)}) \in \mathcal{H} : \tag{4.21}$$
$$E\{\alpha_j^{q\times 1}(Z^{(1)}, \ldots, Z^{(j)})|Z^{(1)}, \ldots, Z^{(j-1)}\} = 0^{q\times 1}\Big\},\ j = 2, \ldots, m,$$

and $\mathscr{T}_j$, $j = 1, \ldots, m$ are mutually orthogonal spaces. Equivalently, the linear space $\mathscr{T}_j$ can be defined as the space

$$\left\{[h_{*j}^{q\times 1}(Z^{(1)}, \ldots, Z^{(j)}) - E\{h_{*j}^{q\times 1}(Z^{(1)}, \ldots, Z^{(j)})|Z^{(1)}, \ldots, Z^{(j-1)}\}]\right\} \tag{4.22}$$

for all square-integrable functions $h_{*j}^{q\times 1}(\cdot)$ of $Z^{(1)}, \ldots, Z^{(j)}$.

In addition, any element $h(Z^{(1)}, \ldots, Z^{(m)}) \in \mathcal{H}$ can be decomposed into orthogonal elements

$$h = h_1 + \ldots + h_m,$$

where

$$\begin{aligned} h_1(Z^{(1)}) &= E\{h(\cdot)|Z^{(1)}\}, \\ h_j(Z^{(1)}, \ldots, Z^{(j)}) &= E\{h(\cdot)|Z^{(1)}, \ldots, Z^{(j)}\} \\ &\quad - E\{h(\cdot)|Z^{(1)}, \ldots, Z^{(j-1)}\}, \end{aligned} \tag{4.23}$$

for $j = 2, \ldots, m$, and $h_j(\cdot)$ is the projection of h onto $\mathscr{T}_j$; i.e., $h_j(\cdot) = \Pi\{h(\cdot)|\mathscr{T}_j\}$.

Proof. That the partition of the tangent space $\mathscr{T}_j$ associated with the nuisance parameter $\eta_j(\cdot)$ is the set of elements given by (4.21) follows by arguments similar to those for the proof of Theorem 4.4. That is, because of properties of score functions for parametric models of conditional densities, the score vector $S_{\gamma_j}(\cdot)$ must be a function only of $Z^{(1)}, \ldots, Z^{(j)}$ and must have conditional expectation

$$E\{S_{\gamma_j}(Z^{(1)}, \ldots, Z^{(j)}) | Z^{(1)}, \ldots, Z^{(j-1)}\} = 0^{s_j \times 1}.$$

Consequently, any q-dimensional element spanned by $S_{\gamma_j}(\cdot)$ must belong to $\mathscr{T}_j$. Conversely, for any bounded element $\alpha_j(Z^{(1)}, \ldots, Z^{(j)})$ in $\mathscr{T}_j$, consider the parametric submodel

$$p_j(z^{(j)}|z^{(1)}, \ldots, z^{(j-1)}, \theta_j) = p_{0j}(z^{(j)}|z^{(1)}, \ldots, z^{(j-1)})\{1+\theta_j^T \alpha_j(z^{(1)}, \ldots, z^{(j)})\}, \tag{4.24}$$

where $p_{0j}(z^{(j)}|z^{(1)}, \ldots, z^{(j-1)})$ denotes the true conditional density of $Z^{(j)}$ given $Z^{(1)}, \ldots, Z^{(j-1)}$ and θ_j is a q-dimensional parameter chosen sufficiently small to guarantee that $p_j(z^{(j)}|z^{(1)}, \ldots, z^{(j-1)}, \theta_j)$ is positive. This class of functions is clearly a parametric submodel since

$$\int p_{0j}(z^{(j)}|z^{(1)}, \ldots, z^{(j-1)})\{1 + \theta_j^T \alpha_j(z^{(1)}, \ldots, z^{(j)})\} d\nu_j(z^{(j)}) = 1,$$

which follows because

$$\begin{aligned} &\textstyle\int p_{0j}(z^{(j)}|z^{(1)}, \ldots, z^{(j-1)}) d\nu_j(z^{(j)}) = 1, \\ &\textstyle\int p_{0j}(z^{(j)}|z^{(1)}, \ldots, z^{(j-1)})\{\theta_j^T \alpha_j(z^{(1)}, \ldots, z^{(j)})\} d\nu_j(z^{(j)}) = 0, \end{aligned} \tag{4.25}$$

where (4.25) is equal to $\theta_j^T E\{\alpha_j(Z^{(1)}, \ldots, Z^{(j)}) | Z^{(1)}, \ldots, Z^{(j-1)}\}$, which must equal zero by the definition of $\mathscr{T}_j$. The score vector for the parametric submodel (4.24) is

$$\begin{aligned} S_{\theta_j}(\cdot) &= \left. \frac{\partial \log p_{0j}(z^{(j)}|z^{(1)}, \ldots, z^{(j-1)})\{1 + \theta_j^T \alpha_j(z^{(1)}, \ldots, z^{(j)})\}}{\partial \theta_j} \right|_{\theta_j = 0} \\ &= \alpha_j(z^{(1)}, \ldots, z^{(j)}). \end{aligned}$$

Thus we have shown that the tangent space for every parametric submodel of $\eta_j(\cdot)$ is contained in $\mathscr{T}_j$ and every bounded element in $\mathscr{T}_j$ belongs to the tangent space for some parametric submodel of $\eta_j(\cdot)$. The argument is completed by noting that every element of $\mathscr{T}_j$ is the limit of bounded elements of $\mathscr{T}_j$.

That the projection of any element $h \in \mathcal{H}$ onto $\mathscr{T}_j$ is given by $h_j(\cdot)$, defined by (4.23), can be verified directly. Clearly $h_j(\cdot) \in \mathscr{T}_j$. Therefore, by the projection theorem for Hilbert spaces, we only need to verify that $h - h_j$ is orthogonal to every element of $\mathscr{T}_j$. Consider any arbitrary element $\ell_j \in \mathscr{T}_j$.

Then, because h_j and l_j are functions of $Z^{(1)}, \ldots, Z^{(j)}$, we can use the law of iterated conditional expectations to obtain

$$\begin{aligned} & E\{(h-h_j)^T \ell_j\} = E[E\{(h-h_j)^T \ell_j | Z^{(1)}, \ldots, Z^{(j)}\}] \\ &= E[\{E(h|Z^{(1)}, \ldots, Z^{(j)}) - h_j\}^T \ell_j] \\ &= E[\{E(h|Z^{(1)}, \ldots, Z^{(j-1)})\}^T \ell_j] \qquad (4.26) \\ &= E\left(E[\{E(h|Z^{(1)}, \ldots, Z^{(j-1)})\}^T \ell_j | Z^{(1)}, \ldots, Z^{(j-1)}]\right) \\ &= E[\{E(h|Z^{(1)}, \ldots, Z^{(j-1)})\}^T E(\ell_j | Z^{(1)}, \ldots, Z^{(j-1)})] = 0. \qquad (4.27) \end{aligned}$$

Note that (4.26) follows from the definition of h_j. The equality in (4.27) follows because $l_j \in \mathscr{T}_j$, which, in turn, implies that $E(l_j|Z^{(1)}, \ldots, Z^{(j-1)}) = 0$.

Finally, in order to prove that $\mathscr{T}_j, j = 1, \ldots, m$ are mutually orthogonal subspaces, we must show that h_j is orthogonal to $h_{j'}$, where $h_j \in \mathscr{T}_j$, $h_{j'} \in \mathscr{T}_{j'}$ and $j \neq j^{'}$, $j, j^{'} = 1, \ldots, m$. This follows using the law of iterated conditional expectations, which we leave for the reader to verify. □

4.5 Semiparametric Restricted Moment Model

We will focus a great deal of attention on studying the geometric properties of influence functions of estimators for parameters in the restricted moment model because of its widespread use in statistical modeling. The relationship of the response variable Y, which may be univariate or multivariate, and covariates X is modeled by considering the conditional expectation of Y given X as a function of X and a finite number of parameters β. This includes linear as well as nonlinear models. For example, if Y is a univariate response variable, then, in a linear model, we would assume that $E(Y|X) = X^T\beta$, where the dimensions of X and β were, say, equal to q. We also gave an example of a log-linear model earlier, in Section 4.1. More generally, the response variable Y may be multivariate, say d-dimensional, and the relationship of the conditional expectation of Y given X may be linear or nonlinear, say $E(Y|X) = \mu(X, \beta)$, where $\mu(x, \beta)$ is a d-dimensional function of the covariates X and the q-dimensional parameter β. Therefore, such models are particularly useful when modeling longitudinal and/or multivariate response data. The goal is to estimate the parameter β using a sample of iid data $(Y_i, X_i), i = 1, \ldots, n$. As we argued in Chapter 1, without any additional restrictions on the joint probability distribution of Y and X, this is a semiparametric model. The restricted moment model can also be expressed as

$$Y = \mu(X, \beta) + \varepsilon,$$

where

$$E(\varepsilon|X) = 0.$$

We will assume, for the time being, that the d-dimensional response variable Y is continuous; i.e., the dominating measure is the Lebesgue measure, which we will denote by ℓ_Y. It will be shown later how this can be generalized to more general dominating measures that will also allow Y to be discrete. The covariates X may be continuous, discrete, or mixed, and we will denote the dominating measure by ν_X.

The observed data are assumed to be realizations of the iid random vectors $(Z_1, \ldots, Z_n)$, where $Z_i = (Y_i, X_i)$. Our aim is to find semiparametric estimators for β and identify, if possible, the most efficient semiparametric estimator.

The density of a single observation, denoted by $p(z)$, belongs to the semiparametric model

$$\mathscr{P} = \left\{ p\{z, \beta, \eta(\cdot)\}, z = (y, x) \right\},$$

defined with respect to the dominating measure $\ell_Y \times \nu_X$. The truth (i.e., the density that generates the data) is denoted by $p_0(z) = p\{z, \beta_0, \eta_0(\cdot)\}$. Because there is a one-to-one transformation of (Y, X) and (ε, X), we can express the density

$$p_{Y,X}(y, x) = p_{\varepsilon,X}\{y - \mu(x, \beta), x\}, \tag{4.28}$$

where $p_{\varepsilon,X}(\varepsilon, x)$ is a density with respect to the dominating measure $\ell_\varepsilon \times \nu_X$.

The restricted moment model only makes the assumption that

$$E(\varepsilon|X) = 0.$$

As illustrated in Example 1 of Section 1.2, the density of (ε, X) can be expressed as

$$p_{\varepsilon,X}(\varepsilon, x) = \eta_1(\varepsilon, x)\ \eta_2\ (x), \tag{4.29}$$

where $\eta_1(\varepsilon, x) = p_{\varepsilon|X}(\varepsilon|x)$ is any nonnegative function such that

$$\int \eta_1(\varepsilon, x) d\varepsilon = 1 \quad \text{for all} \quad x, \tag{4.30}$$

$$\int \varepsilon \eta_1(\varepsilon, x) d\varepsilon = 0 \quad \text{for all} \quad x, \tag{4.31}$$

and $p_X(x) = \eta_2(x)$ is a nonnegative function of x such that

$$\int \eta_2(x) d\nu(x) = 1. \tag{4.32}$$

The set of functions $\eta_1(\varepsilon, x)$ and $\eta_2(x)$, satisfying the constraints (4.30), (4.31), and (4.32), are infinite-dimensional and can be used to characterize the semiparametric model as

$$p\{z, \beta, \eta_1(\cdot), \eta_2(\cdot)\} = \eta_1\{y - \mu(x, \beta), x\}\eta_2(x).$$

The true density generating the data is denoted by

$$\begin{aligned} p_0(z) &= \eta_{10}\{y - \mu\ (x, \beta_0), x\}\ \eta_{20}\ (x) \\ &= p\{z, \beta_0, \eta_{10}(\cdot), \eta_{20}(\cdot)\}. \end{aligned}$$

To develop the semiparametric theory and define the semiparametric nuisance tangent space, we first consider parametric submodels. Instead of arbitrary functions $\eta_1(\varepsilon, x)$ and $\eta_2(x)$ for $p_{\varepsilon|X}(\varepsilon|x)$ and $p_X(x)$ satisfying constraints (4.30), (4.31), and (4.32), we will consider parametric submodels

$$p_{\varepsilon|X}(\varepsilon|x, \gamma_1) \quad \text{and} \quad p_X(x, \gamma_2),$$

where γ_1 is an r_1-dimensional vector and γ_2 is an r_2-dimensional vector. Thus $\gamma = (\gamma_1^T, \gamma_2^T)^T$ is an r-dimensional vector, $r = r_1 + r_2$.

This parametric submodel is given as

$$\mathscr{P}_{\beta,\gamma} = \{p(z, \beta, \gamma_1, \gamma_2) = p_{\varepsilon|X}\{y - \mu(x, \beta)|x, \gamma_1\}p_X(x, \gamma_2) \tag{4.33}$$

for

$$\left(\beta^T, \gamma_1^T, \gamma_2^T\right)^T \in \Omega_{\beta,\gamma} \subset \mathbb{R}^{q+r}\}.$$

Also, to be a parametric submodel, $\mathscr{P}_{\beta,\gamma}$ must contain the truth; i.e.,

$$p_0(z) = p_{\varepsilon|X}\{y - \mu(x, \beta_0)|x, \gamma_{10}\}p_X(x, \gamma_{20}).$$

We begin by defining the parametric submodel nuisance tangent space and will show how this leads us to the semiparametric model nuisance tangent space. The parametric submodel nuisance score vector is given as

$$\begin{aligned} S_\gamma(z, \beta_0, \gamma_0) &= \left\{ \left(\frac{\partial \log p(z, \beta, \gamma)}{\partial \gamma_1} \right)^T, \left(\frac{\partial \log p(z, \beta, \gamma)}{\partial \gamma_2} \right)^T \right\}^T \Bigg| \begin{matrix} \beta = \beta_0, \\ \gamma = \gamma_0 \end{matrix} \\ &= \left\{ S_{\gamma_1}^T(z, \beta_0, \gamma_0), S_{\gamma_2}^T(z, \beta_0, \gamma_0) \right\}^T. \end{aligned}$$

By (4.33), we note that

$$\log p(z, \beta, \gamma_1, \gamma_2) = \log p_{\varepsilon|X}\{y - \mu(x, \beta)|x, \gamma_1\} + \log p_X(x, \gamma_2).$$

Therefore,

$$S_{\gamma_1}(z, \beta_0, \gamma_0) = \frac{\partial \log p_{\varepsilon|X}\{y - \mu(x, \beta_0)|x, \gamma_{10}\}}{\partial \gamma_1},$$

and

$$S_{\gamma_2}(z, \beta_0, \gamma_0) = \frac{\partial \log p_X(x, \gamma_{20})}{\partial \gamma_2}.$$

Since we are taking derivatives with respect to γ_1 and γ_2 and leaving β fixed for the time being at "β_0," we will use the simplifying notation that

$$\varepsilon = y - \mu(x, \beta_0).$$

Also, unless stated otherwise, if parameters are omitted in an expression, they will be understood to be evaluated at the truth. So, for example,

$$S_{\gamma_1}(z, \beta_0, \gamma_0) = \frac{\partial \log p_{\varepsilon|X}\{y - \mu(x, \beta_0)|x, \gamma_{10}\}}{\partial \gamma_1}$$

will be denoted as $S_{\gamma_1}(\varepsilon, x)$.

A typical element in the parametric submodel nuisance tangent space is given by

$$\underbrace{B^{q\times r}}_{\substack{\text{matrix}\\ \text{of constants}}} S_\gamma(\varepsilon, X) = B_1^{q\times r_1} S_{\gamma_1}(\varepsilon, X) + B_2^{q\times r_2} S_{\gamma_2}(X).$$

Therefore, the parametric submodel nuisance tangent space

$$\Lambda_\gamma = \left\{ B^{q\times r} S_\gamma \quad \text{for all} \quad B^{q\times r} \right\}$$

can be written as the direct sum $\Lambda_{\gamma_1} \oplus \Lambda_{\gamma_2}$, where

$$\Lambda_{\gamma_1} = \left\{ B^{q\times r_1} S_{\gamma_1}(\varepsilon, X) \quad \text{for all} \quad B^{q\times r_1} \right\} \tag{4.34}$$

and

$$\Lambda_{\gamma_2} = \left\{ B^{q\times r_2} S_{\gamma_2}(X) \quad \text{for all} \quad B^{q\times r_2} \right\}. \tag{4.35}$$

It is easy to show that the space Λ_{γ_1} is orthogonal to the space Λ_{γ_2}, as we demonstrate in the following lemma.

Lemma 4.1. The space Λ_{γ_1} defined by (4.34) is orthogonal to the space Λ_{γ_2} defined by (4.35).

Proof. Since $p_{\varepsilon|X}(\varepsilon|x, \gamma_1)$ is a conditional density, then by properties of score vectors

$$E\{S_{\gamma_1}(\varepsilon, X)|X\} = 0. \tag{4.36}$$

(Equation (4.36) is also derived explicitly later in (4.37).) Similarly,

$$E\{S_{\gamma_2}(X)\} = 0.$$

Consequently,

$$
\begin{aligned}
&E\{S_{\gamma_1}(\varepsilon, X)S_{\gamma_2}^T(X)\} \\
&= E\left[E\{S_{\gamma_1}(\varepsilon, X)S_{\gamma_2}^T(X)|X\}\right] \\
&= E\left[\underbrace{E\{S_{\gamma_1}(\varepsilon, X)|}_{\|}X\}S_{\gamma_2}^T(X)\right] = 0^{r_1 \times r_2}. \\
&\qquad\qquad 0
\end{aligned}
$$

Convince yourself that (4.36) suffices to show that every element of Λ_{γ_1} is orthogonal to every element of Λ_{γ_2}. □

By definition, the semiparametric nuisance tangent space

$$
\begin{aligned}
\Lambda &= \left\{ \begin{array}{l} \text{mean-square closure of all parametric} \\ \text{submodel nuisance tangent spaces} \end{array} \right\} \\
&= \{\text{mean-square closure of} \Lambda_{\gamma_1} \oplus \Lambda_{\gamma_2}\}.
\end{aligned}
$$

Because γ_1, γ_2 are variationally independent – that is, proper densities in the parametric submodel can be defined by considering any combination of γ_1 and γ_2 – this implies that $\Lambda = \Lambda_{1s} \oplus \Lambda_{2s}$, where

$$
\begin{aligned}
\Lambda_{1s} &= \{\text{mean-square closure of all} \Lambda_{\gamma_1}\}, \\
\Lambda_{2s} &= \{\text{mean-square closure of all} \Lambda_{\gamma_2}\}.
\end{aligned}
$$

We now show how to explicitly derive the spaces Λ_{1s}, Λ_{2s} and the space orthogonal to the nuisance tangent space $\Lambda^{\perp}$.

The Space Λ_{2s}

Since here we are considering marginal distributions of X with no restrictions, finding the space Λ_{2s} is similar to finding the nuisance tangent space for the nonparametric model given in Section 4.4. For completeness, we give the arguments so that the reader can become more facile with the techniques used in such exercises.

Theorem 4.6. The space Λ_{2s} consists of all q-dimensional mean-zero functions of X with finite variance.

Remark 8. In many cases, the structure of the parametric submodel nuisance tangent space will allow us to make an educated guess for the semiparametric nuisance tangent space; i.e., the mean-square closure of parametric submodel nuisance tangent spaces. After we make such a guess, we then need to verify that our guess is correct. □

We illustrate below.

Proof. For any parametric submodel,

(i) $S_{\gamma_2}(Z) = S_{\gamma_2}(X)$ is a function only of X

and

(ii) Any score vector has mean zero

$$E\{S_{\gamma_2}(X)\} = 0.$$

Therefore, any element of $\Lambda_{\gamma_2} = \{B^{q\times r_2} S_{\gamma_2}(X) \text{ for all } B\}$ is a q-dimensional function of X with mean zero. It may be reasonable to guess that Λ_{2s}, the mean-square closure of all Λ_{γ_2}, is the linear subspace of all q-dimensional mean-zero functions of X.

Recall: The Hilbert space is made up of all q-dimensional mean-zero functions of $Z = (Y, X)$; hence, the conjectured space $\Lambda_{2s} \subset \mathcal{H}$. □

We denote the conjectured space as

$$\Lambda_{2s}^{(\text{conj})} = \{\text{all } q\text{-dimensional mean-zero functions of } X\}.$$

In order to verify that our conjecture is true, we must demonstrate:

(a) Any element of Λ_{γ_2}, for any parametric submodel, belongs to $\Lambda_{2s}^{(\text{conj})}$, and conversely,
(b) any element of $\Lambda_{2s}^{(\text{conj})}$ is either an element of Λ_{γ_2} for some parametric submodel or a limit of such elements.

(a) is true because

$$E\{B^{q\times r_2} S_{\gamma_2}(X)\} = 0^{q\times 1}.$$

To verify (b), we start by choosing a bounded element $\alpha(X) \in \Lambda_{2s}^{(\text{conj})}$; i.e.,

$$E\{\alpha^{q\times 1}(X)\} = 0^{q\times 1},$$
$$\int \alpha(x) p_0(x) d\nu(x) = 0.$$

Consider the parametric submodel with density $p_X(x, \gamma_2) = p_0(x)\{1 + \gamma_2^T \alpha(x)\}$, γ_2 is a q-dimensional vector, and γ_2 is taken sufficiently small so that

$$\{1 + \gamma_2^T \alpha(x)\} \geq 0 \quad \text{for all } x.$$

This is necessary to guarantee that $p_X(x, \gamma_2)$ is a proper density in a neighborhood of γ_2 around zero and is true because $\alpha(x)$ is bounded.

The function $p_X(x, \gamma_2)$, in x, is a density function since $p_X(x, \gamma_2) \geq 0$ for all x and

$$\int p_X(x,\gamma_2)d\nu(x) = \int p_0(x)\{1+\gamma_2^T\alpha(x)\}d\nu(x)$$
$$= \underbrace{\int p_0(x)d\nu(x)}_{\|\ 1} + \underbrace{\int \gamma_2^T\alpha(x)p_0(x)\,d\nu(x)}_{\|\ 0} = 1.$$

For this parametric submodel, the score vector is

$$S_{\gamma_2}(x) = \left.\frac{\partial \log p_0(x)\{1+\gamma_2^T\alpha(x)\}}{\partial\gamma_2}\right|_{\gamma_2=0}$$
$$= \alpha(X).$$

Hence, by choosing the constant matrix $B^{q\times q}$ to be the $q\times q$ identity matrix, we deduce that $\alpha(X)$ is an element of this particular parametric submodel nuisance tangent space. Since arbitrary $\alpha(X)\in\Lambda_{2s}^{(\text{conj})}$ can always be taken as limits of bounded mean-zero functions of X, we have thus shown that all elements of $\Lambda_{2s}^{(\text{conj})}$ are either elements of a parametric submodel nuisance tangent space or a limit of such elements; hence, our conjecture has been verified and

$$\Lambda_{2s} = \{\text{all } q\text{-dimensional mean-zero functions of } X\}.\ \square$$

The Space Λ_{1s}

Theorem 4.7. The space Λ_{1s} is the space of all q-dimensional random functions $a(\varepsilon,x)$ that satisfy

$$\text{(i) } E\{a(\varepsilon,X)|X\} = 0^{q\times 1}$$

and

$$\text{(ii) } E\{a(\varepsilon,X)\varepsilon^T|X\} = 0^{q\times d}.$$

Proof. The space Λ_{1s} is the mean-square closure of all parametric submodel nuisance tangent spaces Λ_{γ_1}, where

$$\Lambda_{\gamma_1} = \{B^{q\times r_1}S_{\gamma_1}(\varepsilon,X) \text{ for all } B^{q\times r_1}\}$$

and

$$S_{\gamma_1}(\varepsilon,x) = \frac{\partial \log p_{\varepsilon|X}(\varepsilon|x,\gamma_{10})}{\partial\gamma_1}.$$

Recall: We use ε to denote $\{y-\mu(x,\beta_0)\}$. $\square$

Note the following relationships:

(i) $\int p_{\varepsilon|X}(\varepsilon|x,\gamma_1)d\varepsilon = 1$ for all x,γ_1, implies $\dfrac{\partial}{\partial\gamma_1}\int p_{\varepsilon|X}(\varepsilon|x,\gamma_1)d\varepsilon = 0$ for all x and γ_1. Interchanging integration and differentiation, dividing and multiplying by $p_{\varepsilon|X}(\varepsilon|x,\gamma_{10})$, and evaluating at $\gamma_1 = \gamma_{10}$, we obtain

$$\int \frac{\partial p_{\varepsilon|X}(\varepsilon|x,\gamma_{10})/\partial\gamma_1}{p_{\varepsilon|X}(\varepsilon|x,\gamma_{10})} p_{\varepsilon|X}(\varepsilon|x,\gamma_{10})d\varepsilon = 0$$

for all x; i.e.,

$$E\{S_{\gamma_1}(\varepsilon,X)|X\} = 0. \tag{4.37}$$

(ii) The model restriction $E(\varepsilon|X) = 0$ is equivalent to $\int p_{\varepsilon|X}(\varepsilon|x,\gamma_1)\varepsilon^T d\varepsilon = 0^{1\times d}$ for all x,γ_1. Using arguments similar to (i), where we differentiate with respect to γ_1, interchange integration and differentiation, divide and multiply by $p_{\varepsilon|X}(\varepsilon|x,\gamma_{10})$, and set γ_1 to γ_{10}, we obtain

$$\int S_{\gamma_1}(\varepsilon,x)\varepsilon^T p_{\varepsilon|X}(\varepsilon|x,\gamma_{10})d\varepsilon = 0 \text{ for all } x;$$
$$\text{i.e., } E\{S_{\gamma_1}(\varepsilon,X)\varepsilon^T|X\} = 0^{r_1\times d}.$$

Consequently, any element of $\Lambda_{\gamma_1} = B^{q\times r_1}S_{\gamma_1}(\varepsilon,X)$, say $a(\varepsilon,X)$, must satisfy

$$\text{(i) } E\{a(\varepsilon,X)|X\} = 0 \tag{4.38}$$

and

$$\text{(ii) } E\{a(\varepsilon,X)\varepsilon^T|X\} = 0^{q\times d}. \tag{4.39}$$

A reasonable conjecture is that Λ_{1s} is the space of all q-dimensional functions of (ε,X) that satisfy (4.38) and (4.39).

To verify this, consider the parametric submodel

$$p_{\varepsilon|X}(\varepsilon|x,\gamma_1) = p_{0\varepsilon|X}(\varepsilon|x)\{1+\gamma_1^T a(\varepsilon,x)\}$$

for some bounded function $a(\varepsilon,X)$ satisfying (4.38) and (4.39) and a q-dimensional parameter γ_1 chosen sufficiently small so that

$$\{1+\gamma_1^T a(\varepsilon,x)\} \geq 0 \quad \text{for all} \quad \varepsilon,x.$$

This parametric submodel contains the truth; i.e., $(\gamma_1 = 0)$. Also, the class of densities in this submodel consists of proper densities,

$$\int p_{\varepsilon|X}(\varepsilon|x,\gamma_1)d\varepsilon = 1 \text{ for all } x,\gamma_1,$$

and $E(\varepsilon|X) = 0$,

$$\int \varepsilon p_{\varepsilon|X}(\varepsilon|x,\gamma_1)d\varepsilon = 0^{d\times 1} \text{ for all } x,\gamma_1.$$

For this parametric submodel, the score vector is

$$S_{\gamma_1}(\varepsilon,x) = \frac{\partial \log p_{\varepsilon|X}(\varepsilon|x,\gamma_{10})}{\partial \gamma_1} = a(\varepsilon,x).$$

Premultiplying this score vector by the $q \times q$ identity matrix leads us to the conclusion that $a(\varepsilon,x)$ is an element of this parametric submodel nuisance tangent space.

Finally, any $a(\varepsilon,X) \in \mathcal{H}$ satisfying (4.38) and (4.39) can be obtained, in the limit, by a sequence of bounded $a(\varepsilon,X)$ satisfying (4.38) and (4.39). □

Recap:

$$\begin{aligned}\Lambda_{1s} = \{&a^{q\times 1}(\varepsilon,X) \text{ such that}\\ &E\{a(\varepsilon,X)|X\} = 0 \text{ and}\\ &E\{a(\varepsilon,X)\varepsilon^T|X\} = 0\},\\ \Lambda_{2s} = \{&\alpha^{q\times 1}(X) \text{ such that}\\ &E\{\alpha(X)\} = 0\}. \square\end{aligned}$$

It is now easy to demonstrate that Λ_{1s} is orthogonal to Λ_{2s}.

Lemma 4.2. $\Lambda_{1s} \perp \Lambda_{2s}$

Proof.

$$\begin{aligned}E\{\alpha^T(X)a(\varepsilon,X)\} &= E[E\{\alpha^T(X)a(\varepsilon,X)|X\}]\\ &= E[\alpha^T(X)E\underbrace{\{a(\varepsilon,X)|X\}}_{\|}] = 0. \square\\ &\qquad\qquad\quad 0\end{aligned}$$

The nuisance tangent space for the semiparametric model is $\Lambda = \Lambda_{1s} \oplus \Lambda_{2s}$. Note that Λ_{1s} is the intersection of two linear subspaces; namely,

$$\Lambda_{1sa} = \left\{ a_a^{q\times 1}(\varepsilon,X) : E\{a_a(\varepsilon,X)|X\} = 0^{q\times 1} \right\},$$

and

$$\Lambda_{1sb} = \left\{ a_b^{q\times 1}(\varepsilon,X) : E\{a_b^{q\times 1}\ (\varepsilon,X)\varepsilon^T|X\} = 0^{q\times d} \right\}.$$

Therefore, the nuisance tangent space $\Lambda = \Lambda_{2s} \oplus (\Lambda_{1sa} \cap \Lambda_{1sb})$.

There are certain relationships between the spaces Λ_{2s}, Λ_{1sa}, and Λ_{1sb} that will allow us to simplify the representation of the nuisance tangent space Λ and also allow us to derive the space orthogonal to the nuisance tangent space $\Lambda^\perp$ more easily. We give these relationships through a series of lemmas.

Lemma 4.3.

$$\Lambda_{1sa} = \Lambda_{2s}^{\perp}.$$

Lemma 4.4.

$$\Lambda_{2s} \subset \Lambda_{1sb}.$$

and

Lemma 4.5.

$$\Lambda = \Lambda_{2s} \oplus (\Lambda_{1sa} \cap \Lambda_{1sb}) = \Lambda_{1sb}.$$

Proof. Lemma 4.3
We first show that the space Λ_{1sa} is orthogonal to Λ_{2s}. Let $a(\varepsilon, X)$ be an arbitrary element of Λ_{1sa} and $\alpha(X)$ be an arbitrary element of Λ_{2s}. Then

$$\begin{aligned} E\{\alpha^T(X)a(\varepsilon, X)\} &= E[E\{\alpha^T(X)a(\varepsilon, X)|X\}] \\ &= E[\alpha^T(X)E\underbrace{\{a(\varepsilon, X)|X\}}_{\|}] = 0. \\ &\qquad\qquad\qquad\quad 0 \end{aligned}$$

To complete the proof, we must show that any element $h \in \mathcal{H}$ can be written as $h = h_1 \oplus h_2$, where $h_1 \in \Lambda_{2s}$ and $h_2 \in \Lambda_{1sa}$. We write $h = E(h|X) + \{h - E(h|X)\}$. That $E(h|X) \in \Lambda_{2s}$ and $\{h - E(h|X)\} \in \Lambda_{1sa}$ follow immediately. The result above also implies that $\Pi(h|\Lambda_{2s}) = E(h|X)$ and $\Pi(h|\Lambda_{1sa}) = \{h - E(h|X)\}$. □

Proof. Lemma 4.4
Consider any element $\alpha(X) \in \Lambda_{2s}$. Then

$$E\{\alpha(X)\varepsilon^T|X\} = \alpha(X)E(\varepsilon^T|X) = 0^{q\times d},$$

which follows from the model restriction $E(\varepsilon^T|X) = 0^{1\times d}$. Hence $\alpha(X) \in \Lambda_{1sb}$. □

Proof. Lemma 4.5
Consider any element $h_1 \in \Lambda_{2s}$. By Lemma 4.4, $h_1 \in \Lambda_{1sb}$. Let h_2 be any element in $(\Lambda_{1sa} \cap \Lambda_{1sb})$. Then, by definition, $h_2 \in \Lambda_{1sb}$. Hence $(h_1 + h_2) \in \Lambda_{1sb}$ since Λ_{1sb} is a linear space.

Conversely, let h be any element of Λ_{1sb}. Since by Lemmas 4.3 and 4.4 $E(h|X) \in \Lambda_{2s} \subset \Lambda_{1sb}$, this implies that $\{h - E(h|X)\} \in \Lambda_{1sb}$ since Λ_{1sb} is a linear space. Therefore h can be written as $E(h|X) + \{h - E(h|X)\}$, where $E(h|X) \in \Lambda_{2s}$ and $\{h - E(h|X)\} \in \Lambda_{1sb}$. But by Lemma 4.3, $\{h - E(h|X)\}$ is also an element of Λ_{1sa} and hence $\{h - E(h|X)\} \in (\Lambda_{1sa} \cap \Lambda_{1sb})$, completing the proof. □

Consequently, we have shown that the nuisance tangent space Λ for the semiparametric restricted moment model is given by

$$\Lambda = \Lambda_{1sb} = \{h(\varepsilon, X) \text{ such that } E\{h(\varepsilon, X)\varepsilon^T | X\} = 0^{q\times d}\}. \tag{4.40}$$

Influence Functions and the Efficient Influence Function for the Restricted Moment Model

The key to deriving the space of influence functions is first to identify elements of the Hilbert space that are orthogonal to Λ. Equivalently, the space $\Lambda^{\perp}$ is the linear space of residuals

$$h(\varepsilon, X) - \Pi(h(\varepsilon, X)|\Lambda)$$

for all

$$h(\varepsilon, X) \in \mathcal{H}.$$

Using (4.40), this equals

$$\left[h(\varepsilon, X) - \{\Pi(h|\Lambda_{1sb})\}\right]. \tag{4.41}$$

Theorem 4.8. The space orthogonal to the nuisance tangent space, $\Lambda^{\perp}$, or equivalently $\Lambda_{1sb}^{\perp}$, is

$$\{A^{q\times d}(X)\varepsilon \text{ for all } A^{q\times d}(X)\}, \tag{4.42}$$

where $A^{q\times d}(X)$ is the matrix of arbitrary $q \times d$-dimensional functions of X. Moreover, the projection of any arbitrary element $h(\varepsilon, X) \in \mathcal{H}$ onto Λ_{1sb} satisfies

$$h(\varepsilon, X) - \Pi(h|\Lambda_{1sb}) = g^{q\times d}(X)\varepsilon, \tag{4.43}$$

where

$$g(X) = E\{h(\varepsilon, X)\varepsilon^T | X\}\ \{E(\varepsilon\varepsilon^T | X)\}^{-1},$$

which implies that

$$\Pi[h|\Lambda_{1sb}] = h - E(h\varepsilon^T | X)\{E(\varepsilon\varepsilon^T | X)\}^{-1}\varepsilon. \tag{4.44}$$

Proof. Theorem 4.8
In order to prove that the space given by (4.42) is the orthogonal complement of Λ_{1sb}, we first prove that this space is orthogonal to Λ_{1sb}. That is, for any $A(X)\varepsilon$, we must show

$$E\{a_b^T(\varepsilon, X)A(X)\varepsilon\} = 0 \text{ for all } a_b \in \Lambda_{1sb}. \tag{4.45}$$

By a conditioning argument, this expectation equals

$$E[E\{a_b^T(\varepsilon, X)A(X)\varepsilon|X\}]. \tag{4.46}$$

But, by the definition of Λ_{1sb}, $E\{a_b(\varepsilon, X)\varepsilon^T|X\} = 0^{q\times d}$ or, equivalently, $E\{a_{bj}(\varepsilon, X)\varepsilon_{j'}|X\} = 0$, for all $j = 1, \ldots, q$ and $j^{'} = 1, \ldots, d$, where $\{a_{bj}(\varepsilon, X)$ is the j-th element of $a_b(\varepsilon, X)$ and $\varepsilon_{j'}$ is the $j^{'}$-th element of ε. Consequently, the inner expectation of (4.46), which can be written as

$$\sum_{j,j'} A_{jj'}(X)E\{a_{bj}(\varepsilon, X)\varepsilon_{j'}|X\},$$

where $A_{jj'}(X)$ is the $(j, j^{'})$-th element of $A(X)$, must also equal zero. This, in turn, proves (4.45).

Now that we have shown the orthogonality of the spaces Λ_{1sb} and the space (4.42), in order to prove that the space (4.42) is the orthogonal complement of Λ_{1sb}, it suffices to show that any $h \in \mathcal{H}$ can be written as $h_1 + h_2$, where $h_1 \in$ (4.42) and $h_2 \in \Lambda_{1sb}$. Or, equivalently, for any $h \in \mathcal{H}$, there exists $g^{q\times d}(X)$ such that

$$\{h(\varepsilon, X) - g(X)\varepsilon\} \in \Lambda_{1sb}. \tag{4.47}$$

That such a function $g(X)$ exists follows by solving the equation

$$E[\{h - g(X)\varepsilon\}\varepsilon^T|X] = 0^{q\times d},$$

or

$$E(h\varepsilon^T|X) - g(X)E(\varepsilon\varepsilon^T|X) = 0,$$

which yields

$$g(X) = E(h\varepsilon^T|X)\{E(\varepsilon\varepsilon^T|X)\}^{-1},$$

where, to avoid any technical difficulties, we will assume that the conditional variance matrix $E(\varepsilon\varepsilon^T|X)$ is positive definite and hence invertible. □

We have thus demonstrated that, for the semiparametric restricted moment model, any element of the Hilbert space perpendicular to the nuisance tangent space is given by

$$A^{q\times d}(X)\ \varepsilon \text{ or } A(X)\{Y - \mu(X, \beta_0)\}. \tag{4.48}$$

Influence functions of RAL estimators for β (i.e., $\varphi(\varepsilon, X)$) are normalized versions of elements perpendicular to the nuisance tangent space. That is, the space of influence functions, as well as being orthogonal to the nuisance tangent space, must also satisfy condition (i) of Theorem 4.2, namely that

$$E\{\varphi(\varepsilon, X)S_\beta^T(\varepsilon, X)\} = I^{q\times q},$$

where $S_\beta(\varepsilon, X)$ is the score vector with respect to the parameter β and $I^{q\times q}$ is the $q \times q$ identity matrix. If we start with any $A(X)$, and define $\varphi(\varepsilon, X) = C^{q\times q}A(X)\varepsilon$, where $C^{q\times q}$ is a $q\times q$ constant matrix (i.e., normalization factor), then condition (i) of Theorem 4.2 is satisfied by solving

$$E\{CA(X)\varepsilon S_\beta^T(\varepsilon, X)\} = I^{q\times q}$$

or

$$C = [E\{A(X)\varepsilon S_\beta^T(\varepsilon, X)\}]^{-1}. \tag{4.49}$$

Since a typical element orthogonal to the nuisance tangent space is given by $A(X)\{Y - \mu(X, \beta_0)\}$, and since a typical influence function is given by $CA(X)\{Y - \mu(X, \beta_0)\}$, where C is defined by (4.49), this motivates us to consider an m-estimator for β of the form

$$\sum_{i=1}^{n} CA(X_i)\{Y_i - \mu(X_i, \beta)\} = 0.$$

Because C is a multiplicative constant matrix, then, as long as C is invertible, this is equivalent to solving the equation

$$\sum_{i=1}^{n} A(X_i)\{Y_i - \mu(X_i, \beta)\} = 0.$$

This logic suggests that estimators can often be motivated by identifying elements orthogonal to the nuisance tangent space, a theme that will be used frequently throughout the remainder of the book.

We showed in Section 4.1 that solutions to such linear estimating equations result in semiparametric GEE estimators with influence function (4.4) that is proportional to $A(X)\{Y - \mu(X, \beta)\}$, where the proportionality constant is given by $\{E(AD)\}^{-1}$. In the next section, we will show that this proportionality constant satisfies equation (4.49). This will be a consequence of the results obtained in deriving the efficient influence function, as we now demonstrate.

The Efficient Influence Function

To derive an efficient semiparametric estimator, we must find the efficient influence function. For this, we need to derive the efficient score (i.e., the residual after projecting the score vector with respect to β onto the nuisance tangent space Λ) $S_{\text{eff}}(\varepsilon, X) = S_\beta(\varepsilon, X) - \Pi\{S_\beta(\varepsilon, X)|\Lambda\}$, which by (4.44) equals

$$S_{\text{eff}}(\varepsilon, X) = E\{S_\beta(\varepsilon, X)\varepsilon^T|X\}V^{-1}(X)\varepsilon, \tag{4.50}$$

where $V(X)$ denotes $E(\varepsilon\varepsilon^T|X)$.

Recall that the restricted moment model was characterized by $\{\beta, \eta_1(\varepsilon, x), \eta_2(x)\}$, where β is the parameter of interest and $\eta_1(\cdot), \eta_2(\cdot)$ are infinite-dimensional nuisance parameters. When studying the nuisance tangent space, we fixed β at the truth and varied $\eta_1(\cdot)$ and $\eta_2(\cdot)$. To compute S_β, we fix the nuisance parameters at the truth and vary β.

A typical density in our model was given as

$$\eta_1\{y - \mu(x, \beta), x\}\eta_2(x),$$

where

$$\eta_1(\varepsilon, x) = p_{\varepsilon|X}(\varepsilon|x)$$

and

$$\eta_2(x) = p_X(x).$$

If we fix the nuisance parameter at the truth (i.e., $\eta_{10}(\varepsilon, x)$ and $\eta_{20}(x)$), then

$$\begin{aligned} S_\beta(y, x, \beta_0, \eta_0(\cdot)) &= \left. \frac{\partial \log \eta_{10}\{y - \mu(x, \beta), x\}}{\partial \beta} \right|_{\beta=\beta_0}, \\ &= \left. \frac{\partial \eta_{10}(\varepsilon, x)/\partial \beta}{\eta_{10}(\varepsilon, x)} \right|_{\beta=\beta_0}. \end{aligned} \tag{4.51}$$

Due to the model restriction,

$$\int \{y - \mu(x, \beta)\} \eta_{10}\{y - \mu(x, \beta), x\} dy = 0 \quad \text{for all} \quad x, \beta,$$

we obtain

$$\left. \frac{\partial}{\partial \beta^T} \int \{y - \mu(x, \beta)\} \eta_{10}\{y - \mu(x, \beta), x\} dy \right|_{\beta=\beta_0} = 0.$$

Taking the derivative inside the integral, we obtain

$$\int \left\{ -\frac{\partial \mu(x, \beta_0)}{\partial \beta^T} \right\}^{d \times q} \eta_{10}(\varepsilon, x) d\varepsilon + \int \varepsilon S_\beta^T \{\varepsilon, x, \beta_0, \eta_0(\cdot)\} \eta_{10}(\varepsilon, x) d\varepsilon = 0$$

for all x, or

$$\frac{-\partial \mu(X, \beta_0)}{\partial \beta^T} + E\left\{ \varepsilon S_\beta^T(\varepsilon, X) | X \right\} = 0.$$

After solving the preceding equation and taking the transpose, we obtain

$$D^T(X) = E\{S_\beta(\varepsilon, X) \varepsilon^T | X\}, \tag{4.52}$$

where

$$D(X) = \frac{\partial \mu(X, \beta_0)}{\partial \beta^T}.$$

By (4.50) and (4.52), we obtain that the efficient score is

$$S_{\text{eff}}(\varepsilon, X) = D^T(X) V^{-1}(X) \varepsilon, \tag{4.53}$$

and the optimal estimator is obtained by solving the estimating equation

$$\sum_{i=1}^{n} D^T(X_i)V^{-1}(X_i)\{Y_i - \mu(X_i, \beta)\} = 0 \tag{4.54}$$

(optimal GEE).

We also note that the normalization constant matrix C given in (4.49) involves the expectation $E\{A(X)\varepsilon S_\beta^T(\varepsilon, X)\}$, which by a conditioning argument can be derived as $E[A(X)E\{\varepsilon S_\beta^T(\varepsilon, X)|X\}]$, which equals $E\{A(X)D(X)\}$ by (4.52). Hence, $C = [E\{A(X)D(X)\}]^{-1}$. This implies that a typical influence function is given by

$$[E\{A(X)D(X)\}]^{-1}A(X)\{Y - \mu(X, \beta)\},$$

which is the influence function for the GEE estimator given in (4.4). Similarly, the efficient influence function can be obtained by using the appropriate normalization constant with the efficient score (4.53) to yield

$$[E\{D^T(X)V^{-1}(X)D(X)\}]^{-1}D^T(X)V^{-1}(X)\{Y - \mu(X, \beta)\}. \tag{4.55}$$

The semiparametric efficiency bound is given as

$$\mathcal{V} = [E\{S_{\text{eff}}(\varepsilon, X)S_{\text{eff}}^T(\varepsilon, X)\}]^{-1},$$

which by (4.53) is equal to

$$\begin{aligned} &[E\{D^T(X)V^{-1}(X)\varepsilon\varepsilon^T V^{-1}(X)D(X)\}]^{-1} \\ &= [E\{D^T(X)V^{-1}(X)E(\varepsilon\varepsilon^T|X)V^{-1}(X)D(X)\}]^{-1} \\ &= [E\{D^T(X)V^{-1}(X)D(X)\}]^{-1}. \end{aligned} \tag{4.56}$$

This, of course, is also the variance of the efficient influence function (4.54).

A Different Representation for the Restricted Moment Model

Up to now, we have defined probability models for the restricted moment model when the response variable Y was continuous. This allowed us to consider conditional densities for ε given X, where $\varepsilon = Y - \mu(X, \beta)$. We were able to do this because we assumed that Y was a continuous variable with respect to Lebesgue measure and hence the transformed variable ε was also a continuous variable with respect to Lebesgue measure. This turned out to be useful, as we could then describe the semiparametric model through the finite-dimensional parameter of interest β together with the infinite-dimensional nuisance parameters $\eta_1(\varepsilon, x)$ and $\eta_2(x)$, where β was variationally independent of $\eta_1(\varepsilon, x)$ and $\eta_2(x)$. That is, any combination of β, $\eta_1(\varepsilon, x)$, and $\eta_2(x)$ would lead to a valid density describing our semiparametric model.

There are many problems, however, where we want to use the restricted moment model with a dependent variable Y that is not a continuous random

variable. Strictly speaking, the response variable CD4 count used in the log-linear model example of Section 4.1 is not a continuous variable. A more obvious example is when we have a binary response variable Y taking on values 1 (response) or 0 (nonresponse). A popular model for modeling the probability of response as a function of covariates X is the logistic regression model. In such a model, we assume

$$P(Y = 1|X) = \frac{\exp(\beta^T X^*)}{1 + \exp(\beta^T X^*)},$$

where $X^* = (1, X^T)^T$, allowing the introduction of an intercept term. Since Y is a binary indicator, this implies that $E(Y|X) = P(Y = 1|X)$, and hence the logistic regression model is just another example of a restricted moment model with $\mu(X, \beta) = \frac{\exp(\beta^T X^*)}{1+\exp(\beta^T X^*)}$.

The difficulty that occurs when the response variable Y is not a continuous random variable is that the transformed variable $Y - \mu(X, \beta)$ may no longer have a dominating measure that allows us to define densities. In order to address this problem, we will work directly with densities defined on (Y, X), namely $p(y, x)$ with respect to some dominating measure $\nu_Y \times \nu_X$. As you will see, many of the arguments developed previously will carry over to this more general setting.

We start by first deriving the nuisance tangent space. As before, we need to find parametric submodels. Let $p(y, x)$ be written as $p(y|x)\ p(x)$, where $p(y|x)$ is the conditional density of Y given X and $p(x)$ is the marginal density of X, and denote the truth as $p_0(y, x) = p_0(y|x)p_0(x)$. The parametric submodel can be written generically as

$$p(y|x, \beta, \gamma_1)p(x, \gamma_2),$$

where for some $\beta_0, \gamma_{10}, \gamma_{20}$

$$p_0(y|x) = p(y|x, \beta_0, \gamma_{10})$$

and

$$p_0(x) = p(x, \gamma_{20}).$$

The parametric submodel nuisance tangent space is the space spanned by the score vector with respect to the nuisance parameters γ_1 and γ_2. As in the previous section, the parametric submodel nuisance tangent space can be written as the direct sum of two orthogonal spaces

$$\Lambda_{\gamma_1} \oplus \Lambda_{\gamma_2}, \quad \Lambda_{\gamma_1} \perp \Lambda_{\gamma_2},$$

where

$$\begin{aligned}
\Lambda_{\gamma_1} &= \left\{ B^{q\times r_1} S_{\gamma_1}(Y, X) \quad \text{for all} \quad B^{q\times r_1} \right\}, \\
\Lambda_{\gamma_2} &= \left\{ B^{q\times r_2} S_{\gamma_2}(X) \quad \text{for all} \quad B^{q\times r_2} \right\}, \\
S_{\gamma_1}(y, x) &= \frac{\partial \log p(y|x, \beta_0, \gamma_{10})}{\partial \gamma_1},
\end{aligned}$$

and

$$S_{\gamma_2}(x) = \frac{\partial \log p(x, \gamma_{20})}{\partial \gamma_2}.$$

Hence, the semiparametric nuisance tangent space Λ equals $\Lambda_{1s} \oplus \Lambda_{2s}$, $\Lambda_{1s} \perp \Lambda_{2s}$, where

$$\Lambda_{1s} = \{\text{mean-square closure of all } \Lambda_{\gamma_1}\}$$

and

$$\Lambda_{2s} = \{\text{mean-square closure of all } \Lambda_{\gamma_2}\}.$$

We showed previously that

$$\Lambda_{2s} = \left\{ \begin{array}{l} \text{all } q\text{-dimensional mean-zero} \\ \text{measurable functions of } X \\ \text{with finite second moments; i.e.,} \\ \alpha^{q\times 1}(X) : E\{\alpha(X)\} = 0 \end{array} \right\}.$$

We now consider the space Λ_{1s}. Again, we proceed by making educated guesses for the nuisance tangent space by considering the structure of the parametric submodel nuisance tangent space and then verify that our guess is correct. For the restricted moment model, if we fix β at the truth, (β_0), then the conditional densities $p(y|x, \beta_0, \gamma_1)$ must satisfy

$$\int p(y|x, \beta_0, \gamma_1) d\nu(y) = 1 \quad \text{for all} \quad x, \gamma_1 \tag{4.57}$$

and

$$\int y p(y|x, \beta_0, \gamma_1) d\nu(y) = \mu(x, \beta_0) \quad \text{for all} \quad x, \gamma_1. \tag{4.58}$$

Using standard arguments, where we take derivatives of (4.57) and (4.58) with respect to γ_1, interchange integration and differentiation, divide and multiply by $p(y|x, \beta_0, \gamma_1)$, and set γ_1 at the truth, we obtain

$$\int S_{\gamma_1}(y, x) p_0(y|x) d\nu(y) = 0^{r_1 \times 1} \quad \text{for all} \quad x$$

and

$$\int y S^T_{\gamma_1}(y, x) p_0(y|x) d\nu(y) = 0^{d \times r_1} \quad \text{for all} \quad x.$$

That is, $E\{S_{\gamma_1}(Y, X)|X\} = 0^{r_1 \times 1}$ and $E\{Y S^T_{\gamma_1}(Y, X)|X\} = 0^{d \times r_1}$. This implies that any element of Λ_{γ_1}, namely $B^{q \times r_1} S_{\gamma_1}(Y, X)$, would satisfy

$E\{B^{q\times r_1}S_{\gamma_1}(Y,X)|X\} = 0^{q\times 1}$ and $E\{B^{q\times r_1}S_{\gamma_1}(Y,X)Y^T|X\} = 0^{q\times d}$. This leads us to the conjecture that

$$\Lambda_{1s}^{(\text{conj})} = \Big\{ a^{q\times 1}(Y,X) : E\{a(Y,X)|X\} = 0^{q\times 1}$$

and

$$E\{a(Y,X)Y^T|X\} = 0^{q\times d}\Big\}.$$

To verify this conjecture, we consider an arbitrary bounded element $a^{q\times 1}(Y,X) \in \Lambda_{1s}^{(\text{conj})}$. We construct the parametric submodel $p_0(y|x)\{1 + \gamma_1^T a(y,x)\}$ with γ_1 chosen sufficiently small to ensure $\{1 + \gamma_1^T a(y,x)\} \geq 0$ for all y, x. This parametric submodel contains the truth when $\gamma_1 = 0$ and satisfies the constraints of the restricted moment model, namely

$$\int p(y|x,\gamma_1)d\nu(y) = 1 \quad \text{for all} \quad x,\gamma_1,$$
$$\int yp(y|x,\gamma_1)d\nu(y) = \mu(x,\beta_0) \quad \text{for all} \quad x,\gamma_1.$$

The score vector $S_{\gamma_1}(Y,X)$ for this parametric submodel is $a(Y,X)$. Also any element of $\Lambda_{1s}^{(conj)}$ can be derived as limits of bounded elements in $\Lambda_{1s}^{(conj)}$. Therefore, we have established that any element of a parametric submodel nuisance tangent space Λ_{γ_1} is an element of $\Lambda_{1s}^{(conj)}$, and any element of $\Lambda_{1s}^{(conj)}$ is either an element of a parametric submodel nuisance tangent space or a limit of such elements. Thus, we conclude that $\Lambda_{1s} = \Lambda_{1s}^{(conj)}$.

Note that Λ_{1s} can be expressed as $\Lambda_{1sa} \cap \Lambda_{1sb}$, where

$$\Lambda_{1sa} = \{a^{q\times 1}(Y,X) : E\{a(Y,X)|X\} = 0^{q\times 1}\}$$

and

$$\Lambda_{1sb} = \{a^{q\times 1}(Y,X) : E\{a(Y,X)\{Y - \mu(X,\beta_0)\}|X\} = 0^{q\times d}\}.$$

Therefore the nuisance tangent space $\Lambda = \Lambda_{2s} \oplus (\Lambda_{1sa} \cap \Lambda_{1sb})$. This representation is useful because $\Lambda_{2s} \oplus \Lambda_{1sa} = \mathcal{H}$ (the whole Hilbert space) and $\Lambda_{2s} \subset \Lambda_{1sb}$. Consequently, we use Lemmas 4.3–4.5 to show that the semiparametric nuisance tangent space $\Lambda = \Lambda_{1sb}$.

Using the exact same proof as for Theorem 4.8, we can show that

$$h(Y,X) - \Pi[h|\Lambda_{1sb}] \quad \text{or} \quad \Pi[h|\Lambda_{1sb}^{\perp}]$$
$$= E\{h(Y,X)(Y - \mu(X,\beta_0))^T|X\}V^{-1}(X)\{Y - \mu(X,\beta_0)\}.$$

To complete the development of the semiparametric theory, we still need to derive the efficient score or

$$S_{\text{eff}}(Y,X) = \Pi[S_\beta(Y,X)|\Lambda^\perp]$$
$$= E[S_\beta(Y,X)\{(Y-\mu(X,\beta_0)\}^T|X]V^{-1}(X)\{Y-\mu(X,\beta_0)\}. \quad (4.59)$$

Since

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

then for any parametric submodel where $p(y|x,\beta,\gamma_1)$ satisfies

$$\int yp(y|x,\beta,\gamma_1)d\nu(y) = \mu(x,\beta) \text{ for all } x,\gamma_1 \quad (4.60)$$

and

$$p(y|x,\beta_0,\gamma_{10}) = p_0(y|x),$$

assuming at least one such parametric submodel exists, we can differentiate both sides of (4.60) with respect to β^T, interchange integration and differentiation, divide and multiply by $p_0(y|x)$, and set β and γ_1 equal to β_0 and γ_{10}, respectively, to obtain

$$E\{YS_\beta^T(Y,X)|X\}$$
$$= E[\{Y-\mu(X,\beta_0)\}S_\beta^T(Y,X)|X]$$
$$= D(X), \quad (4.61)$$

where

$$D(X) = \frac{\partial\mu(X,\beta_0)}{\partial\beta^T}.$$

Equation (4.61) follows because

$$E\{\mu(X,\beta_0)S_\beta^T(Y,X)|X\} = \mu(X,\beta_0)E\{S_\beta^T(Y,X)|X\} = 0.$$

Taking transposes yields

$$E[S_\beta(Y,X)\{Y-\mu(X,\beta_0)\}^T|X] = D^T(X).$$

Consequently, the efficient score (4.59) is given by

$$S_{\text{eff}}(Y,X) = D^T(X)V^{-1}(X)\{Y-\mu(X,\beta_0)\}.$$

It still remains to show that a parametric submodel exists that satisfies (4.60). This is addressed by the following argument.

Existence of a Parametric Submodel for the Arbitrary Restricted Moment Model

A class of joint densities for (Y,X) can be defined with dominating measure $\nu_Y \times \nu_X$ that satisfy $E(Y|X) = \mu(X,\beta)$ by considering the conditional density

of Y given X multiplied by the marginal density of X, where a class of conditional densities for Y given X can be constructed using exponential tilting. To illustrate, let us consider, for simplicity, the case where Y is a univariate bounded random variable. We will assume that, at the truth, the conditional density of Y given X is given by $p_0(y|x)$, where $\int yp_0(y|x)d\nu(y) = \mu(x, \beta_0)$ for all x. The question is whether we can define a class of conditional densities $p(y|x, \beta)$ with respect to ν_Y such that $\int yp(y|x, \beta)d\nu(y) = \mu(x, \beta)$ for all x and for β in a neighborhood of β_0. We consider the conditional densities

$$p(y|x, \beta) = \frac{p_0(y|x) \exp\{c(x, \beta)y\}}{\int \exp\{c(x, \beta)y\}p_0(y|x)d\nu(y)},$$

where $c(x, \beta)$, if possible, is chosen so that $\int yp(y|x, \beta)d\nu(y) = \mu(x, \beta)$ for β in a neighborhood of β_0. We first note that we can take $c(x, \beta_0)$ to be equal to zero because, by definition, $\int yp_0(y|x)d\nu(y) = \mu(x, \beta_0)$. To illustrate that $c(x, \beta)$ exists and can be uniquely defined in a neighborhood of β_0, we fix the value of x and consider the function

$$\int y \frac{p_0(y|x) \exp(cy)}{\int \exp(cy)p_0(y|x)d\nu(y)} d\nu(y)$$

as a function in c, which can also be written as

$$\frac{E_0\{Y \exp(cY)|X = x\}}{E_0\{\exp(cY)|X = x\}}, \tag{4.62}$$

where the conditional expectation $E_0(\cdot|X = x)$ is taken with respect to the conditional density $p_0(y|x)$. Taking the derivative of (4.62) with respect to c, we obtain

$$\frac{E_0\{Y^2 \exp(cY)|X = x\}}{E_0\{\exp(cY)|X = x\}} - \left[\frac{E_0\{Y \exp(cY)|X = x\}}{E_0\{\exp(cY)|X = x\}}\right]^2.$$

This derivative, being the conditional variance of Y given $X = x$ with respect to the conditional density $\frac{p_0(y|x) \exp(cy)}{\int \exp(cy)p_0(y|x)d\nu(y)}$, must therefore be positive, implying that the function (4.62) is strictly monotonically increasing in c. Hence, in a neighborhood of β about the value β_0, a unique inverse for c exists in a neighborhood of zero that satisfies the equation

$$\frac{E_0\{Y \exp(cY)|X = x\}}{E_0\{\exp(cY)|X = x\}} = \mu(x, \beta).$$

We define this solution as $c(x, \beta)$.

The arguments above can be generalized to multivariate Y by using the inverse function theorem.

4.6 Adaptive Semiparametric Estimators for the Restricted Moment Model

Using the theory we have developed for the semiparametric restricted moment model, we now know that the class of all influence functions for semiparametric RAL estimators for β must be of the form

$$[E\{A(X)D(X)\}]^{-1}A(X)\{Y-\mu(X,\beta_0)\}$$

for any arbitrary $q \times d$ matrix, $A(X)$, of functions of X. We also showed that the solution to the estimating equation

$$\sum_{i=1}^{n} A(X_i)\{Y_i-\mu(X_i,\beta)\}=0 \tag{4.63}$$

results in an estimator for β that is RAL with influence function given by (4.4). The estimating equation (4.63) is an example of what Liang and Zeger (1986) referred to as a linear estimating equation or a GEE estimator. Using semiparametric theory, we showed that this class of estimators encompasses, at least asymptotically, all possible semiparametric RAL estimators for β in a restricted moment model. That is, any semiparametric RAL estimator must have an influence function that is contained within the class of influence functions for GEE estimators. Consequently, it is reasonable to restrict attention to only such estimators and, moreover, the efficient RAL estimator must be asymptotically equivalent to the efficient GEE estimator given by the solution to (4.54).

It is important to note that the optimal estimating equation depends on using the correct $V(X)$, where

$$V(X)=E(\varepsilon\varepsilon^T|X)$$

is the conditional variance of ε given X, or equivalently, the conditional variance of Y given X. Since, in a semiparametric model, the distribution of ε given X is left unspecified, the function $V(x)$ is unknown to us. We can try to estimate this conditional variance of Y as a function of $X=x$ using the data, but generally, without additional assumptions, this requires smoothing estimators that are not very stable, especially if X is multidimensional. Consequently, substituting such nonparametric type smoothing estimators $\hat{V}(X)$ into equation (4.54) leads to estimators for β that perform poorly with moderate sample sizes.

Another strategy is to posit some relationship for $V(x)$, either completely specified or as a function of a finite (small) number of additional parameters ξ as well as the parameters β in the restricted moment model. This is referred to as a working variance assumption because the model $V(x,\xi,\beta)$ for the conditional variance of Y given $X=x$ may not contain the truth. For example,

suppose, for simplicity, we take the response variable Y to be one-dimensional (i.e., $d = 1$) and assume the variance function

$$V(x, \xi) = \exp(\xi_0 + \xi_1^T x),$$

where ξ_1 is a vector of dimension equal to the number of covariates that make up the vector X. For this illustration, we chose a log-linear relationship to ensure a positive variance function. We might choose such a model for the variance not necessarily because we believe this is the true relationship but rather because if we believe that the variance is related to the covariates, then this model may capture some of this relationship, at least to a first order.

Another model, if one believes that the conditional variance of Y given X may be related to the conditional mean of Y given X, is to assume that

$$V(x, \xi, \beta) = \xi_0^2 \{\mu(x, \beta)\}^{\xi_1}, \tag{4.64}$$

where ξ_0 and ξ_1 are scalar constants. Again, we may not believe that this captures the true functional relationship of the conditional variance to the conditional mean but may serve as a useful approximation. Nonetheless, if, for the time being, we accepted $V(x, \xi, \beta)$ as a working model, then the parameters ξ in $V(x, \xi, \beta)$ can be estimated separately using the squared residuals $\{Y_i - \mu(X_i, \hat{\beta}_n^{initial})\}^2$, $i = 1, \ldots, n$, where $\hat{\beta}_n^{initial}$ is some initial consistent estimator for β. For instance, we can find an initial estimator for β by solving equation (4.1) using $A(X, \beta) = D(X, \beta)$ (which is equivalent to a working variance $V(X)$ proportional to the identity matrix). Using this initial estimator, we can then find an estimator for ξ by solving the equation

$$\sum_{i=1}^{n} \mathcal{Q}(X_i, \xi, \hat{\beta}_n^{initial}) \left[\{Y_i - \mu(X_i, \hat{\beta}_n^{initial})\}^2 - V(X_i, \xi, \hat{\beta}_n^{initial}) \right] = 0,$$

where $\mathcal{Q}(X, \xi, \beta)$ is an arbitrary vector of functions of X, ξ, and β of dimension equal to the dimension of ξ. One possibility is to choose $\mathcal{Q}(X, \xi, \beta) = \partial V(X, \xi, \beta)/\partial \xi$. Denote the resulting estimator by $\hat{\xi}_n$. Under weak regularity conditions, $\hat{\xi}_n$ will converge in probability to some constant ξ^* whether the variance function was correctly specified or not. Without going into the technicalities, it can be shown that substituting $V(X_i, \hat{\xi}_n, \hat{\beta}_n^{initial})$ into equation (4.54) will result in an RAL estimator for β, namely $\hat{\beta}_n$ (not to be confused with $\hat{\beta}_n^{initial}$), with influence function $[E\{A(X)D(X)\}]^{-1} A(X)\{Y - \mu(X, \beta)\}$, where $A(X) = D^T(X)V^{-1}(X, \xi^*, \beta_0)$. Consequently, solutions to such estimating equations, where we use a working variance assumption, will lead to what is called a *locally efficient* estimator for β. That is, if the posited relationship for $V(x, \xi, \beta)$ is indeed true (i.e., if the true conditional variance of Y given $X = x$ is contained in the model $V(x, \xi, \beta)$, with $V(x, \xi_0, \beta_0)$ denoting the truth), then $V(x, \hat{\xi}^n, \hat{\beta}_n^{initial})$ will converge to $V(x, \xi_0, \beta_0) = \text{var}(Y|X = x)$ and the resulting estimator is semiparametric efficient; otherwise, it is not. However, even if the posited model is not correct (i.e.,

$V(x, \xi^*, \beta_0) \neq \text{var}(Y|X = x)$), $V(X, \xi^*, \beta_0)$ is still a function of X and the resulting estimator for β is consistent and asymptotically normal. Such adaptive estimators for β have been shown empirically to have high relative efficiency compared with the optimal semiparametric efficiency bound if the working variance model provides a good approximation to the truth.

When using a working variance, one must be careful in estimating the asymptotic variance of the estimator $\hat{\beta}_n$. Since the asymptotic variance of the efficient influence function is given by (4.56), there is a natural temptation to estimate the asymptotic variance by using an estimator for $[E\{D^T(X)V^{-1}(X, \xi^*, \beta_0)D(X)\}]^{-1}$, namely

$$\left\{ n^{-1} \sum_{i=1}^{n} D^T(X_i, \hat{\beta}_n) V^{-1}(X_i, \hat{\xi}_n, \hat{\beta}_n^{initial}) D(X_i, \hat{\beta}_n) \right\}^{-1}.$$

This estimator would only be a consistent estimator for the asymptotic variance of $\hat{\beta}_n$ if the working variance contained the truth. Otherwise, it would be asymptotically biased. A consistent estimator for the asymptotic variance can be obtained by using the sandwich variance given by (4.14) with $A(X) = D^T(X, \hat{\beta}_n)V^{-1}(X, \hat{\xi}_n, \hat{\beta}_n^{initial})$.

Example 1. Logistic regression model
We argued earlier that the logistic regression model is an example of a restricted moment model where the response variable Y is a binary variable taking on the values 1 or 0 and $\mu(X, \beta) = \frac{\exp(\beta^T X^*)}{1+\exp(\beta^T X^*)}$, where $X^* = (1, X^T)^T$. Accordingly, from the theory just developed, we know that the influence functions of all RAL estimators for β can be derived as the solution to the generalized estimating equations

$$\sum_{i=1}^{n} A(X_i)\{Y_i - \mu(X_i, \beta)\} = 0$$

for arbitrary $A^{q\times 1}(X)$, and the efficient estimator is obtained by choosing $A(X) = D^T(X)V^{-1}(X)$, where $D(X) = \frac{\partial \mu(X,\beta_0)}{\partial \beta^T}$ and $V(X) = \text{var}(Y|X)$. Because Y is binary,

$$V(X) = \text{var}(Y|X) = \mu(X, \beta_0)\{1 - \mu(X, \beta_0)\} = \frac{\exp(\beta_0^T X^*)}{\{1 + \exp(\beta_0^T X^*)\}^2}.$$

Taking derivatives, we also obtain that $D^T(X) = X^* V(X)$. Hence the optimal estimator for β can be derived by choosing $A(X) = D^T(X)V^{-1}(X) = X^*$, leading us to the optimal estimating equation

$$\sum_{i=1}^{n} X_i^* \left\{ Y_i - \frac{\exp(\beta^T X_i^*)}{1 + \exp(\beta^T X_i^*)} \right\} = 0. \tag{4.65}$$

Because the conditional distribution of Y given X is fully described with the finite number of parameters β, we could also estimate the parameter β using maximum likelihood. It is an easy exercise to show that the solution to (4.65) also leads to the maximum likelihood estimator for β. □

Example 2. Log-linear model
In Section 4.1, we considered a log-linear model to model CD4 count as a function of covariates; see (4.10). We also proposed an ad hoc semiparametric GEE estimator for β as the solution to (4.11) without any motivation. Let us study this problem more carefully. We know that the optimal semiparametric estimator for β for model (4.10) is given as the solution to the equation

$$\sum_{i=1}^{n} D^T(X_i, \beta)V^{-1}(X_i)\{Y_i - \mu(X_i, \beta)\} = 0, \tag{4.66}$$

where the gradient matrix $D(X, \beta)$ was derived in Section 4.1 for the log-linear model to be $D(X, \beta) = \frac{\partial \mu(X,\beta)}{\partial \beta^T} = \mu(X, \beta)(1, X^T)$. Although the semiparametric restricted moment model makes no assumptions about the variance function $V(X) = \text{var}(Y|X)$, we argued that to find a locally efficient estimator for β we might want to make some assumptions regarding the function $V(X)$ and derive an adaptive estimator.

Since CD4 count is a count we might be willing to assume that it follows a Poisson distribution. If indeed the distribution of Y given X follows a Poisson distribution with mean $\mu(X, \beta)$, then we immediately know that $V(X) = \mu(X, \beta)$. Although this is probably too strong an assumption to make in general, we may believe that a good approximation is that the variance function $V(X)$ is at least proportional to the mean; i.e., that $V(X) = \sigma^2\mu(X, \beta)$, where σ^2 is some unknown scale factor. In that case, $D^T(X, \beta)V^{-1}(X) = \sigma^{-2}(1, X^T)^T$ and the locally efficient estimator for β would be the solution to (4.66), which, up to a proportionality constant, would be the solution to the estimating equation

$$\sum_{i=1}^{n}(1, X_i^T)^T\{Y_i - \mu(X_i, \beta)\} = 0, \tag{4.67}$$

which is the same as the estimator proposed in Section 4.1; see (4.11).

Thus, we have shown that the locally efficient semiparametric estimator for β, when the conditional distribution of the response variable given the covariates follows a Poisson distribution or, more generally, if the conditional variance of the response variable given the covariates is proportional to the conditional mean of Y given the covariates, is given by (4.67). For a more detailed discussion on log-linear models, see McCullagh and Nelder (1989, Chapter 6).

If the conditional variance of Y given X is not proportional to the conditional mean $\mu(X, \beta)$, then the estimator (4.67) is no longer semiparametric

efficient; nonetheless, it will still be a consistent, asymptotically normal estimator for β with an asymptotic variance that can be estimated using the sandwich estimator (4.14).

Another possibility, which would give somewhat greater flexibility, is to model the variance function using (4.64), as this model contains the Poisson variance structure mentioned above, and use the adaptive methods described in this section to estimate β. □

Extensions of the Restricted Moment Model

When considering the restricted moment model, we have concentrated on models where $E(Y|X) = \mu(X, \beta)$ or, equivalently, $E\{\varepsilon(Y, X, \beta)|X\} = 0$, where $\varepsilon(Y, X, \beta) = Y - \mu(X, \beta)$. Using this second representation, the theory developed for the restricted moment problem can be applied (or extended) to models where $E\{\varepsilon(Y, X, \beta)|X\} = 0$ for arbitrary functions $\varepsilon(Y, X, \beta)$.

This allows us to consider models, for example, where we model both the conditional variance and the conditional mean of Y given X. Say we want a model where we assume that $E(Y|X) = \mu(X, \beta)$ and $\text{var}(Y|X) = V(X, \beta, \xi)$ and our interest is in estimating the parameters β and ξ using a sample of data that are realizations of $(Y_i, X_i), i = 1, \ldots, n$. For simplicity, take Y to be a univariate response random variable, although this can be easily generalized to multivariate response random vectors as well. We could then define the bivariate vector $\varepsilon(Y, X, \beta, \xi) = \{\varepsilon_1(Y, X, \beta, \xi), \varepsilon_2(Y, X, \beta, \xi)\}^T$, where

$$\varepsilon_1(Y, X, \beta, \xi) = Y - \mu(X, \beta)$$

and

$$\varepsilon_2(Y, X, \beta, \xi) = \{Y - \mu(X, \beta)\}^2 - V(X, \beta, \xi).$$

With such a representation, it is clear that our model for the conditional mean and conditional variance of Y given X is equivalent to $E\{\varepsilon(Y, X, \beta, \xi)|X\} = 0$.

This representation also allows us to consider models for the conditional quantiles of Y as a function of X. For example, suppose we wanted to consider a model for the median of a continuous random variable Y as a function of X. Say we wanted a model where we assumed that the conditional median of Y given X was equal to $\mu(X, \beta)$ and we wanted to estimate β using a sample of data $(Y_i, X_i), i = 1, \ldots, n$. This could be accomplished by considering $\varepsilon(Y, X, \beta) = I\{Y \leq \mu(X, \beta)\} - .5$ because the conditional expectation of $\varepsilon(Y, X, \beta)$ given X is given by

$$E\{\varepsilon(Y, X, \beta)|X\} = P\{Y \leq \mu(X, \beta)|X\} - .5$$

and, by definition, the conditional median is the value $\mu(X, \beta)$ such that the conditional probability that Y is less than or equal to $\mu(X, \beta)$, given X, is equal to .5, which would imply that $E\{\varepsilon(Y, X, \beta)|X\} = 0$.

Using arguments similar to those developed in this chapter, we can show that we can restrict attention to semiparametric estimators that are solutions to the estimating equations

$$\sum_{i=1}^{n} A(X_i)\varepsilon(Y_i, X_i, \beta) = 0$$

for arbitrary $A(X)$ and that the efficient semiparametric estimator for β falls within this class. For models where we include both the conditional mean and conditional variance, this leads to the so-called quadratic estimating equations or GEE2. We give several exercises along these lines.

4.7 Exercises for Chapter 4

1. Prove that the linear subspaces $\mathscr{T}_j, j = 1, \ldots, m$ are mutually orthogonal subspaces, where $\mathscr{T}_j$ is defined by (4.21) of Theorem 4.5. That is, show that h_j is orthogonal to $h_{j'}$, where $h_j \in \mathscr{T}_j$, $h_{j'} \in \mathscr{T}_{j'}$ and $j \neq j^{'}$, $j, j^{'} = 1, \ldots, m$.
2. Let Y be a one-dimensional response random variable. Consider the model

$$Y = \mu(X, \beta) + \varepsilon,$$

where $\beta \in \mathbb{R}^q$, and $E\{h(\varepsilon)|X\} = 0$ for some arbitrary function $h(\cdot)$. Up to now, we considered the identity function $h(\varepsilon) = \varepsilon$, but this can be generalized to arbitrary $h(\varepsilon)$. For example, if we define $h(\varepsilon) = \{I(\varepsilon \leq 0) - 1/2\}$, then this is the median regression model. That is, if we define $F(y|x) = P(Y \leq y|X = x)$, then med $(Y|x) = F^{-1}(1/2, x)$, the value $m(x)$ such that $F(m(x)|x) = 1/2$. Therefore, the model with this choice of $h(\cdot)$ is equivalent to

$$\text{med}\,(Y|X) = \mu(X, \beta).$$

Assume no other restrictions are placed on the model but $E\{h(\varepsilon)|X\} = 0$ for some function $h(\cdot)$. For simplicity, assume h is differentiable, but this can be generalized to nondifferentiable h such as in median regression.
   a) Find the space $\Lambda^{\perp}$ (i.e., the space perpendicular to the nuisance tangent space).
   b) Find the efficient score vector for this problem.
   c) Describe how you would construct a locally efficient estimator for β from a sample of data $(Y_i, X_i), i = 1, \ldots, n$.
   d) Find an estimator for the asymptotic variance of the estimator defined in part (c).
3. Letting Y be a one-dimensional response variable, consider the semiparametric model where

$$E(Y|X) = \mu(X, \beta)$$

and

$$\text{Var}\,(Y|X) = V(X, \beta), \quad \beta \in \mathbb{R}^q,$$

where $V(x, \beta) > 0$ for all x and β.

a) Derive the space $\Lambda^{\perp}$ (i.e., the space orthogonal to the nuisance tangent space).
b) Find the efficient score vector.
c) Derive a locally efficient estimator for β using a sample of data $(Y_i, X_i), i = 1, \ldots, n$.

5

Other Examples of Semiparametric Models

In this chapter, we will consider other widely used semiparametric models. We begin with the "location-shift" regression model and the proportional hazards regression model. These models, like the restricted moment regression model discussed in detail in Chapter 4, are best represented through a parameter of interest β and an infinite-dimensional nuisance parameter η. This being the case, the class of influence functions of RAL estimators for β will be defined as elements of a Hilbert space orthogonal to the nuisance tangent space. Later in the chapter, we will also discuss the problem of estimating the mean treatment difference between two treatments in a randomized pretest-posttest study or, more generally, in a randomized study with covariate adjustment. We will show that this seemingly easy problem can be cast as a semiparametric problem for which the goal is to find an efficient estimator for the mean treatment difference. We will see that this problem is best represented by defining the model through an infinite-dimensional parameter θ, where the parameter of interest, the mean treatment difference, is given by $\beta(\theta)$, a function of θ. With this representation, it will be more convenient to define the tangent space and its orthogonal complement and then find the efficient estimator by considering the residual after projecting the influence function of a simple, but inefficient, RAL estimator onto the orthogonal complement of the tangent space. This methodology will be described in detail in Section 5.4 that follows.

5.1 Location-Shift Regression Model

Let Y_i denote a continuous response variable and X_i a vector of covariates measured on the i-th individual of an iid sample $i = 1, \ldots, n$. For simplicity, we will only consider a univariate response variable here, but all the arguments that follow could be generalized to multivariate response variables as well.

A popular way of modeling the relationship of Y_i as a function of X_i is with a location-shift regression model. That is, consider the model where

$$Y_i = \mu(X_i, \beta) + \varepsilon_i,\ i = 1, \ldots, n,$$

$(Y_i, X_i), i = 1, \ldots, n$ are iid, Y_i is assumed to be continuous, and ε_i (also continuous) is independent of X_i (denoted as $\varepsilon_i \perp\!\!\!\perp X_i$). These models may be linear or nonlinear models depending on whether $\mu(X_i, \beta)$ is linear in β or not. The semiparametric properties of this model were studied by Bickel et al. (1993, Section 4.3) and Manski (1984).

In this model, there is an assumed basic underlying distributional shape that the density of the response variable Y_i follows, that of the distribution of ε_i, but where the location of this distribution is determined according to the value of the covariates X_i (i.e., shifted by $\mu(X_i, \beta)$). The q-dimensional parameter β determines the magnitude of the shift in the location of the distribution as a function of the covariates, and it is this parameter that is of primary interest. We make no additional assumptions on the distribution of ε_i or X_i. To avoid identifiability problems, we will also assume that if $(\alpha_1, \beta_1) \neq (\alpha_2, \beta_2)$, then

$$\alpha_1 + \mu(X, \beta_1) \neq \alpha_2 + \mu(X, \beta_2),$$

where α_1, α_2 are any scalar constants and β_1, β_2 are values in the parameter space contained in $\mathbb{R}^q$.

For example, if we consider a linear model where $\mu(X_i, \beta) = X_i^T \beta$, we must make sure not to include an intercept term, as this will be absorbed into the error term ε_i; i.e.,

$$Y_i = \beta_1 X_{i1} + \cdots + \beta_q X_{iq} + \varepsilon_i,\ i = 1, \ldots, n. \tag{5.1}$$

The location-shift regression model is semiparametric because no restrictions are placed on the distribution of ε_i or X_i. Such a model can be characterized by

$$\{\beta, p_\varepsilon(\varepsilon), p_X(x)\},$$

where $p_\varepsilon(\varepsilon)$ and $p_X(x)$ are arbitrary densities of ε and X, respectively. We assume that ε and hence Y is a continuous random variable with dominating Lebesgue measure ℓ_ε or ℓ_Y, respectively. The dominating measure for X is denoted by ν_X. An arbitrary density in this model for a single observation is given by

$$p_{Y,X}(y, x) = p_\varepsilon\{y - \mu(x, \beta)\} p_X(x)$$

with respect to the dominating measure $\ell_Y \times \nu_X$.

It has been my experience that there is confusion between the location-shift regression model and the restricted moment model. In many introductory courses in statistics, a linear regression model is defined as

$$Y_i = \alpha + X_i^T \beta + \varepsilon_i,\ i = 1, \ldots, n, \tag{5.2}$$

where $\varepsilon_i,\ i = 1, \ldots, n$ are assumed to be iid mean-zero random variables. Sometimes, in addition, it may be assumed that ε_i are normally distributed

with mean zero and some common variance σ^2. What is often not made clear is that there is an implicit assumption that the covariates $X_i,\ i = 1, \ldots, n$ are fixed. Consequently, the error terms ε_i being iid implies that the distribution of ε_i is independent of X_i. Thus, such models are examples of what we are now calling the location-shift regression models. In contrast, a linear restricted moment model can also be written as (5.2), where $\varepsilon_i,\ i = 1, \ldots, n$ are iid random variables. However, the restricted moment model makes the assumption that $E(\varepsilon_i|X_i) = 0$, which then implies that $E(\varepsilon_i) = 0$ but does not necessarily assume that ε_i is independent of X_i.

The location-shift regression model, although semiparametric, is more restrictive than the restricted moment model considered in Chapter 4. For example, if we consider the linear restricted moment model

$$Y_i = \alpha + \beta_1 X_{i1} + \cdots + \beta_q X_{iq} + \varepsilon_i,$$

where

$$E\ (\varepsilon_i|X_i) = 0,$$

or equivalently

$$Y_i = \beta_1 X_{i1} + \cdots + \beta_q X_{iq} + (\alpha + \varepsilon_i),$$

where

$$E\{\alpha + \varepsilon_i|X_i\} = \alpha,$$

then this model includes a larger class of probability distributions than the linear location-shift regression model; namely,

$$Y_i = \beta_1 X_{i1} + \ldots + \beta_q X_{iq} + \varepsilon_i^*,$$

where ε_i^* is independent of X_i. Because of independence, we obtain $E(\varepsilon_i^*|X_i) = E(\varepsilon_i^*) =$ constant, which satisfies the linear moment restriction, but, conversely, $E(\alpha + \varepsilon_i|X_i) = \alpha$, which holds true for the restricted moment model, does not imply that $(\alpha + \varepsilon_i) = \varepsilon_i^* \perp\!\!\!\perp X_i$.

Since the location-shift regression model is more restrictive than the restricted moment model, we would expect the class of semiparametric RAL estimators for β for the location-shift regression model to be larger than the class of semiparametric RAL estimators for β for the restricted moment model and the semiparametric efficiency bound for the location-shift regression model to be smaller than the semiparametric efficiency bound for the restricted moment model.

The Nuisance Tangent Space and Its Orthogonal Complement for the Location-Shift Regression Model

The key to finding semiparametric RAL estimators for β and identifying the efficient such estimator is to derive the space of influence functions of RAL estimators for β. This will require us to find the space orthogonal to the

nuisance tangent space. The nuisance tangent space is defined as the mean-square closure of all parametric submodel nuisance tangent spaces. The nuisance tangent space and its orthogonal complement for the semiparametric location-shift regression model are given as follows.

Theorem 5.1. Using the convention that $\varepsilon(\beta) = Y - \mu(X, \beta)$ and $\varepsilon = \varepsilon(\beta_0) = Y - \mu(X, \beta_0)$, the nuisance tangent space for the location-shift regression model is given by

$$\Lambda = \Lambda_{1s} \oplus \Lambda_{2s},$$

where

$$\Lambda_{1s} = \left[a_1^{q\times 1}(\varepsilon) : E\{a_1^{q\times 1}(\varepsilon)\} = 0^{q\times 1} \right],$$

$$\Lambda_{2s} = \left[a_2^{q\times 1}(X) : E\{a_2^{q\times 1}(X)\} = 0^{q\times 1} \right],$$

and $\Lambda_{1s} \perp \Lambda_{2s}$.

Proof. Consider the parametric submodel with density

$$p_\varepsilon\{y - \mu(x, \beta), \gamma_1\} p_X(x, \gamma_2), \tag{5.3}$$

where γ_{10} and γ_{20} denote the "truth." If we fix β at the truth β_0, then $p_\varepsilon(\varepsilon, \gamma_1)$ allows for an arbitrary marginal density of ε. Consequently, using logic developed in Chapter 4, the mean-square closure for parametric submodel nuisance tangent spaces

$$\Lambda_{\gamma_1} = \{B^{q\times r_1} S_{\gamma_1}(\varepsilon) \text{ for all } B^{q\times r_1}\},$$

where $S_{\gamma_1}(\varepsilon) = \partial \log p_\varepsilon(\varepsilon, \gamma_{10})/\partial\gamma_1$, is the space Λ_{1s}, defined as

$$\Lambda_{1s} = \left[a_1^{q\times 1}(\varepsilon) : E\{a_1^{q\times 1}(\varepsilon)\} = 0^{q\times 1} \right].$$

Similarly, the mean-square closure for parametric submodel nuisance tangent spaces

$$\Lambda_{\gamma_2} = \{B^{q\times r_2} S_{\gamma_2}(X) \text{ for all } B^{q\times r_2}\},$$

where $S_{\gamma_2}(x) = \partial \log p_X(x, \gamma_{20})/\partial\gamma_2$, is the space Λ_{2s}, defined as

$$\Lambda_{2s} = \left[a_2^{q\times 1}(X) : E\{a_2^{q\times 1}(X)\} = 0^{q\times 1} \right].$$

Since the density (5.3) is a product involving variationally independent parameters γ_1 and γ_2, the nuisance tangent space (i.e., the mean-square closure of all parametric submodel nuisance tangent spaces associated with arbitrary nuisance parameters γ_1 and γ_2) is given by

$$\Lambda = \Lambda_{1s} \oplus \Lambda_{2s}.$$

Because ε is independent of X, it is easy to verify that $\Lambda_{1s} \perp \Lambda_{2s}$. □

Influence functions of RAL estimators for β lie in the space orthogonal to the nuisance tangent space. We can find all elements of $\Lambda^{\perp}$ by considering

$$\Lambda^{\perp} = [\{h - \Pi\ (h|\Lambda)\} \text{ for all } h \in \mathcal{H}]\,.$$

Because $\Lambda_{1s} \perp \Lambda_{2s}$, we obtain the following intuitive result.

Theorem 5.2. Let $\Lambda = \Lambda_{1s} \oplus \Lambda_{2s}$, where the closed linear subspaces Λ_{1s} and Λ_{2s} are orthogonal; i.e., $\Lambda_{1s} \perp \Lambda_{2s}$. Then

$$\Pi(h|\Lambda) = \Pi(h|\Lambda_{1s}) + \Pi(h|\Lambda_{2s}). \tag{5.4}$$

Proof. The projection of any element $h \in \mathcal{H}$ onto the closed linear space Λ is the unique element $\Pi(h|\Lambda)$ such that the residual $h - \Pi(h|\Lambda)$ is orthogonal to every element in

$$\Lambda = \{(a_1 + a_2),\ \text{where } a_1 \in \Lambda_{1s} \text{ and } a_2 \in \Lambda_{2s}\}.$$

To verify that (5.4) is true, we first note that $\Pi(h|\Lambda_{1s}) \in \Lambda_{1s}$ and $\Pi(h|\Lambda_{2s}) \in \Lambda_{2s}$, implying that $\Pi(h|\Lambda_{1s}) + \Pi(h|\Lambda_{2s}) \in \Lambda$. To complete the proof, we must show that the inner product

$$\langle h - \Pi(h|\Lambda_{1s}) - \Pi(h|\Lambda_{2s}), a_1 + a_2 \rangle = 0 \tag{5.5}$$

for all $a_1 \in \Lambda_{1s}$ and $a_2 \in \Lambda_{2s}$. The inner product (5.5) can be written as

$$\begin{aligned} &\langle h - \Pi(h|\Lambda_{1s}), a_1 \rangle && (5.6)\\ -\ &\langle \Pi(h|\Lambda_{2s}), a_1 \rangle && (5.7)\\ +\ &\langle h - \Pi(h|\Lambda_{2s}), a_2 \rangle && (5.8)\\ -\ &\langle \Pi(h|\Lambda_{1s}), a_2 \rangle. && (5.9) \end{aligned}$$

Since $h - \Pi(h|\Lambda_{1s})$ is orthogonal to Λ_{1s} and $h - \Pi(h|\Lambda_{2s})$ is orthogonal to Λ_{2s}, this implies that the inner products (5.6) and (5.8) are equal to zero. Also, since $\Pi(h|\Lambda_{2s}) \in \Lambda_{2s}$ and $\Pi(h|\Lambda_{1s}) \in \Lambda_{1s}$, then $\Lambda_{1s} \perp \Lambda_{2s}$ implies that the inner products (5.7) and (5.9) are also equal to zero. □

In the proof of Lemma 4.3, we showed that for any $h(\varepsilon, X) \in \mathcal{H}$ the projection of $h(\varepsilon, X)$ onto Λ_{2s} is given by $\Pi(h|\Lambda_{2s}) = E\{h(\varepsilon, X)|X\}$. Similarly, we can show that $\Pi(h|\Lambda_{1s}) = E\{h(\varepsilon, X)|\varepsilon\}$. Consequently, the space orthogonal to the nuisance tangent space is given by

$$\Lambda^{\perp} = \Big([h(\varepsilon, X) - E\{h(\varepsilon, X)|\varepsilon\} - E\{h(\varepsilon, X)|X\}] \text{ for all } h \in \mathcal{H}\Big). \tag{5.10}$$

Semiparametric Estimators for β

Since influence functions of RAL estimators for β are, up to a proportionality constant, defined by the elements orthogonal to the nuisance tangent space, we now use the result in (5.10) that defines elements orthogonal to the nuisance tangent space to aid us in finding semiparametric estimators of β in the location-shift regression model.

We begin by considering any arbitrary q-dimensional function of ε, X, say $g^{q\times 1}(\varepsilon, X)$. In order that this be an arbitrary element of $\mathcal{H}$, it must have mean zero. Therefore, we center $g(\varepsilon, X)$ and define

$$h(\varepsilon, X) = g(\varepsilon, X) - E\{g(\varepsilon, X)\}.$$

We now consider an arbitrary element of $\Lambda^{\perp}$, which by (5.10) is given by $\{h - \Pi\ (h|\Lambda)\}$, which equals

$$g(\varepsilon, X) - G_{\varepsilon}(X) - G_X(\varepsilon) + E\{g(\varepsilon, X)\}, \tag{5.11}$$

where $G_{\varepsilon}(X) = E\{g(\varepsilon, X)|X\}$, $G_X(\varepsilon) = E\{g(\varepsilon, X)|\varepsilon\}$, and the function $g(\varepsilon, X)$ should not equal $g_1(\varepsilon) + g_2(X)$ in order to ensure that the residual (5.11) is not trivially equal to zero.

Because of the independence of ε and X, we obtain, for fixed $X = x$, that $G_{\varepsilon}(x) = E\{g(\varepsilon, x)\}$ and, for fixed $\varepsilon = \varepsilon^*$, that $G_X(\varepsilon^*) = E\{g(\varepsilon^*, X)\}$. Consequently, consistent and unbiased estimators for $G_{\varepsilon}(x)$ and $G_X(\varepsilon^*)$ are given by

$$\hat{G}_{\varepsilon}(x) = n^{-1} \sum_{i=1}^{n} g(\varepsilon_i, x), \tag{5.12}$$

$$\hat{G}_X(\varepsilon^*) = n^{-1} \sum_{i=1}^{n} g(\varepsilon^*, X_i), \tag{5.13}$$

respectively.

Because influence functions of RAL estimators for β are proportional to elements orthogonal to the nuisance space, if we knew $G_{\varepsilon}(x)$, $G_X(\varepsilon^*)$, and $E\{g(\varepsilon, X)\}$, then a natural estimator for β would be obtained by solving the estimating equation

$$\sum_{i=1}^{n} [g\{\varepsilon_i(\beta), X_i\} - G_{\varepsilon}(X_i) - G_X\{\varepsilon_i(\beta)\} + E\{g(\varepsilon, X)\}] = 0^{q\times 1},$$

where $\varepsilon_i(\beta) = Y_i - \mu(X_i, \beta)$. Since $G_{\varepsilon}(x)$, $G_X(\varepsilon^*)$, and $E\{g(\varepsilon, X)\}$ are not known, a natural strategy for obtaining an estimator for β is to substitute estimates of these quantities in the preceding estimating equation, leading to the estimating equation

$$\sum_{i=1}^{n} \Big(g\{\varepsilon_i(\beta), X_i\} - \hat{G}_{\varepsilon(\beta)}(X_i) - \hat{G}_X\{\varepsilon_i(\beta)\} + \hat{E}[g\{\varepsilon(\beta), X\}] \Big) = 0^{q\times 1}, \tag{5.14}$$

where $\hat{G}_{\varepsilon(\beta)}(X_i)$, $\hat{G}_X\{\varepsilon_i(\beta)\}$ are defined by (5.12) and (5.13), respectively, and $\hat{E}[g\{\varepsilon(\beta), X\}] = n^{-1}\sum_{i=1}^{n} g\{\varepsilon_i(\beta), X_i\}$.

The estimator for β that solves (5.14) should be a consistent and asymptotically normal semiparametric estimator for β with influence function proportional to (5.11). Rather than trying to prove the asymptotic properties of this estimator, we will instead focus on deriving a class of locally efficient estimators for β and investigate the asymptotic properties of this class of estimators.

Efficient Score for the Location-Shift Regression Model

The efficient estimator for β has an influence function proportional to the efficient score (i.e., the residual after projecting the score vector with respect to β onto the nuisance tangent space). To obtain the score vector with respect to β, we consider the density for a single observation (Y, X), which is given by

$$p_{Y,X}(y, x, \beta) = p_\varepsilon\{y - \mu(x, \beta)\}p_X(x).$$

Therefore,

$$\begin{aligned} S_\beta^{q\times 1}(y, x) &= \left.\frac{\partial \log p_{Y,X}(y, x, \beta)}{\partial \beta}\right|_{\beta=\beta_0} \\ &= -D^{T^{q\times 1}}(x, \beta_0) S_\varepsilon(\varepsilon), \end{aligned}$$

where

$$D(x, \beta_0) = \frac{\partial \mu(x, \beta_0)}{\partial \beta^T}$$

and

$$S_\varepsilon(\varepsilon) = \frac{\partial \log p_\varepsilon(\varepsilon)}{d\varepsilon}. \tag{5.15}$$

We first prove that $E\{S_\varepsilon(\varepsilon)\} = 0$.

Theorem 5.3. If the random variable ε is continuous with support on the real line, then

$$E\{S_\varepsilon(\varepsilon)\} = 0,$$

where $S_\varepsilon(\varepsilon)$ is defined by (5.15).

Proof. Because ε is a continuous random variable whose distribution is dominated by Lebesgue measure and has support on the real line, this means that $\int p_\varepsilon(\varepsilon - \mu)d\varepsilon = 1$ for all scalar constants μ. This implies that $d/d\mu\{\int p_\varepsilon(\varepsilon - \mu)d\varepsilon\} = 0$ for all μ. Interchanging differentiation and integration, we obtain $\int -p'_\varepsilon(\varepsilon - \mu)d\varepsilon = 0$ for all μ, where $p'_\varepsilon(x) = dp_\varepsilon(x)/dx$. Multiplying and dividing by $p_\varepsilon(\varepsilon)$ and taking $\mu = 0$, we obtain $-\int S_\varepsilon(\varepsilon)p_\varepsilon(\varepsilon)d\varepsilon = 0$, or $E\{S_\varepsilon(\varepsilon)\} = 0$. □

Because $E\{S_\varepsilon(\varepsilon)\} = 0$, we use (5.11) to deduce that the efficient score is given by

$$\begin{aligned} S_\beta(\varepsilon, X) - \Pi\left[S_\beta|\Lambda\right] &= -D^T(X,\beta_0)S_\varepsilon(\varepsilon) + E\{D^T(X,\beta_0)\}S_\varepsilon(\varepsilon) \\ &= -\{D^T(X,\beta_0) - E\{D^T(X,\beta_0)\}\}S_\varepsilon(\varepsilon). \end{aligned} \tag{5.16}$$

If the distributions of X, $p_X(x)$, and ε, $p_\varepsilon(\varepsilon)$, were known to us, then we would also know $E\{D^T(X,\beta)\}$ and $S_\varepsilon(\varepsilon)$. If this were the case, then (5.16) suggests finding the efficient estimator for β by solving the equation

$$\sum_{i=1}^{n}[D^T(X_i,\beta) - E\{D^T(X,\beta)\}]S_\varepsilon\{Y_i - \mu(X_i,\beta)\} = 0. \tag{5.17}$$

Note 1. The sign in (5.16) was reversed, but this is not important, as the estimator remains the same. □

However, since $E\{D^T(X,\beta)\}$ is not known to us, a natural strategy is to substitute an estimator for $E\{D^T(X,\beta)\}$ in (5.17), leading to the estimator for β that solves the estimating equation

$$\sum_{i=1}^{n}\{D^T(X_i,\beta) - \bar{D}^T(\beta)\}S_\varepsilon\{Y_i - \mu(X_i,\beta)\} = 0, \tag{5.18}$$

where $\bar{D}^T(\beta) = n^{-1}\sum_{i=1}^{n} D^T(X_i,\beta)$.

Locally Efficient Adaptive Estimators

The function $S_\varepsilon(\varepsilon)$ depends on the underlying density $p_\varepsilon(\varepsilon)$, which is unknown to us. Consequently, we may posit some underlying density for ε to start with, some working density $p_\varepsilon(\varepsilon)$ that may or may not be correct. Therefore, in what follows, we consider the asymptotic properties of the estimator for β, denoted by $\hat{\beta}_n$, which is the solution to equation (5.18) for an arbitrary function of ε, $S_\varepsilon(\varepsilon)$, which may not be a score function or, for that matter, may not have mean zero at the truth; i.e., $E\{S_\varepsilon(\varepsilon)\} \neq 0$. To emphasize the fact that we may not be using the correct score function $S_\varepsilon(\varepsilon)$, we will substitute an arbitrary function of ε, which we denote by $\kappa(\varepsilon)$, for $S_\varepsilon(\varepsilon)$ in equation (5.18).

We now investigate (heuristically) the asymptotic properties of the estimator $\hat{\beta}_n$, which solves (5.18) for an arbitrary function $\kappa(\cdot)$ substituted for $S_\varepsilon(\cdot)$.

Theorem 5.4. The estimator $\hat{\beta}_n$, which is the solution to the estimating equation

$$\sum_{i=1}^{n}\{D^T(X_i,\beta) - \bar{D}^T(\beta)\}\kappa\{Y_i - \mu(X_i,\beta)\} = 0, \tag{5.19}$$

is asymptotically normal. That is,

$$n^{1/2}(\hat{\beta}_n - \beta_0) \xrightarrow{\mathcal{D}} N(0, \Sigma),$$

where

$$\Sigma = E\left\{\frac{\partial \kappa(\varepsilon)}{\partial \varepsilon}\right\}^{-2} \operatorname{var}\{\kappa(\varepsilon)\}\left[\operatorname{var}\left\{D^T(X, \beta_0)\right\}\right]^{-1}.$$

Proof. To derive the asymptotic distribution of $\hat{\beta}_n$, we expand $\hat{\beta}_n$ about β_0 in equation (5.19) and, after multiplying by $n^{-1/2}$, we obtain

$$\begin{aligned}
0 &= n^{-1/2}\sum_{i=1}^{n}\{D^T(X_i, \hat{\beta}_n) - \bar{D}^T(\hat{\beta}_n)\}\kappa\{Y_i - \mu(X_i, \hat{\beta}_n)\} \\
&= n^{-1/2}\sum_{i=1}^{n}\{D^T(X_i, \beta_0) - \bar{D}^T(\beta_0)\}\kappa\{Y_i - \mu(X_i, \beta_0)\} \qquad (5.20) \\
&\quad + \left\{n^{-1}\sum_{i=1}^{n}\frac{\partial}{\partial \beta^T}\left[\left\{D^T(X_i, \beta_n^*) - \bar{D}^T(\beta_n^*)\right\}\kappa\{Y_i - \mu(X_i, \beta_n^*)\}\right]\right\} \qquad (5.21) \\
&\quad n^{1/2}(\hat{\beta}_n - \beta_0),
\end{aligned}$$

where β_n^* is an intermediate value between $\hat{\beta}_n$ and β_0. Consequently,

$$\begin{aligned}
n^{1/2}(\hat{\beta}_n - \beta_0) = \Bigg\{-n^{-1}\sum_{i=1}^{n}\frac{\partial}{\partial \beta^T}\Big[\left\{D^T(X_i, \beta_n^*) - \bar{D}^T(\beta_n^*)\right\} \\
\kappa\{Y_i - \mu(X_i, \beta_n^*)\}\Big]\Bigg\}^{-1} \times n^{-1/2}\sum_{i=1}^{n}\left\{D^T(X_i, \beta_0) - \bar{D}^T(\beta_0)\right\} \\
\kappa\{Y_i - \mu(X_i, \beta_0)\}. \qquad (5.22)
\end{aligned}$$

Under suitable regularity conditions, the sample average (5.21) will converge in probability to

$$E\left[\left\{D^T(X, \beta_0) - E\{D^T(X, \beta_0)\}\right\}^{q\times 1}\left(\frac{\partial \kappa(\varepsilon)}{\partial \varepsilon}\right)D(X, \beta_0)^{1\times q}\right] \qquad (5.23)$$

$$+ E\left[\left\{\frac{\partial}{\partial \beta^T}D^T(X, \beta_0) - E\left(\frac{\partial}{\partial \beta^T}D^T(X, \beta_0)\right)\right\}\kappa(\varepsilon)\right]. \qquad (5.24)$$

Because of the independence of ε and X, (5.23) is equal to

$$E\left\{\frac{\partial \kappa(\varepsilon)}{\partial \varepsilon}\right\}\operatorname{var}\left\{D^T(X, \beta_0)\right\},$$

where

$$\operatorname{var}\{D^T(X, \beta_0)\} = E\{D^T(X, \beta_0)D(X, \beta_0)\} - E\{D^T(X, \beta_0)\}E\{D(X, \beta_0)\}$$

is the variance matrix of $D^T(X,\beta_0)$, and (5.24) = 0. Therefore (5.22) can be written as

$$n^{1/2}(\hat{\beta}_n-\beta_0)=\left[-E\left\{\frac{\partial\kappa(\varepsilon)}{\partial\varepsilon}\right\}\mathrm{var}\left\{D^T(X,\beta_0)\right\}\right]^{-1} n^{-1/2}\sum_{i=1}^{n}\left\{D^T(X_i,\beta_0)-\bar{D}^T(\beta_0)\right\}\kappa(\varepsilon_i)+o_p(1). \quad (5.25)$$

Note that

$$\begin{aligned} & n^{-1/2}\sum_{i=1}^{n}\left\{D^T(X_i,\beta_0)-\bar{D}^T(\beta_0)\right\}\kappa(\varepsilon_i) \\ &\qquad = n^{-1/2}\sum_{i=1}^{n}\left\{D^T(X_i,\beta_0)-\bar{D}^T(\beta_0)\right\}\left\{\kappa(\varepsilon_i)-\bar{\kappa}\right\} \\ &= n^{-1/2}\sum_{i=1}^{n}\left\{D^T(X_i,\beta_0)-\mu_D\right\}\left\{\kappa(\varepsilon_i)-\mu_\kappa\right\} \\ &\quad + n^{1/2}\left\{\bar{D}^T(\beta_0)-\mu_D\right\}\left\{\bar{\kappa}-\mu_\kappa\right\} \\ &= n^{-1/2}\sum_{i=1}^{n}\left\{D^T(X_i,\beta_0)-\mu_D\right\}\left\{\kappa(\varepsilon_i)-\mu_\kappa\right\}+o_p(1), \end{aligned} \quad (5.26)$$

where

$$\bar{\kappa}=n^{-1}\sum_{i=1}^{n}\kappa(\varepsilon_i),\ \mu_D=E\left\{D^T(X_i,\beta_0)\right\},\ \text{and}\ \mu_\kappa=E\left\{\kappa(\varepsilon_i)\right\}.$$

Therefore, by (5.25) and (5.26), we obtain that

$$n^{1/2}(\hat{\beta}_n-\beta_0)=\left[-E\left\{\frac{d\kappa(\varepsilon)}{d\varepsilon}\right\}\mathrm{var}\left\{D^T(X,\beta_0)\right\}\right]^{-1} \times n^{-1/2}\sum_{i=1}^{n}\left\{D^T(X_i,\beta_0)-\mu_D\right\}\left\{\kappa(\varepsilon_i)-\mu_\kappa\right\}+o_p(1).$$

Consequently, the influence function of $\hat{\beta}_n$ is

$$\left[-E\left\{\frac{d\kappa(\varepsilon)}{\partial\varepsilon}\right\}\mathrm{var}\left\{D^T(X,\beta_0)\right\}\right]^{-1}\left\{D^T(X_i,\beta_0)-\mu_D\right\}\left\{\kappa(\varepsilon_i)-\mu_\kappa\right\}, \quad (5.27)$$

and $n^{1/2}(\hat{\beta}_n-\beta_0)$ is asymptotically normal with mean zero and variance matrix equal to the variance matrix of the influence function, which equals

$$E\left\{\frac{\partial\kappa(\varepsilon)}{\partial\varepsilon}\right\}^{-2}\mathrm{var}\{\kappa(\varepsilon)\}\left[\mathrm{var}\left\{D^T(X,\beta_0)\right\}\right]^{-1}. \ \square$$

Remark 1. If we start with an arbitrary function

$$g(\varepsilon, X) = D^T(X, \beta_0)\kappa(\varepsilon)$$

regardless of whether $\kappa(\varepsilon) = S_\varepsilon(\varepsilon) = \partial \log \frac{p_\varepsilon(\varepsilon)}{\partial \varepsilon}$ or some other function of ε, then by (5.11),

$$\begin{aligned} & D^T(X, \beta_0)\kappa(\varepsilon) - D^T(X, \beta_0)\mu_\kappa - \mu_D\kappa(\varepsilon) + \mu_D\mu_\kappa \\ = & \left\{D^T(X, \beta_0) - \mu_D\right\}\{\kappa(\varepsilon) - \mu_\kappa\} \end{aligned}$$

is orthogonal to the nuisance tangent space. We also notice that this is proportional to the influence function of $\hat{\beta}_n$ given by (5.27).

If, however, we choose the true density $p_\varepsilon(\varepsilon)$, then we obtain an efficient estimator. Consequently, the estimator for β given by (5.19) is a locally efficient semiparametric estimator for β. □

If we wanted to derive a globally efficient semiparametric estimator, then we would need to estimate the score function $\partial \log p_\varepsilon(\varepsilon)/\partial\varepsilon$ nonparametrically and substitute this for $\kappa(\varepsilon)$ in the estimating equation (5.19). Although this may be theoretically possible (see Bickel et al., 1993, Section 7.8), generally such nonparametric estimators are unstable (unless the sample size is very large). Another strategy would be to posit some parametric model, say $p_\varepsilon(\varepsilon, \xi)$, for the density of ε in terms of a finite number of parameters ξ. The parameter ξ can be estimated, say, by using pseudo-likelihood techniques; i.e., by maximizing the pseudo-likelihood,

$$\prod_{i=1}^{n} p_\varepsilon\{Y_i - \mu(X_i, \hat{\beta}_n^I), \xi\},$$

as a function in ξ using some initial consistent estimator $\hat{\beta}_n^I$ of β. Letting $\hat{\xi}_n$ denote such an estimator, we can estimate β by substituting $S_\varepsilon(\varepsilon, \hat{\xi}_n) = \partial \log p_\varepsilon(\varepsilon, \hat{\xi}_n)/\partial\varepsilon$ into equation (5.18). Such an adaptive estimator is locally efficient in the sense that if the true density of ε is an element of the posited parametric model, then the resulting estimator is efficient; otherwise, it will still be a consistent asymptotically normal semiparametric estimator for β. The idea here is that with a flexible parametric model, we can get a reasonably good approximation for the underlying density of ε, so even if we don't estimate this density consistently, it would hopefully be close enough so that the resulting estimator will have good efficiency properties.

In the following example, we consider the linear model and use this to contrast the estimators obtained for the restricted moment model and the location-shift regression model.

Example 1. Consider the linear model (5.1)

$$Y_i = X_i^T\beta + \varepsilon_i,$$

where Y_i is a one-dimensional random variable. For this model,

$$D^T(X_i, \beta) = X_i^{q \times 1}.$$

The estimator for β given by (5.18) is the solution to

$$\sum_{i=1}^{n} (X_i - \bar{X}) S_\varepsilon (Y_i - X_i^T \beta) = 0. \tag{5.28}$$

Suppose we believe our data are normally distributed; i.e., $\varepsilon_i \sim N(\mu_\varepsilon, \sigma^2)$ or

$$p_\varepsilon(\varepsilon) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{ -\frac{1}{2\sigma^2} (\varepsilon - \mu_\varepsilon)^2 \right\}.$$

Then

$$S_\varepsilon(\varepsilon) = -\left\{ \frac{\varepsilon - \mu_\varepsilon}{\sigma^2} \right\}.$$

Substituting into (5.28) yields

$$\sum_{i=1}^{n} (X_i - \bar{X}) \frac{(Y_i - X_i^T \beta - \mu_\varepsilon)}{\sigma^2} = 0.$$

Since σ^2 is a multiplicative constant and $\sum_{i=1}^{n} (X_i - \bar{X}) = 0$, the estimator for β is equivalent to solving

$$\sum_{i=1}^{n} (X_1 - \bar{X})(Y_i - X_i^T \beta) = 0.$$

This gives the usual least-squares estimator for the regression coefficients in a linear model with an intercept term; namely,

$$\hat{\beta}_n = \left\{ \sum_{i=1}^{n} (X_i - \bar{X})(X_i - \bar{X})^T \right\}^{-1} \sum_{i=1}^{n} (X_i - \bar{X}) Y_i. \tag{5.29}$$

We mentioned previously that the location-shift regression model is contained within the restricted moment model

$$Y_i = \alpha + \beta_1 X_{i1} + \cdots + \beta_q X_{iq} + \varepsilon_i,$$

where $E(\varepsilon_i | X_i) = 0$. For the location-shift regression model, ε_i is independent of X_i, which implies that the variance function

$$V(X_i) = \text{var}\ (Y_i | X_i) = \sigma^2 \ \text{(a constant independent of } X_i).$$

The efficient estimator for β among semiparametric estimators for β in the restricted moment model, given by (4.54), is

$$\sum_{i=1}^{n} D^T(X_i)V^{-1}(X_i)\{Y_i - \mu(X_i, \beta)\} = 0.$$

If $V(X_i)$ is assumed constant, then the efficient restricted moment estimator for $(\alpha, \beta^T)^T$ is given as the solution to

$$\sigma^{-2}\sum_{i=1}^{n}(1, X_i^T)^T(Y_i - \alpha - X_i^T\beta) = 0.$$

Some usual matrix algebra yields

$$\hat{\beta}_n = \left\{\sum_{i=1}^{n}(X_i - \bar{X})(X_i - \bar{X})^T\right\}^{-1}\sum_{i=1}^{n}(X_i - \bar{X})Y_i,$$

which is identical to the estimator (5.29), the locally efficient estimator for β in the location-shift regression model.

Remarks

1. The estimator $\hat{\beta}_n$ given by (5.29) is the efficient estimator for β among semiparametric estimators for the location-shift regression model if the error distribution is indeed normally distributed.
2. If, however, the error distribution was not normally distributed, other semiparametric estimators (semiparametric for the location-shift model) would be more efficient; namely, the solution to the equation

$$\sum_{i=1}^{n}(X_i - \bar{X})S_\varepsilon(Y_i - X_i^T\beta) = 0, \tag{5.30}$$

   where $S_\varepsilon(\varepsilon) = \frac{d\log p_\varepsilon(\varepsilon)}{d\varepsilon}$.
3. In contrast, among the estimators for β for the restricted moment model, when the variance function $V(X_i)$ is assumed constant, the efficient estimator for β is given by (5.29).
4. The estimator for β given by (5.30) may not be consistent and asymptotically normal if ε_i is not independent of X_i but $E(\varepsilon_i|X_i) = 0$, whereas (5.29) is.

5.2 Proportional Hazards Regression Model with Censored Data

The proportional hazards model, which was introduced by Cox (1972), is the most widely used model for analyzing survival data. This model is especially

useful for survival data that are right censored, as is often the case in many clinical trials where the primary endpoint is "time to an event" such as time to death, time to relapse, etc. As we will see shortly, the proportional hazards model is a semiparametric model that can be represented naturally with a finite number of parameters of interest and an infinite-dimensional nuisance parameter. Because of this natural representation with a parametric and nonparametric component, it was one of the first models to be studied using semiparametric theory, in the landmark paper of Begun et al. (1983).

In order to follow the arguments in this section, the reader must be familiar with the theory of counting processes and associated martingale processes that have been developed for studying the theoretical properties of many statistics used in censored survival analysis. Some excellent books for studying counting processes and their application to censored survival-data models include those by Fleming and Harrington (1991) and Andersen et al. (1992). The reader who has no familiarity with this area can skip this section, as it is self-contained and will not affect the understanding of the remainder of the book.

The primary goal of the proportional hazards model is to model the relationship of "time to event" as a function of a vector of covariates X. Throughout this section, we will often refer to the "time to event" as the "survival time" since, in many applications where these models are used, the primary endpoint is time to death. But keep in mind that the time to any event could be used. Let T denote the underlying survival time of an arbitrary individual in our population. The random variable T will be assumed to be a continuous positive random variable. Unlike previous models where we considered the conditional mean of the response variable, or shifts in the location of the distribution of the response variable as a function of covariates, in the proportional hazards model it is the hazard rate of failure that is modeled as a function of covariates X (q-dimensional). Specifically, the proportional hazards model of Cox (1972) assumes that

$$\begin{aligned}\lambda(u|X) &= \lim_{h\to 0}\left\{\frac{P(u \le T < u+h|T \ge u, X)}{h}\right\} \\ &= \lambda(u)\exp(\beta^T X), \end{aligned} \tag{5.31}$$

where $\lambda(u|x)$ is the conditional hazard rate of failure at time u given $X = x$. The baseline hazard function $\lambda(u)$ can be viewed as the underlying hazard rate if all the covariates X are equal to zero and this underlying hazard rate is left unspecified (nonparametric). This baseline hazard rate is assumed to be increased or decreased proportionately (the same proportionality constant through time) as a function of the covariates X through the relationship $\exp(\beta^T X)$. Consequently, the parameters β in this model measure the strength of the association of this proportionality increase or decrease in the hazard rate as a function of the covariates, and it is these parameters that will be of primary interest to us.

In many experimental settings where survival data are obtained, such as in clinical trials, not all the survival data are available for the individuals in the study. Some of the survival data may be right censored. That is, for some individuals, we may only know that they survived to some time. For example, in a clinical trial where patients enter the study during some accrual period and where the study is analyzed before all the patients die, a patient who has not died has a survival time that is right censored; that is, we only know that this patient survived the period of time between their entry into the study and the time the study was analyzed. This is an example of what is referred to as administrative censoring. Other reasons for censoring include patient dropout, where we only know that a patient was still alive at the time they dropped out. To accommodate censoring, we define the random variable C that corresponds to the potential time that an individual is followed for survival. We denote the conditional density of C given X as $p_{C|X}(c|x)$. In this setting, we observe, for each individual, the variables

$$\begin{aligned} V &= \min(T, C) : \text{(time on study)}, \\ \Delta &= I(T \leq C) : \text{(failure indicator)}, \\ \text{and } X &= q\text{-dimensional covariate vector.} \end{aligned}$$

In addition, we assume that T and C are conditionally independent given X. This is denoted as $T \perp\!\!\!\perp C|X$. This assumption is necessary to allow for the identifiability of the conditional distribution of T given X when we have censored data; see Tsiatis (1998) for more details. Otherwise, no other assumptions are made on the conditional distribution of C given X.

The data from a survival study are represented as $(Z_1, \ldots, Z_n)$, iid, where

$$Z_i = (V_i, \Delta_i, X_i), \ i = 1, \ldots, n.$$

The goal is to find semiparametric consistent, asymptotically normal, and efficient estimators for β using the sample $(V_i, \Delta_i, X_i), \ i = 1, \ldots, n$ without making any additional assumptions on the underlying baseline hazard function $\lambda(u)$, on the conditional distribution of C given X, or on the distribution of X. To do so, we will use the semiparametric theory that we have developed. Specifically, we will find the space of influence functions of RAL estimators for β for this model, which, in turn, will motivate a class of RAL estimators among which we will derive the efficient estimator. This will be accomplished by deriving the semiparametric nuisance tangent space, its orthogonal complement, where influence functions lie, and the efficient score. We begin by first considering the density of the observable data.

The density of a single data item is given by

$$\begin{aligned} p_{V,\Delta,X}(v, \delta, x) = &\left\{\lambda(v)\exp(\beta^T x)\right\}^{\delta} \exp\left\{-\Lambda(v)\exp(\beta^T x)\right\} \\ &\left\{p_{C|X}(v|x)\right\}^{1-\delta} \left\{\int_v^{\infty} p_{C|X}(u|x)du\right\}^{\delta} p_X(x), \end{aligned} \tag{5.32}$$

where the cumulative baseline hazard function is defined as

$$\Lambda(v) = \int_0^v \lambda(u)du,$$

and $p_X(x)$ denotes the marginal density of X.

It will be convenient to use a hazard representation for the conditional distribution of C given X. Define

$$\lambda_{C|X}(u|x) = \lim_{h\to 0}\left\{\frac{P(u \le C < u+h | C \ge u, X = x)}{h}\right\}$$

and

$$\Lambda_{C|X}(v|x) = \int_0^v \lambda_{C|X}(u|x)du.$$

Using the fact that

$$p_{C|X}(v|x) = \lambda_{C|X}(v|x)\exp\{-\Lambda_{C|X}(v|x)\}$$

and

$$\int_v^\infty p_{C|X}(u|x)du = \exp\{-\Lambda_{C|X}(v|x)\},$$

we write the density (5.32) as

$$\begin{aligned} p_{V,\Delta,X}(v,\delta,x) = &\left\{\lambda(v)\exp(\beta^T x)\right\}^\delta \exp\left\{-\Lambda(v)\exp(\beta^T x)\right\} \\ &\times \left\{\lambda_{C|X}(v|x)\right\}^{1-\delta}\exp\left\{-\Lambda_{C|X}(v|x)\right\} \times p_X(x). \end{aligned} \tag{5.33}$$

The model is characterized by the parameter (β, η), where β denotes the q-dimensional regression parameters of interest and the nuisance parameter

$$\eta = \{\lambda(v), \lambda_{C|X}(v|x), p_X(x)\},$$

where $\lambda(v)$ is an arbitrary positive function of v, $\lambda_{C|X}(v|x)$ is an arbitrary positive function of v and x, and $p_X(x)$ is any density such that

$$\int p_X(x)d\nu_X(x) = 1.$$

The log of the density in (5.33) is

$$\begin{aligned} &\delta\left\{\log\lambda(v) + \beta^T x\right\} - \Lambda(v)\exp(\beta^T x) \\ &\quad + (1-\delta)\log\lambda_{C|X}(v|x) - \Lambda_{C|X}(v|x) + \log p_X(x). \end{aligned} \tag{5.34}$$

The Nuisance Tangent Space

For this problem, the Hilbert space $\mathcal{H}$ consists of all q-dimensional measurable functions $h(v, \delta, x)$ of (V, Δ, X) with mean zero and finite variance equipped with the covariance inner product. In order to define the nuisance tangent space within the Hilbert space $\mathcal{H}$, we must derive the mean-square closure of all parametric submodel nuisance tangent spaces. Toward that end, we define an arbitrary parametric submodel by substituting $\lambda(v, \gamma_1)$, $\lambda_{C|X}(v|x, \gamma_2)$ and $p_X(x, \gamma_3)$ into (5.33), where γ_1, γ_2, and γ_3 are finite-dimensional nuisance parameters of dimension r_1, r_2, and r_3, respectively. Since the nuisance parameters γ_1, γ_2, and γ_3 are variationally independent and separate from each other in the log-likelihood of (5.34), this implies that the parametric submodel nuisance tangent space, and hence the nuisance tangent space, can be written as a direct sum of three orthogonal spaces,

$$\Lambda = \Lambda_{1s} \oplus \Lambda_{2s} \oplus \Lambda_{3s},$$

where

$$\Lambda_{js} = \text{mean-square closure of } \{B^{q\times r_j} S_{\gamma j}(V, \Delta, X) \text{ for all } B^{q\times r_j}\},\ j = 1, 2, 3,$$

and $S_{\gamma_j}(v, \delta, x) = \partial \log p_{V,\Delta,X}(v, \delta, x, \gamma)/\partial\gamma_j$ are the parametric submodel score vectors, for $j = 1, 2, 3$.

Since the space Λ_{3s} is associated with arbitrary marginal densities $p_X(x)$ of X, we can use arguments already developed in Chapter 4 to show that

$$\Lambda_{3s} = \left[\alpha^{q\times 1}(X) : E\left\{\alpha^{q\times 1}(X)\right\} = 0^{q\times 1}\right] \tag{5.35}$$

and that the projection of an arbitrary element $h(V, \Delta, X) \in \mathcal{H}$ is given by

$$\Pi(h|\Lambda_{3s}) = E(h|X).$$

We will now show in a series of lemmas how to derive Λ_{1s} and Λ_{2s} followed by the key theorem that derives the space orthogonal to the nuisance tangent space; i.e., $\Lambda^{\perp}$.

The Space Λ_{2s} Associated with $\lambda_{C|X}(v|x)$

Using counting process notation, let $N_C(u) = I(V \le u, \Delta = 0)$ denote (count) the indicator of whether a single individual is observed to be censored by or at time u and $Y(u) = I(V \ge u)$ denote the indicator of being at risk at time u. The corresponding martingale increment is

$$dM_C(u, x) = dN_C(u) - \lambda_{0C|X}(u|x)Y(u)du,$$

where $\lambda_{0C|X}(u|x)$ denotes the true conditional hazard function of C at time u given $X = x$.

Lemma 5.1. The space Λ_{2s} associated with the arbitrary conditional density of C given X is given as the class of elements

$$\left\{ \int \alpha^{q\times 1}(u,X)dM_C(u,X) \text{ for all functions } \alpha^{q\times 1}(u,x) \right\},$$

where $\alpha^{q\times 1}(u,x)$ is any arbitrary q-dimensional function of u and x.

Proof. In order to derive the space Λ_{2s}, consider the parametric submodel

$$\lambda_{C|X}(v|x,\gamma_2) = \lambda_{0C|X}(v|x)\exp\{\gamma_2^T \alpha^{q\times 1}(v,x)\}.$$

That this is a valid parametric submodel follows because the truth is contained within this model (i.e., when $\gamma_2 = 0$) and a hazard function must be some positive function of v.

The contribution to the log-likelihood (5.34) for this parametric submodel is

$$(1-\delta)\left\{\log \lambda_{0C|X}(v|x) + \gamma_2^T \alpha(v,x)\right\} - \int_0^v \lambda_{0C|X}(u|x)\exp\left\{\gamma_2^T \alpha(u,x)\right\} du.$$

Taking the derivative with respect to γ_2 and evaluating it at the truth ($\gamma_2 = 0$), we obtain the nuisance score vector

$$\begin{aligned} S_{\gamma_2} &= (1-\Delta)\alpha(V,X) - \int_0^V \alpha(u,X)\lambda_{0C|X}(u|X)du \\ &= (1-\Delta)\alpha(V,X) - \int \alpha(u,X)\lambda_{0C|X}(u|X)I(V \geq u)du. \end{aligned}$$

Using counting process notation, S_{γ_2} can be written as a stochastic integral,

$$S_{\gamma_2} = \int \alpha(u,X)dM_C(u,X).$$

From this last result, we conjecture that the space Λ_{2s} consists of all elements in the class

$$\left\{ \int \alpha^{q\times 1}(u,X)dM_C(u,X) \text{ for all functions } \alpha^{q\times 1}(u,x) \right\}.$$

We have already demonstrated that any element in the class above is an element of a parametric submodel nuisance tangent space. Therefore, to complete our argument and verify our conjecture, we need to show that the linear space spanned by the score vector with respect to γ_2 for any parametric submodel belongs to the space above.

Consider any arbitrary parametric submodel $\lambda_{C|X}(u|x,\gamma_2)$, with γ_{20} denoting the truth. The score vector is given by

$$\frac{\partial}{\partial\gamma_2}\left\{(1-\Delta)\log\lambda_{C|X}(V|x,\gamma_2) - \int \lambda_{C|X}(u|X,\gamma_2)I(V\geq u)du\right\}\Big|_{\gamma_2=\gamma_{20}}$$
$$= (1-\Delta)\frac{\frac{\partial}{\partial\gamma_2}\lambda_{C|X}(V|X,\gamma_{20})}{\lambda_{C|X}(V|X,\gamma_{20})}$$
$$- \int \frac{\frac{\partial}{\partial\gamma_2}\lambda_{C|X}(u|X,\gamma_{20})}{\lambda_{C|X}(u|X,\gamma_{20})}\lambda_{C|X}(u|X,\gamma_{20})I(V\geq u)du$$
$$= \int\left\{\frac{\partial\log\lambda_{C|X}(u|X,\gamma_{20})}{\partial\gamma_2}\right\}dM_C(u,X).$$

Multiplying this score vector by a conformable matrix results in an element of the form

$$\int \alpha(u,X)dM_C(u,X). \qquad \square$$

The Space Λ_{1s} Associated with $\lambda(v)$

Let $N(u)$ denote the counting process that counts whether an individual was observed to die before or at time u (i.e., $N(u) = I(V \leq u, \Delta = 1)$) and, as before, $Y(u) = I(V \geq u)$ is the "at risk" indicator at time u. Let $dM(u,X)$ denote the martingale increment $dN(u) - \lambda_0(u)\exp(\beta_0^T X)Y(u)du$, where $\lambda_0(u)$ denotes the true underlying hazard rate in the proportional hazards model.

Lemma 5.2. The space Λ_{1s}, the part of the nuisance tangent space associated with the nuisance parameter $\lambda(\cdot)$, the underlying baseline hazard rate of failure, is

$$\Lambda_{1s} = \left\{\int a^{q\times 1}(u)dM(u,X) \text{ for all } q\text{-dimensional functions } \ a^{q\times 1}(u)\right\},$$

where $a^{q\times 1}(u)$ denotes an arbitrary q-dimensional function of u.

Proof. Consider the parametric submodel

$$\lambda(v,\gamma_1) = \lambda_0(v)\exp\{\gamma_1^T a^{q\times 1}(v)\}$$

for any arbitrary q-dimensional function $a^{q\times 1}(v)$ of v. For this parametric submodel, the contribution to the log-density is

$$\delta\{\log\lambda_0(v) + \gamma_1^T a(v) + \beta^T x\} - \int_0^v \lambda_0(u)\exp\{\gamma_1^T a(u) + \beta^T x\}du. \qquad (5.36)$$

Taking derivatives of (5.36) with respect to γ_1, setting $\gamma_1 = 0$ and $\beta = \beta_0$, we obtain the score function

$$S_{\gamma_1} = \int a^{q\times 1}(u)dM(u,X).$$

From this, we conjecture that Λ_{1s} is

$$\Lambda_{1s} = \left\{ \int a^{q\times 1}(u) dM(u, X) \text{ for all } q\text{-dimensional functions } \ a^{q\times 1}(u) \right\}.$$

We have already demonstrated that any element in Λ_{1s} is an element of a parametric submodel nuisance tangent space. Therefore, to complete our argument and verify our conjecture, we need to show that the linear space spanned by the score vector with respect to γ_1 for any parametric submodel belongs to Λ_{1s}. Consider the parametric submodel with the nuisance parameter γ_1, which appears in the log-density (5.34) as

$$\delta \left\{ \log \lambda(v, \gamma_1) + \beta^T x \right\} - \Lambda(v, \gamma_1) \exp(\beta^T x), \tag{5.37}$$

where $\Lambda(v, \gamma_1) = \int_0^v \lambda(u, \gamma_1) du$ and γ_{10} denotes the truth; i.e., $\lambda(v, \gamma_{10}) = \lambda_0(v)$. Taking the derivative of (5.37) with respect to γ_1, setting $\gamma_1 = \gamma_{10}$ and $\beta = \beta_0$, we deduce that the score vector with respect to γ_1 is

$$S_{\gamma_1} = \int \left\{ \frac{\partial \log \lambda(u, \gamma_{10})}{\partial \gamma_1} \right\} dM(u, X).$$

Multiplying this score vector by a conformable matrix leads to an element in Λ_{1s}, thus verifying our conjecture. □

Finding the Orthogonal Complement of the Nuisance Tangent Space

We have demonstrated that the nuisance tangent space can be written as a direct sum of three orthogonal linear spaces, namely

$$\Lambda = \Lambda_{1s} \oplus \Lambda_{2s} \oplus \Lambda_{3s},$$

where Λ_{2s} and Λ_{1s} were derived in Lemmas 5.1 and 5.2, respectively, and Λ_{3s} is given in (5.35). Influence functions of RAL estimators for β belong to the space orthogonal to Λ. We now consider finding the orthogonal complement to Λ.

Theorem 5.5. The space orthogonal to the nuisance tangent space is given by

$$\Lambda^{\perp} = \left[\int \{\alpha(u, X) - a^*(u)\} dM(u, X) \text{ for all } \alpha^{q\times 1}(u, x) \right], \tag{5.38}$$

where

$$a^*(u) = \frac{E\left\{\alpha(u, X) \exp(\beta_0^T X) Y(u)\right\}}{E\left\{\exp(\beta_0^T X) Y(u)\right\}}. \tag{5.39}$$

Proof. We begin by noting some interesting geometric properties for the nuisance tangent space that we can take advantage of in deriving its orthogonal complement.

If we put no restrictions on the densities $p_{V,\Delta,X}(v,\delta,x)$ that generate our data (completely nonparametric), then it follows from Theorem 4.4 that the corresponding tangent space would be the space of all q-dimensional measurable functions of (V,Δ,X) with mean zero; i.e., the entire *Hilbert space* $\mathcal{H}$.

The proportional hazards model we are considering puts no restrictions on the marginal distribution of $X, p_X(x)$ or the conditional distribution of C given X. Therefore, the only restrictions on the class of densities for (V,Δ,X) come from those imposed on the conditional distribution of T given X via the proportional hazards model. Suppose, for the time being, we put no restriction on the conditional hazard of T given X and denote this by $\lambda_{T|X}(v|x)$. If this were the case, then there would be no restrictions on the distribution of (V,Δ,X).

The distribution of (V,Δ,X) given by the density $P_{V,\Delta,X}(v,\delta,x)$ can be written as $P_{V,\Delta|X}(v,\delta|x)p_X(x)$, where the conditional density $P_{V,\Delta|X}(v,\delta|x)$ can also be characterized through the cause-specific hazard functions

$$\lambda^*_\Delta(v|x) = \lim_{h\to 0}\left\{\frac{P(v\le V<v+h,\Delta=\delta|V\ge v,X=x)}{h}\right\},$$

for $\Delta = 0,1$. Under the assumption $T \perp\!\!\!\perp C|X$, the cause-specific hazard functions equal the net-specific hazard functions; namely,

$$\lambda^*_1(v|x) = \lambda_{T|X}(v|x), \lambda^*_0(v|x) = \lambda_{C|X}(v|x). \tag{5.40}$$

That (5.40) is true follows from results in Tsiatis (1998). Therefore, putting no restrictions on $\lambda_{T|X}(v|x)$ or $\lambda_{C|X}(v|x)$ implies that no restrictions are placed on the conditional distribution of (V,Δ) given X. Hence, the log-density for such a saturated (nonparametric) model could be written (analogously to (5.34)) as

$$\begin{aligned}&\delta\log\lambda_{T|X}(v|x) - \Lambda_{T|X}(v|x)\\ &\quad+(1-\delta)\log\lambda_{C|X}(v|x) - \Lambda_{C|X}(v|x)\\ &\qquad+\log p_X(x).\end{aligned}$$

The tangent space for this model can be written as a direct sum of three orthogonal spaces,

$$\Lambda^*_{1s}\oplus\Lambda_{2s}\oplus\Lambda_{3s},$$

where Λ_{2s} and Λ_{3s} are defined as before, but now Λ^*_{1s} is the space associated with $\lambda_{T|X}(v|x)$, which is now left arbitrary.

Arguments that are virtually identical to those used to find the space Λ_{2s} in Lemma 5.1 can be used to show that

$$\Lambda_{1s}^* = \left\{ \int \alpha^{q\times 1}(u,X) dM(u,X) \text{ for all } \alpha^{q\times 1}(u,x) \right\},$$

where

$$dM(u,X) = dN(u) - \lambda_{0T|X}(u|X)Y(u)du.$$

Note 2. Notice the difference: For Λ_{1s} we used $a^{q\times 1}(u)$ (i.e., a function of u only), whereas for Λ_{1s}^* we used $\alpha^{q\times 1}(u,X)$ (i.e., a function of both u and X) in the stochastic integral above. □

Because the tangent space $\Lambda_{1s}^* \oplus \Lambda_{2s} \oplus \Lambda_{3s}$ is that for a nonparametric model (i.e., a model that allows for all densities of (V,Δ,X)), and because the tangent space for a nonparametric model is the entire Hilbert space, this implies that

$$\mathcal{H} = \Lambda_{1s}^* \oplus \Lambda_{2s} \oplus \Lambda_{3s}, \tag{5.41}$$

where $\Lambda_{1s}^*, \Lambda_{2s}$, and Λ_{3s} are mutually orthogonal subspaces.

Since the nuisance tangent space $\Lambda = \Lambda_{1s} \oplus \Lambda_{2s} \oplus \Lambda_{3s}$, this implies that $\Lambda_{1s} \subset \Lambda_{1s}^*$. Also, the orthogonal complement $\Lambda^\perp$ must be orthogonal to $\Lambda_{2s} \oplus \Lambda_{3s} = \Lambda_{1s}^*$; i.e., $\Lambda^\perp \subset \Lambda_{1s}^*$. $\Lambda^\perp$ must also be orthogonal to Λ_{1s}; consequently, $\Lambda^\perp$ consists of elements of Λ_{1s}^* that are orthogonal to Λ_{1s}.

In order to identify elements of $\Lambda^\perp$ (i.e., elements of Λ_{1s}^* that are orthogonal to Λ_{1s}), it suffices to take an arbitrary element of Λ_{1s}^*, namely

$$\int \alpha^{q\times 1}(u,X) dM(u,X)$$

and find its residual after projecting it onto Λ_{1s}. To find the projection, we must derive $a^*(u)$ so that

$$\left\{ \int \alpha(u,X) dM(u,X) - \int a^*(u) dM(u,X) \right\}$$

is orthogonal to every element of Λ_{1s}. That is,

$$E\left[\int \{\alpha(u,X) - a^*(u)\}^T dM(u,X) \int a(u) dM(u,X) \right] = 0$$

for all $a(u)$.

The covariance of martingale stochastic integrals such as those above can be computed by finding the expectation of the predictable covariation process; see Fleming and Harrington (1991). Namely,

$$E\left[\int \{\alpha(u,X) - a^*(u)\}^T a(u) \lambda_0(u) \exp(\beta_0^T X) Y(u) du \right] \tag{5.42}$$

$$= \int E\left[\{\alpha(u,X) - a^*(u)\}^T \exp(\beta_0^T X) Y(u) \right] \lambda_0(u) a(u) du = 0 \tag{5.43}$$

for all $a(u)$. Since $a(u)$ is arbitrary, this implies that

$$E\left[\{\alpha(u,X)-a^*(u)\}^T\exp(\beta_0^TX)Y(u)\right]=0^{q\times 1} \tag{5.44}$$

for all u. We can prove (5.44) by contradiction because if (5.44) was not equal to zero, then we could make the integral (5.43) nonzero by choosing $a(u)$ to be equal to whatever the expectation is in (5.44).

Solving (5.44), we obtain

$$E\left\{\alpha(u,X)\exp(\beta_0^TX)Y(u)\right\}=a^*(u)E\left\{\exp(\beta_0^TX)Y(u)\right\}$$

or

$$a^*(u)=\frac{E\left\{\alpha(u,X)\exp(\beta_0^TX)Y(u)\right\}}{E\left\{\exp(\beta_0^TX)Y(u)\right\}}.$$

Therefore, the space orthogonal to the nuisance tangent space is given by

$$\Lambda^\perp=\left[\int\{\alpha(u,X)-a^*(u)\}\,dM(u,X)\text{ for all }\alpha^{q\times 1}(u,x)\right],$$

where

$$a^*(u)=\frac{E\left\{\alpha(u,X)\exp(\beta_0^TX)Y(u)\right\}}{E\left\{\exp(\beta_0^TX)Y(u)\right\}}.\ \square$$

Finding RAL Estimators for β

As we have argued repeatedly, influence functions of RAL estimators for β are defined, up to a proportionality constant, through the elements orthogonal to the nuisance tangent space. For the proportional hazards model, these elements are defined through (5.38) and (5.39). Often, knowing the functional form of such elements leads us to estimating equations that will result in RAL estimators for β with a particular influence function. We now illustrate.

Since $a^*(u)$, given by (5.39), is defined through a ratio of expectations, it is natural to estimate this using a ratio of sample averages, namely

$$\hat{a}^*(u)=\frac{n^{-1}\sum_{i=1}^n\alpha(u,X_i)\exp(\beta_0^TX_i)Y_i(u)}{n^{-1}\sum_{i=1}^n\exp(\beta_0^TX_i)Y_i(u)}.$$

Note that

$$\begin{aligned}&n^{-1/2}\sum_{i=1}^n\int\{\alpha(u,X_i)-a^*(u)\}\,dM_i(u,X_i)\\&\quad=n^{-1/2}\sum_{i=1}^n\int\{\alpha(u,X_i)-\hat{a}^*(u)\}\,dM_i(u,X_i)+o_p(1).\end{aligned} \tag{5.45}$$

This follows because

$$n^{-1/2}\sum_{i=1}^{n}\int\{\hat{a}^*(u)-a^*(u)\}\,dM_i(u,X_i)\xrightarrow{P}0,$$

a consequence of Lenglart's inequality; see Fleming and Harrington (1991).

Using straightforward algebra, we obtain that

$$\sum_{i=1}^{n}\int\left\{\alpha(u,X_i)-\frac{\sum_{j=1}^{n}\alpha(u,X_j)\exp(\beta_0^TX_j)Y_j(u)}{\sum_{j=1}^{n}\exp(\beta_0^TX_j)Y_j(u)}\right\}\underbrace{dM_i(u,X_i)}_{||}$$
$$dN_i(u)-\lambda_0(u)\exp(\beta_0^TX_i)Y_i(u)$$

is identical to

$$\sum_{i=1}^{n}\int\left\{\alpha(u,X_i)-\frac{\sum_{j=1}^{n}\alpha(u,X_j)\exp(\beta_0^TX_j)Y_j(u)}{\sum_{j=1}^{n}\exp(\beta_0^TX_j)Y_j(u).}\right\}dN_i(u).$$

Using standard expansions of the estimating equation (which we leave to the reader as as exercise), where we expand the estimating equation about the truth, we can show that the estimator for β, which is the solution to the estimating equation

$$\sum_{i=1}^{n}\int\left\{\alpha(u,X_i)-\frac{\sum\alpha(u,X_j)\exp(\beta^TX_j)Y_j(u)}{\sum\exp(\beta^TX_j)Y_j(u)}\right\}dN_i(u)=0,$$

will have an influence function "proportional" to the element of $\Lambda^{\perp}$,

$$\int\{\alpha(u,X_i)-a^*(u)\}\,dM_i(u,X_i).$$

That is, we can find a semiparametric estimator for β by choosing any q-dimensional function $\alpha(u,x)$ of u and x and solving the estimating equation

$$\sum_{i=1}^{n}\Delta_i\left\{\alpha(V_i,X_i)-\frac{\sum_{j=1}^{n}\alpha(V_i,X_j)\exp(\beta^TX_j)Y_j(V_i)}{\sum_{j=1}^{n}\exp(\beta^TX_j)Y_j(V_i)}\right\}=0. \tag{5.46}$$

By considering all functions $\alpha^{q\times1}(u,x)$, the corresponding estimators will define a class of semiparametric estimators that contains all the influence functions of RAL estimators for β.

Efficient Estimator

To find the efficient estimator, we must derive the efficient score. This entails computing

$$S_\beta(V_i, \Delta_i, X_i)$$

and projecting this onto the nuisance tangent space. Going back to the log-density (5.34) and taking the derivative with respect to β, it is straightforward to show that

$$S_\beta = \int X^{q\times 1} dM(u, X).$$

We note that S_β is an element of Λ_{1s}^*, with $\alpha(u, X_i) = X_i$.

Therefore, the efficient score, derived as the residual after projecting S_β onto Λ (or in this case Λ_{1s}), is given as

$$S_{\text{eff}} = \int \left\{ X - \frac{E\{X \exp(\beta_0^T X) Y(u)\}}{E\{\exp(\beta_0^T X) Y(u)\}} \right\} dM(u, X).$$

The estimator for β, which has an efficient influence function (i.e., proportional to S_{eff}), is given by substituting X_i for $\alpha(u, X_i)$ in (5.46); namely,

$$\sum_{i=1}^{n} \Delta_i \left[X_i - \left\{ \frac{\sum_{j=1}^{n} X_j \exp(\beta X_j) Y_j(V_i)}{\sum_{j=1}^{n} \exp(\beta X_j) Y_j(V_i)} \right\} \right] = 0. \tag{5.47}$$

The estimator for β, given as the solution to (5.47), is the estimator proposed by Cox for maximizing the partial likelihood, where the notion of partial likelihood was first introduced by Cox (1975). The martingale arguments above are essentially those used by Andersen and Gill (1982), where the theoretical properties of the proportional hazards model are derived in detail. The argument above shows that Cox's maximum partial likelihood estimator is a *globally efficient* semiparametric estimator for β for the proportional hazards model.

5.3 Estimating the Mean in a Nonparametric Model

Up to this point, we have identified influence functions and estimators by considering elements in the Hilbert space $\mathcal{H}$ that are orthogonal to the nuisance tangent space Λ. In Theorems 3.4 and 4.3, we also defined the space of influence functions of RAL estimators for β for parametric and semiparametric models, respectively, by considering the linear variety $\varphi(Z) + \mathscr{T}^\perp$, where $\varphi(Z)$ is any influence function of an RAL estimator for β and $\mathscr{T}$ is the tangent space. For some problems, this representation of influence functions may be more useful in identifying semiparametric estimators, including the efficient

estimator; i.e., when the parameter of interest is represented as $\beta(\theta)$, some function of the parameter θ, where the infinite-dimensional parameter θ describes the entire parameter space, when the tangent space $\mathscr{T}$ is easier to derive, and when some simple semiparametric but (possibly) inefficient RAL estimator, with influence function $\varphi(Z)$, exists. We will illustrate this with two examples.

In the first example, we consider the problem of finding estimators for the mean of a random variable Z (i.e., $\mu = E(Z)$) with a sample of iid random variables $Z_1, \ldots, Z_n$ with common density $p(z)$ with respect to some dominating measure ν_Z, where $p(z)$ could be any density (i.e., $p(z)$ can be any positive function of z such that $\int p(z)d\nu(z) = 1$ and $\mu = \int zp(z)d\nu(z) < \infty$); that is, the so-called nonparametric model for the distribution of Z.

Goal. We want to find the class of RAL semi(non)parametric estimators for $\mu = E(Z)$.

Basically, the model above puts no restriction on the density of Z other than finite moments. This is referred to as the nonparametric model. In Theorem 4.4, we proved that the tangent space $\mathscr{T}$ was the entire Hilbert space $\mathcal{H}$. Consequently, $\mathscr{T}^{\perp}$, the orthogonal complement of $\mathscr{T}$, is the single "zero" element of the Hilbert space corresponding to the origin. For this model, the space of influence functions, $\varphi(Z) + \mathscr{T}^{\perp}$, consists of, at most, one influence function, and if an influence function exists for an RAL estimator for μ, then this influence function must be the efficient influence function and the corresponding estimator has to be semiparametric efficient.

An obvious and natural RAL estimator for $\mu = E(Z)$ is the sample mean $\bar{Z} = n^{-1} \sum_{i=1}^{n} Z_i$. This estimator can be trivially shown to be an RAL estimator for μ with influence function $(Z - \mu_0)$. Therefore, $Z - \mu_0$ is the unique and hence efficient influence function among RAL estimators for μ for the nonparametric model and $\bar{Z}$ is the semiparametric efficient estimator.

We now consider the more complicated problem where the primary objective is to estimate the treatment effect in a randomized study with covariate adjustment.

5.4 Estimating Treatment Difference in a Randomized Pretest-Posttest Study or with Covariate Adjustment

The randomized study is commonly used to compare the effects of two treatments on response in clinical trials and other experimental settings. In this design, patients are randomized to one of two treatments with probability δ or $1 - \delta$, where the randomization probability δ is chosen by the investigator. Together with the response variable, other baseline covariate information is

collected on all the individuals in the study. The goal is to estimate the difference in the mean response between the two treatments. One simple estimator for this treatment difference is the difference in the treatment-specific sample average response between the two treatments. A problem that has received considerable interest is whether and how we can use the baseline covariates that are collected prior to randomization to increase the efficiency of the estimator for treatment difference.

Along the same line, we also consider the randomized pretest-posttest design. In this design, a random sample of subjects (patients) are chosen from some population of interest and, for each patient, a pretest measurement, say Y_1, is made, and then the patient is randomized to one of two treatments, which we denote by the treatment indicator $A = (1, 0)$ with probability δ and $(1 - \delta)$, and after some prespecified time period, a posttest measurement Y_2 is made. The goal of such a study is to estimate the effect of treatment intervention on the posttest measurement, or equivalently to estimate the effect of treatment on the change score, which is the difference between the pretest response and posttest response.

We will focus the discussion on the pretest-posttest design. However, we will see later that the results derived for the pretest-posttest study can be generalized to the problem of covariate adjustment in a randomized study.

As an example of a pretest-posttest study, suppose we wanted to compare the effect of some treatment intervention on the quality of life for patients with some disease. We may be interested in comparing the effect that a new treatment has to a placebo or comparing a new treatment to the current best standard treatment. Specifically, such a design is carried out by identifying a group of patients with disease that are eligible for either of the two treatments. These patients are then given a questionnaire to assess their baseline quality of life. Typically, in these studies, patients answer several questions, each of which is assigned a score (predetermined by the investigator), and the quality of life score is a sum of these scores, denoted by Y_1. Afterward, they are randomized to one of the two treatments $A = (0, 1)$ and then followed for some period of time and asked to complete the quality of life questionnaire again, where their quality of life score Y_2 is computed.

The goal in such studies is to estimate the treatment effect, defined as

$$\beta = E(Y_2|A = 1) - E(Y_2|A = 0),$$

or equivalently

$$\beta = E(Y_2 - Y_1|A = 1) - E(Y_2 - Y_1|A = 0),$$

where $Y_2 - Y_1$ is the difference from baseline in an individual's response and is sometimes referred to as the change score.

Note 3. This last equivalence follows because of randomization which guarantees that Y_1 and A are independent; i.e., $E(Y_1|A = 1) = E(Y_1|A = 0)$. □

The data from such a randomized pretest-posttest study can be represented as

$$Z_i = (Y_{1i}, A_i, Y_{2i}),\ i = 1, \ldots, n.$$

Some estimators for β that are commonly used include the difference in the treatment-specific sample averages of posttest response, namely the two-sample comparison

$$\hat{\beta}_n = \frac{\sum_{i=1}^{n} A_i Y_{2i}}{\sum_{i=1}^{n} A_i} - \frac{\sum_{i=1}^{n} (1 - A_i) Y_{2i}}{\sum_{i=1}^{n} (1 - A_i)}, \tag{5.48}$$

or possibly the difference in the treatment-specific sample averages of posttest minus pretest response, namely the two-sample change-score comparison

$$\hat{\beta}_n = \frac{\sum A_i (Y_{2i} - Y_{1i})}{\sum A_i} - \frac{\sum (1 - A_i)(Y_{2i} - Y_{1i})}{\sum (1 - A_i)}. \tag{5.49}$$

An analysis-of-covariance model has also been proposed for such designs, where it is assumed that

$$Y_{2i} = \eta_0 + \beta A_i + \eta_1 Y_{1i} + \varepsilon_i$$

and β, η_0, and η_1 are estimated using least squares.

It is clear that the estimators $\hat{\beta}_n$ given by (5.48) and (5.49) are semiparametric estimators for β in the sense that these are consistent and asymptotically normal estimators for β without having to make any additional parametric assumptions. It also turns out that the least-squares estimator for β in the analysis-of-covariance model is semiparametric, although this is not as obvious.

A study of the relative efficiency of these estimators was made in Yang and Tsiatis (2001). In this paper, the parameter for treatment difference β in the pretest-posttest model was also cast using a simple restricted moment model for which the efficient generalized estimating equation (GEE) could be derived. This GEE estimator was compared with the other more commonly used estimators for treatment difference in the pretest-posttest design.

Rather than considering ad hoc semiparametric estimators for β, we now propose to look at this problem using the semiparametric theory that has been developed in this book. Toward that end, we will show how to derive the space of influence functions of RAL estimators for β. We will use the linear variety $\varphi(Z) + \mathcal{T}^{\perp}$ to represent the space of influence functions.

We begin by finding the influence function of some RAL estimator for β. For simplicity, we consider the influence function of $\hat{\beta}_n$ (two-sample difference) from (5.48). This can be derived as

$$n^{1/2}(\hat{\beta}_n - \beta_0) = n^{1/2} \left(\frac{\sum A_i Y_{2i}}{\sum A_i} - \mu_2^{(1)} \right) - n^{1/2} \left(\frac{\sum (1 - A_i) Y_{2i}}{\sum (1 - A_i)} - \mu_2^{(0)} \right),$$

where $\beta_0 = E(Y_2|A=1) - E(Y_2|A=0) = \mu_2^{(1)} - \mu_2^{(0)}$. After some simple algebra, we obtain

$$n^{-1/2}\left\{\frac{\sum A_i(Y_{2i}-\mu_2^{(1)})}{(\sum A_i/n)}\right\} - n^{-1/2}\left\{\frac{\sum(1-A_i)(Y_{2i}-\mu_2^{(0)})}{(\sum(1-A_i)/n)}\right\}$$
$$\Downarrow p \qquad\qquad \Downarrow p$$
$$\delta \qquad\qquad 1-\delta$$
$$= n^{-1/2}\sum\left\{\frac{A_i}{\delta}(Y_{2i}-\mu_2^{(1)}) - \frac{(1-A_i)}{(1-\delta)}(Y_{2i}-\mu_2^{(0)})\right\} + o_p(1). \tag{5.50}$$

Consequently, the influence function for the i-th observation of $\hat{\beta}_n$ equals

$$\varphi(Z_i) = \frac{A_i}{\delta}(Y_{2i}-\mu_2^{(1)}) - \frac{(1-A_i)}{(1-\delta)}(Y_{2i}-\mu_2^{(0)}). \tag{5.51}$$

Now that we have identified one influence function for an RAL estimator for β, we can identify the linear variety of the space of influence functions by deriving the tangent space $\mathscr{T}$ and its orthogonal complement $\mathscr{T}^{\perp}$.

The Tangent Space and Its Orthogonal Complement

Let us now construct the tangent space $\mathscr{T}$. The data from a single observation in a randomized pretest-posttest study are given by (Y_1, A, Y_2). The only restriction that is placed on the density of the data is that induced by the randomization itself, specifically that the pretest measurement Y_1 is independent of the treatment indicator A and that the distribution of the Bernoulli variable A is given by $P(A=1) = \delta$ and $P(A=0) = 1-\delta$, where δ is the randomization probability, which is known by design. Other than these restrictions, we will allow the density of (Y_1, A, Y_2) to be arbitrary.

To derive the tangent space and its orthogonal complement, we will take advantage of the results of partitioning the Hilbert space for a nonparametric model given by Theorem 4.5 of Chapter 4.

First note that the density of the data for a single observation can be factored as

$$p_{Y_1,A,Y_2}(y_1,a,y_2) = p_{Y_1}(y_1)p_{A|Y_1}(a|y_1)p_{Y_2|Y_1,A}(y_2|y_1,a). \tag{5.52}$$

With no restrictions on the distribution of Y_1, A, Y_2 (i.e., the nonparametric model), we can use the results from Theorem 4.5 to show that the entire Hilbert space can be written as

$$\mathcal{H} = \mathscr{T}_1 \oplus \mathscr{T}_2 \oplus \mathscr{T}_3, \tag{5.53}$$

where

$$\mathscr{T}_1 = \{\alpha_1(Y_1) : E\{\alpha_1(Y_1)\} = 0\},$$
$$\mathscr{T}_2 = \{\alpha_2(Y_1, A) : E\{\alpha_2(Y_1, A)|Y_1\} = 0\},$$
$$\mathscr{T}_3 = \{\alpha_2(Y_1, A, Y_2) : E\{\alpha_2(Y_1, A, Y_2)|Y_1, A\} = 0\},$$

and $\mathscr{T}_1, \mathscr{T}_2, \mathscr{T}_3$ are mutually orthogonal linear spaces.

As mentioned above, in the semiparametric randomized pretest-posttest design, the only restrictions placed on the class of densities for Y_1, A, Y_2 are those induced by randomization itself. Specifically, because of the independence of Y_1 and A and the known distribution of A, we obtain

$$p_{A|Y_1}(a|y_1) = P_A(a) = \delta^a(1-\delta)^{(1-a)};$$

that is, the conditional density is completely known to us and not a function of unknown parameters. Otherwise, the density of Y_1, which is $p_{Y_1}(y_1)$, and the conditional density of Y_2 given Y_1, A; i.e., $p_{Y_2|Y_1,A}(y_2|y_1, a)$, are arbitrary. Consequently, the tangent space for semiparametric models in the randomized pretest-posttest design is given by

$$\mathscr{T} = \mathscr{T}_1 \oplus \mathscr{T}_3.$$

Remark 2. The contribution to the tangent space that is associated with the nuisance parameters for $p_{A|Y_1}(a|y_1)$, which is $\mathscr{T}_2$, is left out because, by the model restriction, this conditional density is completely known to us by design. □

The orthogonality of $\mathscr{T}_1, \mathscr{T}_2, \mathscr{T}_3$ together with (5.53) implies that the space orthogonal to the tangent space is given by

$$\mathscr{T}^{\perp} = \mathscr{T}_2.$$

Using (4.22) from Theorem 4.5, we can represent elements of $\mathscr{T}_2$ as $h_{*2}(Y_1, A) - E\{h_{*2}(Y_1, A)|Y_1\}$ for any arbitrary function $h_{*2}(\cdot)$ of Y_1 and A. Because A is a binary indicator function, any function $h_{*2}(\cdot)$ of Y_1 and A can be expressed as $h_{*2}(Y_1, A) = Ah_1(Y_1) + h_2(Y_1)$, where $h_1(\cdot)$ and $h_2(\cdot)$ are arbitrary function of Y_1. Therefore

$$\begin{aligned} h_*(Y_1, A) - E\{h_*(Y_1, A)|Y_1\} &= Ah_1(Y_1) + h_2(Y_1) - \{E(A|Y_1)h_1(Y_1) + h_2(Y_1)\} \\ &= (A - \delta)h_1(Y_1). \end{aligned}$$

Consequently, we have shown that the orthogonal complement of the tangent space is all elements $\{(A-\delta)h_*(Y_1)\}$ for any arbitrary function $h_*(\cdot)$ of Y_1 and therefore the space of all influence functions, $\{\varphi(Z) + \mathscr{T}^{\perp}\}$, is

$$\left\{ \frac{A}{\delta}(Y_2 - \mu_2^{(1)}) - \frac{(1-A)}{(1-\delta)}(Y_2 - \mu_2^{(0)}) + (A-\delta)h_*(Y_1) \right\} \tag{5.54}$$

for any arbitrary function $h_*(Y_1)$.

An estimator for β with this influence function is given by

$$\frac{\sum A_i Y_{2i}}{\sum A_i} - \frac{\sum (1-A_i) Y_{2i}}{\sum (1-A_i)} + n^{-1} \sum \left(A_i - n^{-1} \sum A_i \right) h_*(Y_{1i}).$$

The estimator above will be a consistent, asymptotically normal semiparametric estimator for β. Moreover, the class of estimators above, indexed by the functions $h_*(Y_1)$, are RAL estimators with influence functions that include the entire class of influence functions. Although all the estimators given above are asymptotically normal, the asymptotic variance will vary according to the choice of $h_*(\cdot)$.

We showed in Theorem 4.3 that the efficient influence function is given by

$$\varphi(Z) - \Pi\{\varphi\ (Z) | \mathscr{T}^{\perp}\},$$

or for our problem

$$\begin{aligned} &\frac{A}{\delta}(Y_2 - \mu_2^{(1)}) - \Pi\left\{ \frac{A}{\delta}(Y_2 - \mu_2^{(1)}) | \mathscr{T}^{\perp} \right\} \\ &\quad - \left(\frac{1-A}{1-\delta} \right) (Y_2 - \mu_2^{(0)}) + \Pi\left\{ \left(\frac{1-A}{1-\delta} \right) (Y_2 - \mu_2^{(0)}) | \mathscr{T}^{\perp} \right\}. \end{aligned} \tag{5.55}$$

Since $\mathscr{T}^{\perp} = \mathscr{T}_2$, we can use (4.23) of Theorem 4.5 to show that for any function $\alpha(Y_1, A, Y_2)$,

$$\Pi\{\alpha(\cdot) | \mathscr{T}^{\perp}\} = \Pi\{\alpha(\cdot) | \mathscr{T}_2\} = E\{\alpha(\cdot) | Y_1, A\} - E\{\alpha(\cdot) | Y_1\}.$$

Therefore

$$\Pi\left\{ \frac{A}{\delta}(Y_2 - \mu_2^{(1)}) | \mathscr{T}^{\perp} \right\} = E\left\{ \frac{A}{\delta}(Y_2 - \mu_2^{(1)}) | Y_1, A \right\} - E\left\{ \frac{A}{\delta}(Y_2 - \mu_2^{(1)}) | Y_1 \right\}, \tag{5.56}$$

where

$$E\left\{ \frac{A}{\delta}(Y_2 - \mu_2^{(1)}) | Y_1, A \right\} = \frac{A}{\delta} \left\{ E(Y_2 | Y_1, A = 1) - \mu_2^{(1)} \right\}, \tag{5.57}$$

and by the law of iterated conditional expectations, $E\left\{ \frac{A}{\delta}(Y_2 - \mu_2^{(1)}) | Y_1 \right\}$ can be computed by taking the conditional expectation of (5.57) given Y_1, yielding

$$E\left\{ \frac{A}{\delta}(Y_2 - \mu_2^{(1)}) | Y_1 \right\} = E(Y_2 | Y_1, A = 1) - \mu_2^{(1)}. \tag{5.58}$$

Therefore, by (5.56)–(5.58), we obtain

$$\Pi\left(\frac{A}{\delta}(Y_2 - \mu_2^{(1)}) | \mathscr{T}^{\perp} \right) = \frac{A - \delta}{\delta} \left\{ E(Y_2 | Y_1, A = 1) - \mu_2^{(1)} \right\}. \tag{5.59}$$

Similarly, we obtain

$$\Pi\left(\frac{1-A}{1-\delta}(Y_2-\mu_2^{(0)})|\mathscr{T}^{\perp}\right)=-\frac{A-\delta}{1-\delta}\left\{E(Y_2|Y_1,A=0)-\mu_2^{(0)}\right\}. \quad (5.60)$$

Using (5.59) and (5.60) and after some algebra, we deduce that the efficient influence function (5.55) equals

$$\begin{aligned}&\left\{\frac{A}{\delta}Y_2-\frac{(A-\delta)}{\delta}E(Y_2|A=1,Y_1)\right\}\\&-\left\{\frac{(1-A)}{(1-\delta)}Y_2+\frac{(A-\delta)}{(1-\delta)}E(Y_2|A=0,Y_1)\right\}\\&-\left(\mu_2^{(1)}-\mu_2^{(0)}\right).\\&\qquad\|\\&\qquad\beta\end{aligned} \quad (5.61)$$

In order to construct an efficient RAL estimator for β (that is, an RAL estimator for β whose influence function is the efficient influence function given by (5.61)), we would need to know $E(Y_2|A=1,Y_1)$ and $E(Y_2|A=0,Y_1)$, which, of course, we don't. One strategy is to posit models for $E(Y_2|A=1,Y_1)$ and $E(Y_2|A=0,Y_1)$ in terms of a finite number of parameters ξ_1 and ξ_0, respectively. That is,

$$E(Y_2|A=j,Y_1)=\zeta_j(Y_1,\xi_j),\ j=0,1.$$

These posited models are restricted moment models for the subset of individuals in treatment groups $A=1$ and $A=0$, respectively. For such models, we can use generalized estimating equations to obtain estimators $\hat{\xi}_{1n}$ and $\hat{\xi}_{0n}$ for ξ_1 and ξ_0 using patients $\{i:A_1=1\}$ and $\{i:A_i=0\}$, respectively. Such estimators are consistent for ξ_1 and ξ_0 if the posited models were correctly specified, but, even if they were incorrectly specified, under suitable regularity conditions, $\hat{\xi}_{jn}$ would converge to ξ_j^* for $j=0,1$. With this in mind, we use the functional form of the efficient influence given by (5.61) to motivate the estimator $\hat{\beta}_n$ for β given by

$$\begin{aligned}\hat{\beta}_n=n^{-1}\sum_{i=1}^{n}\Bigg[&\left\{\frac{A_i}{\delta}Y_{2i}-\frac{(A_i-\delta)}{\delta}\zeta_1(Y_{1i},\hat{\xi}_{1n})\right\}\\&-\left\{\frac{(1-A_i)}{(1-\delta)}Y_{2i}+\frac{(A_i-\delta)}{(1-\delta)}\zeta_0(Y_{1i},\hat{\xi}_{0n})\right\}\Bigg].\end{aligned} \quad (5.62)$$

After some algebra (Exercise 3 at the end of this Chapter), we can show that the influence function of $\hat{\beta}_n$ is

$$\begin{aligned}&\frac{A}{\delta}(Y_2-\mu_2^{(1)})-\frac{(A-\delta)}{\delta}\{\zeta_1(Y_1,\xi_1^*)-\mu_2^{(1)}\}\\&-\frac{(1-A)}{(1-\delta)}(Y_2-\mu_2^{(0)})-\frac{(A-\delta)}{(1-\delta)}\{\zeta_0(Y_1,\xi_0^*)-\mu_2^{(0)}\}.\end{aligned} \quad (5.63)$$

This influence function is in the class of influence functions of RAL estimators for β given by (5.54) regardless of whether we posited the correct model for $E(Y_2|Y_1, A = 1)$ and $E(Y_2|Y_1, A = 0)$ or not. Consequently, the estimator $\hat{\beta}_n$ given by (5.62) is locally efficient in the sense that this estimator leads to a consistent asymptotically normal semiparametric RAL estimator for β even if the posited models for the treatment-specific conditional expectations of Y_2 given Y_1 are incorrect, but this estimator is semiparametric efficient if the posited model is correct.

Our experience working with such adaptive methods is that a reasonable attempt at modeling the conditional expectations using the usual model-building techniques that statisticians learn will lead to estimators with very high efficiency even if the model is not exactly correct. For more details, see Leon, Tsiatis, and Davidian (2003).

Although the development above was specific to the pretest-posttest study, this can be easily generalized to the more general problem of covariate adjustment in a two-arm randomized study. Letting Y_2 be the response variable of interest and A the treatment indicator, assigned at random to patients, the primary goal of a two-arm randomized study is to estimate the parameter $\beta = E(Y_2|A = 1) - E(Y_2|A = 0)$. Let Y_1 be a vector of baseline covariates that are also collected on all individuals prior to randomization. One of the elements of Y_1 could be the same measurement that is used to evaluate response, as was the case in the pretest-posttest study. In such a study, the data are realizations of the iid random vectors $(Y_{1i}, A_i, Y_{2i}), i = 1, \ldots, n$. As before, in a semiparametric model, the only restriction on the class of densities for (Y_1, A, Y_2) is that induced by randomization itself; namely, that A is independent of Y_1 and that the $P(A = 1) = \delta$. Consequently, the factorization of the data (Y_1, A, Y_2) is identical to that given by (5.52) and hence the efficient influence function is given by (5.61), with the only difference being that Y_1 now represents the vector of baseline covariates rather than the single pretest measurement made at baseline in the pretest-posttest study. In fact, the exact same strategy for obtaining locally efficient adaptive semiparametric estimators for β as given by (5.62) for the pretest-posttest study can be used for this more general problem.

5.5 Remarks about Auxiliary Variables

In the randomized pretest-posttest problem presented in the previous section, the parameter of interest, $\beta = E(Y_2|A = 1) - E(Y_2|A = 0)$, is obtained from the joint distribution of (Y_2, A). Nonetheless, we were able to use the random variable Y_1, either the pretest measurement in a pretest-posttest study or a vector of baseline characteristics, to obtain a more efficient estimator of the treatment difference β. One can view the variables making up Y_1 as auxiliary variables, as they are not needed to define the parameter of interest.

Therefore, a natural question, when we are considering estimating the parameter β in a model $p(z, \beta, \eta)$ for the random vector Z, is whether we can gain efficiency by collecting additional auxiliary variables W and deriving estimators for β using the data from both Z and W. We will now argue that if we put no additional restrictions on the distribution of (W, Z) other than those from the marginal model for Z, then the class of RAL estimators is the same as if we never considered W.

To see this, let us first make a distinction between the Hilbert space $\mathcal{H}^Z$ (i.e., all q-dimensional mean-zero square-integrable functions of Z) and $\mathcal{H}^{WZ}$ (i.e., all q-dimensional mean-zero square-integrable functions of W and Z). If we consider estimators for β based only on the data from Z, then we have developed a theory that shows that the class of influence functions of RAL estimators for β can be represented as

$$\varphi(Z) + \mathscr{T}^{Z^\perp}, \tag{5.64}$$

where $\varphi(Z)$ is the influence function of any RAL estimator for β and $\mathscr{T}^{Z^\perp}$ is the space orthogonal to the tangent space defined in $\mathcal{H}^Z$.

However, if, in addition, we consider auxiliary variables W, then the space of influence functions of RAL estimators for β is given by

$$\varphi(Z) + \mathscr{T}^{WZ^\perp}, \tag{5.65}$$

where $\mathscr{T}^{WZ^\perp}$ is the space orthogonal to the tangent space defined in $\mathcal{H}^{WZ}$. We now present the key result that demonstrates that, without additional assumptions on the conditional distribution of auxiliary covariates W given Z, the class of influence functions of RAL estimators for β remains the same.

Theorem 5.6. If no restrictions are put on the conditional density $p_{W|Z}(w|z)$, where the marginal density of Z is assumed to be from the semiparametric model $p_Z(z, \beta, \eta)$, then the orthogonal complement of the tangent space $\mathscr{T}^{WZ}$ for the semiparametric model for the joint distribution of (W, Z) (i.e., $\mathscr{T}^{WZ^\perp}$) is equal to the orthogonal complement of the tangent space $\mathscr{T}^Z$ for the semiparametric model of the marginal distribution for Z alone (i.e., $\mathscr{T}^{Z^\perp}$).

Proof. With no additional restrictions placed on the joint distribution of (W, Z), the model can be represent by the class of densities

$$p_{W,Z}(w, z, \beta, \eta, \eta^*) = p_{W|Z}(w|z, \eta^*)p_Z(z, \beta, \eta), \tag{5.66}$$

where $p_{W|Z}(w|z, \eta^*)$ can be any arbitrary conditional density of W given Z; i.e., any positive function $\eta^*(w, z)$ such that $\int \eta^*(w, z)d\nu_W(w) = 1$ for all z. In Theorem 4.5, we showed that the part of the tangent space associated with the infinite-dimensional parameter η^* was the set of random functions

$$\mathscr{T}_1 = \left[a^q(W, Z) : E\{a^q(W, Z)|Z\} = 0^{q\times 1}\right]$$

and, moreover, that the entire Hilbert space $\mathcal{H}^{WZ}$ can be partitioned as

$$\mathcal{H}^{WZ} = \mathscr{T}_1 \oplus \mathcal{H}^Z, \quad \mathscr{T}_1 \perp \mathcal{H}^Z.$$

Because the density in (5.66) factors into terms involving the parameter η^* and terms involving the parameters (β, η), where η^* and (β, η) are variationally independent, it is then straightforward to show that the tangent space for the model on (W, Z) is given by

$$\mathscr{T}^{WZ} = \mathscr{T}_1 \oplus \mathscr{T}^Z, \quad \mathscr{T}_1 \perp \mathscr{T}^Z.$$

It is also straightforward to show that the space orthogonal to the tangent space $\mathscr{T}^{WZ}$ is the space that is orthogonal both to $\mathscr{T}_1$ and to $\mathscr{T}^Z$; i.e., it must be the space within $\mathcal{H}^Z$ that is orthogonal to $\mathscr{T}^Z$. In other words, the space $\mathscr{T}^{WZ^\perp}$ is the same as the space $\mathscr{T}^{Z^\perp}$. □

Because $\mathscr{T}^{WZ^\perp} = \mathscr{T}^{Z^\perp}$, this means that the space of influence functions of RAL estimators for β given by (5.64), using estimators that are functions of Z alone, is identical to the space of influence functions of RAL estimators for β given by (5.65), using estimators that are functions of W and Z.

This formally validates the intuitive notion that if we are not willing to make any additional assumptions regarding the relationship of the auxiliary variables W and the variables of interest Z, then we need not consider auxiliary variables when estimating β.

The reason we gained efficiency in the pretest-posttest problem was because a relationship was induced between Y_1 and the treatment indicator A due to randomization; namely, that Y_1 was independent of A.

5.6 Exercises for Chapter 5

1. *Heteroscedastic models*
   Consider the semiparametric model for which, for a one-dimensional response variable Y, we assume

   $$Y = \mu(X, \beta) + V^{1/2}(X, \beta)\varepsilon, \ \beta \in \mathbb{R}^q,$$

   where ε is an arbitrary continuous random variable such that ε is independent of X. To avoid identifiability problems, assume that for any scalars α, α'

   $$\alpha + \mu(x, \beta) = \alpha' + \mu(x, \beta') \quad \text{for all} \quad x$$

   implies

   $$\alpha = \alpha' \quad \text{and} \quad \beta = \beta',$$

and for any scalars $\sigma, \sigma' > 0$ that

$$\sigma\{V(x,\beta)\} = \sigma'\{V(x,\beta')\} \quad \text{for all} \quad x$$

implies

$$\sigma = \sigma' \quad \text{and} \quad \beta = \beta'.$$

For this model, describe how you would derive a locally efficient estimator for β from a sample of data

$$(Y_i, X_i), i = 1, \ldots, n.$$

2. In the pretest-posttest study, one estimator for the treatment difference β that has been proposed is the analysis of covariance (ANCOVA) estimator. That is, an analysis-of-covariance model is assumed, where

$$Y_{2i} = \eta_0 + \beta A_i + \eta_1 Y_{1i} + \varepsilon_i \tag{5.67}$$

for the iid data $(Y_{1i}, A_i, Y_{2i}), i = 1, \ldots, n$, and the parameters η_0, β, η_1 are estimated using least squares.
   a) Show that the least-squares estimator $\hat{\beta}_n$ in the model above is a semiparametric estimator for $\beta = E(Y_2|A = 1) - E(Y_2|A = 0)$; that is, that $n^{1/2}(\hat{\beta}_n - \beta)$ converges to a normal distribution with mean zero whether the linear model (5.67) is correct or not.
   b) Find the influence function for $\hat{\beta}_n$ and show that it is in the class of influence functions given in (5.54).
3. In (5.62) we considered locally efficient adaptive estimators for the treatment difference β in a randomized pretest-posttest study. Suppose the estimators $\hat{\xi}_{jn}, j = 0, 1$ were root-n consistent estimators; that is, that there exist $\xi_j^*, j = 0, 1$ such that

$$n^{1/2}(\hat{\xi}_{jn} - \xi_j^*)$$

are bounded in probability for $j = 0, 1$ and that the functions $\zeta_j(x_1, \xi_j), j = 0, 1$ as functions in ξ_j were differentiable in a neighborhood of $\xi_j^*, j = 0, 1$ for all x_1. Then show (heuristically) that the influence function for the estimator $\hat{\beta}_n$ given in (5.62) is that given in (5.63).

6

Models and Methods for Missing Data

6.1 Introduction

In many practical situations, although we may set out in advance to collect data according to some "nice" plan, things may not work out quite as intended. Some examples of this follow.

Nonresponse in sample surveys
We send out questionnaires to a sample of (randomly chosen) individuals. However, some may provide only a partial answer or no answer to some questions, or, perhaps, may not return the questionnaire at all.

Dropout or noncompliance in clinical trials
A study is conducted to compare two or more treatments. In a randomized clinical trial, subjects are enrolled into the study and then randomly assigned to one of the treatments. Suppose, in such a clinical trial, the subjects are supposed to return to the clinic weekly to provide response measurement Y_{ij} (for subject i, week j). However, some subjects "drop out" of the study, failing to show up for any clinic visit after a certain point. Still others may miss clinic visits sporadically or quit taking their assigned treatment.

Surrogate measurements
For some studies, the response of interest or some important covariate may be very expensive to obtain. For example, suppose we are interested in the daily average percentage fat intake of a subject. An accurate measurement requires a detailed "food diary" over a long period, which is both expensive and time consuming. A cheaper measurement (surrogate) is to have subjects recall the food they ate in the past 24 hours. Clearly, this cheaper measurement will be correlated with the expensive one but not perfectly. To reduce costs, a study may be conducted where only a subsample of participants provide the expensive measurement (validation sample), whereas everyone provides data on the inexpensive measurement. The expensive measurement would be

missing for all individuals not in the validation sample. Unlike the previous examples, here the missingness was by design rather than by happenstance.

In almost all studies involving human subjects, some important data may be missing for some subjects for a variety of reasons, from oversight or mistakes by the study personnel to refusal or inability of the subjects to provide information.

Objective: Usually, interest focuses on making an inference about some aspect (parameter) of the distribution of the "full data" (i.e., the data that would have been observed if no data were missing).

Problem: When some of the data are missing, it may be that, depending on how and why they are missing, our ability to make an accurate inference may be compromised.

Example 1. Consider the following (somewhat contrived) problem. A study is conducted to assess the efficacy of a new drug in reducing blood pressure for patients that have hypertension using a randomized design, where half of the patients recruited with hypertension are randomized to receive the new treatment and the other half are given a placebo. The endpoint of interest is the decrease in blood pressure after six months.

Let μ_1 denote the mean decrease in blood pressure (after six months) if all patients in a population were given the new treatment and μ_0 the population mean decrease if given a placebo. The parameter of interest is the mean treatment difference,

$$\delta = \mu_1 - \mu_0.$$

If all the patients randomized into this study are followed for six months and have complete measurements, then an unbiased estimator for δ can be computed easily using the difference in sample averages of blood pressure decrease from the patients in the two treatment groups. Suppose, however, that some of the data are missing. Consider the scenario where all the data for patients randomized to the placebo were collected, but some of the patients assigned treatment dropped out and hence their six-month blood pressure reading is missing. For such a problem, what would we do?

This problem can be defined using the following notation. For individual i in our sample $i = 1, \cdots, n$, let

$$A_i = \begin{cases} 1 \\ \\ 0 \end{cases} \text{ denotes treatment assignment,}$$

and

$$Y_i = \text{reduction in blood pressure after six months.}$$

If we had "*full*" *data* (i.e., if we observed $\{A_i, Y_i\}, i = 1, \cdots, n$), then

$$\hat{\mu}_1 = \frac{\sum_{i=1}^{n} A_i Y_i}{\sum_{i=1}^{n} A_i}, \hat{\mu}_0 = \frac{\sum_{i=1}^{n} (1 - A_i) Y_i}{\sum_{i=1}^{n} (1 - A_i)},$$

would be unbiased estimators for μ_1 and μ_0, respectively.

Let R_i denote the indicator of complete data for the i-th individual; i.e.,

$$R_i = \begin{cases} 1 & \text{if the six-month blood pressure} \\ & \text{measurement was taken} \\ 0 & \text{if it was missing.} \end{cases}$$

In such a case, we would only observe (i.e., *observed data*) $(A_i, R_i, R_i Y_i), i = 1, \cdots, n$; that is, we always observe A_i and R_i but only observe Y_i if $R_i = 1$.

It is important to emphasize the distinction between *full data*, *observed data*, and *complete data*, as we will be using this terminology throughout the remainder of the book. *Full data* are the data that we would want to have collected on all the individuals in the sample $i = 1, \ldots, n$. *Observed data* are the data that are actually observed on the individuals in the study, some of which are missing. *Complete data* are the data from only the subset of patients with no missing data.

In this hypothetical scenario, data may be missing only if $A_i = 1$. Therefore, among patients randomized to the placebo, there are no missing data and therefore we can consistently (unbiasedly) estimate μ_0 using the treatment-specific sample average

$$\hat{\mu}_0 = \frac{\sum_{i=1}^{n} (1 - A_i) Y_i}{\sum_{i=1}^{n} (1 - A_i)}.$$

Let us therefore focus our attention on estimating μ_1. Since we will only be considering this one sample for the time being, we will use the notation

$$(R_i, Y_{1i}),\ i = 1, \cdots, n_1$$

to represent the full data for the n_1 individuals randomized to treatment 1, $n_1 = \sum_{i=1}^{n} A_i$, and focus on this subset of data. The observed data are

$$(R_i, R_i Y_{1i}),\ i = 1, \cdots, n_1.$$

A natural estimator for μ_1, using the observed data, is the complete-case sample average; namely,

$$\hat{\mu}_{1c} = \frac{\sum_{i=1}^{n_1} R_i Y_{1i}}{\sum_{i=1}^{n_1} R_i}.$$

We now examine whether this estimator is reasonable or not under various circumstances.

Intuitively, if we believed that missing observations occurred by chance alone, then we might expect that the complete cases will still be representative of the population from which they were sampled and consequently would still give us an unbiased estimator. If, however, there was a systematic bias, where, say, patients with a worse prognosis (i.e., lower values of Y) were more likely to drop out, then we might expect the complete-case estimator to be overly optimistic.

This can be formalized as follows. Let the full data be denoted by

$$(R_i, Y_{1i}), i = 1, \ldots, n_1,$$

assumed to be iid with $E(Y_{1i}) = \mu_1$.

Note 1. In actuality, we cannot observe Y_{1i} whenever $R_i = 0$; nonetheless, this conceptualization is useful in understanding the consequences of missingness. □

The joint density of (R_i, Y_{1i}) can be characterized by

$$p_{R,Y_1}(r_i, y_{1i}) = p_{R|Y_1}(r_i|y_{1i})\, p_{Y_1}(y_{1i}).$$

Since R_i is binary, we only need

$$P(R_i = 1|Y_{1i} = y_{1i}) = \pi(y_{1i})$$

to specify the conditional density because $p_{R|Y_1}(r_i|y_{1i}) = \{\pi(y_{1i})\}^{r_i}\{1 - \pi(y_{1i})\}^{1-r_i}$.

Again, the observed data are denoted by

$$(R_i, R_i Y_{1i}), i = 1, \ldots, n_1.$$

If the data are missing completely at random, *denoted as MCAR* (i.e., $P(R = 1|Y_1) = \pi(Y_1) = \pi$), then the probability of being observed (or missing) does not depend on Y_{1i}. That is, R and Y_1 are independent; $R \perp\!\!\!\perp Y_1$.

The complete-case estimator

$$\frac{\sum_{i=1}^{n_1} R_i Y_{1i}}{\sum_{i=1}^{n_1} R_i} = \frac{n_1^{-1}\sum_{i=1}^{n_1} R_i Y_{1i}}{n_1^{-1}\sum_{i=1}^{n_1} R_i} \xrightarrow{P} \frac{E(RY_1)}{E(R)},$$

and by the independence of R and Y_1,

$$\frac{E(RY_1)}{E(R)} = \frac{E(R)E(Y_1)}{E(R)} = E(Y_1) = \mu_1.$$

Therefore, if the data are missing completely at random (MCAR), then the complete-case estimator is unbiased, as our intuition would suggest. If, however, the probability of missingness depends on Y_1, which we will refer to as nonmissing at random, (NMAR) (a formal definition will be given in later chapters), then the complete-case estimator is written

$$\frac{\sum_{i=1}^{n_1} R_i Y_{1i}}{\sum_{i=1}^{n_1} R_i} \xrightarrow{P} \frac{E(RY_1)}{E(R)} = \frac{E\{E(RY_1)|Y_1\}}{E\{E(R|Y_1)\}} \tag{6.1}$$

$$= \frac{E\{Y_1\pi(Y_1)\}}{E\{\pi(Y_1)\}} \neq E(Y_1) \quad \text{(necessarily)}.$$

In fact, if $\pi(y)$ is an increasing function in y (i.e., probability of *not* being missing increases with y), then this suggests that individuals with larger values of Y would be overrepresented in the observed data and hence

$$\frac{E\{Y_1\pi(Y_1)\}}{E\{\pi(Y_1)\}} > \mu_1.$$

We leave the proof of this last result as an exercise for the reader.

The difficulty with NMAR is that there is no way of estimating $\pi(y) = P(R = 1|Y_1 = y)$ from the observed data because if $R = 0$ we don't get to observe Y_1. In fact, there is no way that we can distinguish whether the missing data were MCAR or NMAR from the observed data. That is, there is an inherent nonidentifiability problem here.

There is, however, a third possibility to consider. Suppose, in addition to the response variable, we measured additional covariates on the i-th individual, denoted by W_i, which are not missing. For example, some baseline characteristics may also be collected, including baseline blood pressure or possibly some additional variables collected between the initiation of treatment and six months, when the follow-up response is supposed to get collected. Such variables $W_i, i = 1, \ldots, n$ are sometimes referred to as auxiliary covariates, as they represent variables that are not of primary interest for inference. The observable data in such a case are denoted by

$$(R_i, R_i Y_{1i}, W_i), \quad i = 1, \cdots, n_1.$$

Although the auxiliary covariates are not of primary interest, suppose we believe that the reason for missingness depends on W_i, and, moreover, conditional on W_i, Y_{1i} has no additional effect on the probability of missingness. Specifically,

$$P(R_i = 1\big|Y_{1i}, W_i) = \pi(W_i). \tag{6.2}$$

That is, conditional on W_i, Y_{1i} is independent of R_i;

$$R_i \perp\!\!\!\perp Y_{1i}\big|W_i.$$

Remark 1. It may be that W_i is related to both Y_{1i} and R_i, in which case, even though (6.2) is true, a dependence between Y_{1i} and R_i would be induced. For example, consider the hypothetical scenario where W_i denotes the blood pressure for the i-th individual at an interim examination, say at three months, measured on all n individuals in the sample. After observing this response, individuals whose blood pressure was still elevated would be more likely to drop out. Therefore, R_i would be correlated with W_i. Since individuals could not possibly know what their blood pressure is at the end of the study (i.e., at six months), it may be reasonable to assume that $R_i \perp\!\!\!\perp Y_{1i}|W_i$. It may also be reasonable to assume that the three-month blood pressure reading is correlated with the six-month blood pressure reading. Under these circumstances, dropping out of the study R_i would be correlated with the response outcome Y_{1i} but in this case, because they were both related to the interim three-month blood pressure reading.

Therefore, without conditioning additionally on W_i, we would obtain that

$$P(R_i = 1\big|Y_{1i}) = \pi(Y_{1i})$$

depends on the value Y_{1i}. Consequently, if we didn't collect the additional data W_i, then we would be back in the impossible NMAR situation. □

The assumption (6.2) is an example of what is referred to as missing at random, or MAR (not to be confused with MCAR). Basically, missing at random means that the probability of missingness depends on variables that are observed. A general and more precise definition of MAR will be given later.

The MAR assumption alleviates the identifiability problems that were encountered with NMAR because the probability of missingness depends on variables that are observed on all subjects. The available data could also be used to model the relationship for the probability of missingness, or, equivalently, the probability of a complete case, as a function of the covariates W_i. For example, we can posit a model for

$$P(R = 1|W = w) = \pi(w, \gamma)$$

(say, logistic regression) in terms of a parameter vector γ and estimate the parameter γ from the observed data (R_i, W_i), which are measured on all individuals $i = 1, \ldots, n$, using, say, maximum likelihood. That is, the maximum likelihood estimator $\hat{\gamma}$ would be obtained by maximizing

$$\prod_{i=1}^{n} \pi(W_i, \gamma)^{R_i} \{1 - \pi(W_i, \gamma)\}^{1-R_i}.$$

This probability of a complete case as a function of the covariates, together with the MAR assumption, will prove useful for computing inverse probability weighted complete-case estimators, to be described later.

There are several methods for estimating parameters when data are missing at random. These methods include:

(a) likelihood methods;
(b) imputation methods;
(c) inverse probability weighting of complete cases.

Likelihood and imputation methods have been studied in great detail, and some excellent references include the books by Rubin (1987), Little and Rubin (1987), and Schafer (1997). The main theory of inverse probability weighted methods is given in the seminal paper by Robins, Rotnitzky, and Zhao (1994) and will be the primary focus of this book.

To give a flavor of how these methods work, let us return to the problem of estimating $\mu_1 = E(Y_1)$ when auxiliary variables W are collected and the response data are missing at random; i.e.,

$$R \perp\!\!\!\perp Y_1 | W.$$

6.2 Likelihood Methods

Consider a parametric model with density

$$p_{Y_1,W}(y_1, w) = p_{Y_1|W}(y_1|w, \gamma_1)\ p_W(w, \gamma_2), \tag{6.3}$$

where γ_1, γ_2 are unknown parameters describing the conditional distribution of Y_1 given W and the marginal distribution of W, respectively. The parameter μ_1 we are interested in can be written as

$$\begin{aligned}\mu_1 = E(Y_1) &= E\{E(Y_1|W)\} \\ &= \int y p_{Y_1|W}(y|w, \gamma_1) p_W(w, \gamma_2) d\nu_Y(y) d\nu_W(w).\end{aligned}$$

If we could obtain estimators for γ_1, γ_2, say $\hat{\gamma}_1, \hat{\gamma}_2$, respectively, then we could estimate μ_1, by

$$\hat{\mu}_1 = \int y p_{Y_1|W}(y|w, \hat{\gamma}_1) p_W(w, \hat{\gamma}_2) d\nu_Y(y) d\nu_W(w). \tag{6.4}$$

A popular way of obtaining estimators is by maximizing the likelihood. The density of the observed data for one individual can be written as

$$p_{RY_1,W,R}(ry_1, w, r) = \{p_{Y_1,W,R}(y_1, w, r = 1)\}^{I(r=1)}\ \{p_{W,R}(w, r = 0)\}^{I(r=0)}$$

or

$$\begin{aligned}&\left\{p_{Y_1|W,R}(y_1|w, r = 1)\ p_{R|W}(r = 1|w) p_W(w)\right\}^{I(r=1)} \\ &\left\{p_{R|W}(r = 0|w)\ p_W(w)\right\}^{I(r=0)}.\end{aligned}$$

Because of MAR (i.e., $R \perp\!\!\!\perp Y_1|W$),

$$p_{Y_1|W,R}(y_1|w, r = 1) = p_{Y_1|W}(y_1|w).$$

Therefore, the likelihood for one individual in our sample is given by

$$\begin{aligned}&\left\{p_{Y_1|W}(y_1|w, \gamma_1)\right\}^{I(r=1)} p_W(w, \gamma_2)\\ &\quad \times \{\pi(w, \gamma_3)\}^{I(r=1)}\{1 - \pi(w, \gamma_3)\}^{I(r=0)},\end{aligned}$$

where

$$p_{R|W}(r = 1|w) = \pi(w, \gamma_3).$$

Consequently, the likelihood for n independent sets of data is given as

$$\left\{\prod_{i=1}^{n} p_{Y_1|W}(y_{1i}|w_i, \gamma_1)^{I(r_i=1)}\right\}\left\{\prod_{i=1}^{n} p_W(w_i, \gamma_2)\right\} \times \{\text{function of } \gamma_3\}.$$

Because of the way the likelihood factorizes, we find the MLE for γ_1, γ_2 by separately maximizing

$$\prod_{\{i\,:\,R_i=1\}} p_{Y_1|W}(y_{1i}|w_i, \gamma_1) \tag{6.5}$$

and

$$\prod_{i=1}^{n} p_W(w_i, \gamma_2). \tag{6.6}$$

Note 2. We only include complete cases to find the MLE for γ_1, whereas we use all the data to find the MLE for γ_2. □

The estimates for γ_1 and γ_2, found by maximizing (6.5) and (6.6), can then be substituted into (6.4) to obtain the MLE for μ_1.

Remark 2. Although likelihood methods are certainly feasible and the corresponding estimators enjoy the optimality properties afforded to an MLE, they can be difficult to compute in some cases. For example, the integral given in (6.4) can be numerically challenging to compute, especially if W involves many covariates.

6.3 Imputation

Since some of the Y_{1i}'s are missing, a natural strategy is to impute or "estimate" a value for such missing data and then estimate the parameter of interest behaving as if the imputed values were the true values.

For example, if there were no missing values of $Y_{1i}, i = 1, \ldots, n_1$, then we would estimate μ_1 by using

$$\hat{\mu}_1 = n_1^{-1} \sum_{i=1}^{n_1} Y_{1i}. \tag{6.7}$$

However, for values of i such that $R_i = 0$, we don't observe such Y_{1i}.

Suppose we posit some relationship for the distribution of Y_1 given W; e.g., we may assume a parametric model $p_{Y_1|W}(y, |w, \gamma_1)$ as we did in (6.3). By the MAR assumption, we can estimate γ_1 using the complete cases, say, by maximizing (6.5) to derive the MLE $\hat{\gamma}_1$. This would allow us to estimate

$$E(Y_1 \big| W = w) \text{ by } \int y p_{Y_1|W}(y|w, \hat{\gamma}_1) d\nu_Y(y),$$

which we denote by $\mu(w, \hat{\gamma}_1)$.

We then propose to impute the missing data for any individual i such that $R_i = 0$ by substituting the value $\mu(W_i, \hat{\gamma}_1)$ for the missing Y_i in (6.7). The resulting imputed estimator is

$$n_1^{-1} \sum_{i=1}^{n_1} \{R_i Y_{1i} + (1 - R_i)\mu(W_i, \hat{\gamma}_1)\}. \tag{6.8}$$

A heuristic argument why this should yield a consistent estimator is as follows. Assuming $\hat{\gamma}_1$ converges to the truth, then (6.8) should be well approximated by

$$n_1^{-1} \sum_{i=1}^{n_1} \{R_i Y_{1i} + (1 - R_i)\mu(W_i, \gamma_{10})\} + o_p(1), \tag{6.9}$$

where $o_p(1)$ is a term converging in probability to zero and, at the truth, γ_{10},

$$\mu(W_i, \gamma_{10}) = E(Y_{1i} \big| W_i).$$

Therefore, by the weak law of large numbers (WLLN), (6.9) converges in probability to

$$E\{RY_1 + (1 - R)E(Y_1|W)\}.$$

By a conditioning argument, this equals

$$\begin{aligned} & E\left[E\{RY_1 + (1 - R)E(Y_1|W)|R, W\}\right] \\ &= E\{R \underbrace{E(Y_1|R, W)}_{\text{by MAR} = E(Y_1|W)} + (1 - R)E(Y_1|W)\} \\ &= E\{RE(Y_1|W) + (1 - R)E(Y_1|W)\} \\ &= E\{E(Y_1|W)\} = E(Y_1) = \mu_1. \qquad \square \end{aligned}$$

Remarks

1. We could have modeled the conditional expectation directly, say as

$$E(Y_1|W) = \mu(W, \gamma),$$

and estimated γ by using GEEs with complete cases.

2. Later, we will consider other imputation techniques, where we impute a missing Y_{1i} by using a random draw from the conditional distribution of $p_{Y_1|W_i}(y_1|W_i, \hat{\gamma}_1)$ or possibly using more than one draw (multiple imputation).

6.4 Inverse Probability Weighted Complete-Case Estimator

When data are MAR (i.e., $Y_1 \perp\!\!\!\perp R|W$), we have already argued that Y_1 may not be independent of R. Therefore, the naive complete-case estimator (6.1)

$$\frac{\sum_{i=1}^{n_1} R_i Y_{1i}}{\sum_{i=1}^{n_1} R_i}$$

may be biased. Horvitz and Thompson (1952), and later Robins, Rotnitzky, and Zhao (1994), suggested using inverse weighting of complete cases as a method of estimation. Let us denote the probability of observing a complete case by

$$\underbrace{P(R=1|W,Y_1) = P(R=1|W)}_{\Downarrow \text{ follows by MAR}} = \pi(W).$$

The basic intuition is as follows. For any individual randomly chosen from a population with covariate value W, the probability that such an individual will have complete data is $\pi(W)$. Therefore, any individual with covariate W with complete data can be thought of as representing $\frac{1}{\pi(W)}$ individuals at random from the population, some of which may have missing data. This suggests using

$$\hat{\mu}_1 = n_1^{-1} \sum_{i=1}^{n_1} \frac{R_i Y_{1i}}{\pi(W_i)}$$

as an estimator for μ_1. By WLLN, $\hat{\mu}_1 \xrightarrow{P} E\left\{\frac{RY_1}{\pi(W)}\right\}$, which, by conditioning, equals

$$\begin{aligned} E\left[E\left\{\frac{RY_1}{\pi(W)}\middle| Y_1, W\right\}\right] &= E\left[\frac{Y_1}{\pi(W)} E(R|Y_1, W)\right] \\ &= E\left\{\frac{Y_1}{\pi(W)}\pi(W)\right\} = E(Y_1) = \mu_1. \end{aligned}$$

In many practical situations, $\pi(W) = P(R = 1|W)$ would not be known to us. In such cases, we may posit a model $P(R = 1|W = w) = \pi(w, \gamma_3)$, and γ_3 can be estimated by maximizing the likelihood

$$\prod_{i=1}^{n_1} \{\pi(W_i, \gamma_3)\}^{R_i} \{1 - \pi(W_i\gamma_3)\}^{1-R_i}$$

to obtain the estimator $\hat{\gamma}_3$. (Often, logistic regression models are used.) The resulting inverse probability weighted complete-case (IPWCC) estimator is given by

$$n_1^{-1} \sum_{i=1}^{n_1} \frac{R_i Y_{1i}}{\pi(W_i, \hat{\gamma_3})}. \tag{6.10}$$

Remark 3. There is a technical condition that $\pi(w)$ be strictly greater than zero for all values of w in the support of W in order that the IPWCC estimator (6.10) be consistent and asymptotically normal. This will be discussed in greater detail in later chapters. A cautionary note, however, even if this technical condition holds true, is that if $\pi(W_i, \hat{\gamma_3})$ is very small, then this gives undue influence to the i-th observation in the IPWCC estimator (6.10). This could result in a very unstable estimator with poor performance with small to moderate sample sizes. □

6.5 Double Robust Estimator

For both the likelihood estimator and the imputation estimator, we had to specify a model for $p_{Y_1|W}(y_1|w, \gamma_1)$. If this model was incorrectly specified, then both of these estimators would be biased. For the IPWCC estimator, we had to specify a model for the probability of missingness $P(R = 1|W = w) = \pi(w, \gamma_3)$. If this model was incorrectly specified, then the IPWCC estimator would be biased. More recently, augmented inverse probability weighted complete-case estimators have been suggested that are doubly robust. Scharfstein, Rotnitzky, and Robins (1999) first introduced the notion of double robust estimators. Double robust estimators were also studied by Lipsitz, Ibrahim, and Zhao (1999), Robins (1999), Robins, Rotnitzky, and van der Laan (2000), Lunceford and Davidian (2004), Neugebauer and van der Laan (2005), Robins and Rotnitzky (2001), and van der Laan and Robins (2003); the last two references give the theoretical justification for these methods. An excellent overview is also given by Bang and Robins (2005). To obtain such an estimator, a model is specified for

$$E(Y_1|W) = \mu(W, \gamma_1) \tag{6.11}$$

and for

$$P(R = 1|W = w) = \pi(w, \gamma_3), \tag{6.12}$$

where the parameters γ_1 and γ_3 can be estimated as we have already discussed. Denote these estimators by $\hat{\gamma}_1$ and $\hat{\gamma}_3$.

If (6.11) is correctly specified, then $\hat{\gamma}_1$ will converge in probability to γ_{10} (the truth), in which case

$$\mu(W, \hat{\gamma}_1) \xrightarrow{P} \mu(W, \gamma_{10}) = E(Y_1|W),$$

whereas if (6.11) is incorrectly specified, then $\hat{\gamma}_1 \xrightarrow{P} \gamma_1^*$ and

$$\mu(W, \hat{\gamma}_1) \xrightarrow{P} \mu(W, \gamma_1^*) \neq E(Y_1|W).$$

Similarly, if (6.12) is correctly specified, then

$$\pi(W, \hat{\gamma}_3) \xrightarrow{P} \pi(W, \gamma_{30}) = P(R = 1|W),$$

and if incorrectly specified

$$\pi(W, \hat{\gamma}_3) \xrightarrow{P} \pi(W, \gamma_3^*) \neq P(R = 1|W).$$

Consider the estimator

$$\hat{\mu}_1 = n_1^{-1} \sum_{i=1}^{n_1} \left[\frac{R_i Y_{1i}}{\pi(W_i, \hat{\gamma}_3)} - \frac{\{R_i - \pi(W_i, \hat{\gamma}_3)\}}{\pi(W_i, \hat{\gamma}_3)} \mu(W_i, \hat{\gamma}_1) \right]. \tag{6.13}$$

This estimator is referred to as an augmented inverse probability weighted complete-case (AIPWCC) estimator. Notice that the first term in (6.13) is the inverse probability weighted complete-case estimator derived in Section 6.4. This term only involves contributions from complete cases. The second term includes additional contributions from individuals with some missing data. This is the so-called augmented term. We will now show that this AIPWCC estimator is doubly robust in the sense that it is consistent if either the model (6.11) for Y_1 given W is correctly specified (i.e., $\mu(W, \gamma_{10}) = E(Y_1|W)$) or the missingness model (6.12) is correctly specified (i.e., $\pi(W, \gamma_{30}) = P(R = 1|W)$).

The estimator (6.13) can be reexpressed as

$$n_1^{-1} \sum_{i=1}^{n_1} \left[Y_{1i} + \frac{\{R_i - \pi(W_i, \hat{\gamma}_3)\}}{\pi(W_i, \hat{\gamma}_3)} \{Y_{1i} - \mu(W_i, \hat{\gamma}_1)\} \right].$$

(a) Suppose the model (6.12) is correctly specified but the model (6.11) might not be. Then the estimator is approximated by

$$n_1^{-1} \sum_{i=1}^{n_1} \left[Y_{1i} + \frac{\{R_i - P(R = 1|W_i)\}}{P(R = 1|W_i)} \{Y_{1i} - \mu(W_i, \gamma_1^*)\} \right] + o_p(1),$$

which converges to

$$
\begin{aligned}
&E\left[Y_1 + \left\{\frac{R - P(R=1|W)}{P(R=1|W)}\right\}\{Y_1 - \mu(W, \gamma_1^*)\}\right] \\
&= E(Y_1) + \underbrace{E\left[\left\{\frac{R - P(R=1|W)}{P(R=1|W)}\right\}\{Y_1 - \mu(W, \gamma_1^*)\}\right]}_{\text{Conditioning on } Y_1, W \quad E[E[\ \ |Y_1, W]]} \\
&\qquad E\left[\underbrace{\left\{\frac{E(R|Y_1, W) - P(R=1|W)}{P(R=1|W)}\right\}}_{\left\{\frac{P(R=1|W) - P(R=1|W)}{P(R=1|W)}\right\} \quad = \quad 0}\{Y_1 - \mu(W, \gamma_1^*)\}\right] \\
&= E(Y_1) = \mu_1.
\end{aligned}
$$

(b) Suppose the model (6.11) is correctly specified but the model (6.12) might not be. Then the estimator is

$$
n_1^{-1}\sum_{i=1}^{n_1}\left[Y_{1i} + \left\{\frac{R_i - \pi(W_i, \gamma_3^*)}{\pi(W_i, \gamma_3^*)}\right\}\{Y_{1i} - E(Y_1|W_i)\}\right] + o_p(1),
$$

which converges to

$$
\begin{aligned}
&E\left[Y_1 + \left\{\frac{R - \pi(W, \gamma_3^*)}{\pi(W, \gamma_3^*)}\right\}\{Y_1 - E(Y_1|W)\}\right] \\
&= E(Y_1) + \underbrace{E\left[\left\{\frac{R - \pi(W, \gamma_3^*)}{\pi(W, \gamma_3^*)}\right\}\{Y_1 - E(Y_1|W)\}\right]}_{\text{Condition on } R, W \quad E[E[\ \ |R, W]]} \\
&= E\left[\left\{\frac{R - \pi(W, \gamma_3^*)}{\pi(W, \gamma_3^*)}\right\}\{E(Y_1|R, W) - E(Y_1|W)\}\right] \\
&\qquad\qquad \{E(Y_1|W) - E(Y_1|W)\} \quad = \quad 0 \\
&= E(Y_1) = \mu_1.
\end{aligned}
$$

Consequently, if either the model for $E(Y_1|W)$ or $P(R = 1|W)$ is correctly specified, then the double robust estimator consistently estimates μ_1.

We will demonstrate later that if both models are correctly specified, then the double robust estimator is more efficient than the IPWCC estimator.

Remark 4. Another advantage to the double robust estimator is that, based on some of our experience, this estimator obviates some of the instability problems that can result from small weights as noted in Remark 3 for the IPWCC estimator. □

6.6 Exercises for Chapter 6

1. In (6.1) we showed that the sample average of the response variable Y_1 among complete cases, when the missingness mechanism is NMAR, will converge in probability to

$$\frac{E\{Y_1\pi(Y_1)\}}{E\{\pi(Y_1)\}},$$

where $\pi(Y_1)$ denotes the probability of being included in the sample as a function of the response variable Y_1 (i.e., $P(R = 1|Y_1) = \pi(Y_1)$). If $\pi(y)$ is an increasing function in y (i.e., the probability of *not* being missing increases with y), then prove that

$$\frac{E\{Y_1\pi(Y_1)\}}{E\{\pi(Y_1)\}} > \mu_1.$$

7

Missing and Coarsening at Random for Semiparametric Models

7.1 Missing and Coarsened Data

In Chapter 6, we described three different missing-data mechanisms:

1. *MCAR* (missing completely at random): The probability of missingness is independent of the data.
2. *MAR* (missing at random): The probability of missingness depends only on the observed data.
3. *NMAR* (nonmissing at random): The probability of missingness may also depend on the unobservable part of the data.

NMAR is clearly the most problematic. Since missingness may depend on data that are unobserved, we run into nonidentifiability problems. Because of these difficulties, we do not find methods that try to model the missingness mechanism for NMAR models very attractive since the correctness of the model cannot be verified using the observed data. Another approach, which we believe is more useful, is to use NMAR models as part of a sensitivity analysis. There has been some excellent work along these lines; see, for example, Scharfstein, Rotnitzky, and Robins (1999), Robins, Rotnitzky, and Scharfstein (2000), and Rotnitzky et al. (2001).

In this book, we will not consider NMAR models; instead, we will focus our attention on models for missing data that are either MAR or MCAR. Restricting attention only to such models still allows us a great deal of flexibility. Although the primary interest is to make inference on parameters that involve the distribution of Z had Z been observed on the entire sample, the MAR assumption allows us to consider cases where the probability of missingness may also depend on other auxiliary variables W that are collected on the sample. If the reasons for missingness depend on the observed data, including the auxiliary variables, then the MAR assumption may be tenable for a wide variety of problems. We gave an example of this when we introduced the notion of MAR in Chapter 6.

Remark 1. Although we have made a distinction between auxiliary variables W and primary variables Z, for ease of notation, we will not introduce additional notation unless absolutely necessary. That is, we can define the full data Z to be all the data that are collected on individuals in our sample. This may include auxiliary as well as primary variables. For example, the full data Z may be partitioned as $(Z_1^T, Z_2^T)^T$, where Z_1 are the primary variables and Z_2 are the auxiliary variables. The model for the primary variables can be denoted by a semiparametric model $p_{Z_1}(z_1, \beta, \eta_1)$, where β is the parameter of interest and η_1 are nuisance parameters. Since the auxiliary variables Z_2 are not of primary interest, we might not want to put any additional restrictions on the conditional distribution of the auxiliary variables Z_2 given Z_1. This situation can be easily accommodated by considering the semiparametric model $p_Z(z, \beta, \eta)$, where $\eta = (\eta_1, \eta_2)$ and

$$p_Z(z, \beta, \eta) = p_{Z_1}(z_1, \beta, \eta_1) p_{Z_2|Z_1}(z_2|z_1, \eta_2),$$

where the nuisance function η_2 allows for any arbitrary conditional density of Z_2 given Z_1. By so doing, auxiliary variables can be introduced as part of a full-data semiparametric model. □

Again, we emphasize that the underlying objective is to make inference on parameters that describe important aspects of the distribution of the data Z had Z not been missing. That is, had Z been completely observed for the entire sample, then the data would be realizations of the iid random quantities $Z_1, \ldots, Z_n$, each with density $p_Z(z, \beta, \eta)$, where $\beta^{q\times 1}$ is assumed finite-dimensional, and η denotes the nuisance parameter, which for semiparametric models is infinite-dimensional. It is the parameter β in this model that is of primary interest. The fact that some of the data are missing is a difficulty that we have to deal with by thinking about and modeling the missingness process. The model for missingness, although important for conducting correct inference, is not of primary inferential interest.

Although we have only discussed missing data thus far, it is not any more difficult to consider the more general notion of "coarsening" of data. The concept of coarsened data was first introduced by Heitjan and Rubin (1991) and studied more extensively by Heitjan (1993) and Gill, van der Laan, and Robins (1996). When we think of missing data, we generally consider the case where the data on a single individual can be represented by a random vector with, say, l components, where a subset of these components may be missing for some of the individuals in the sample. When we refer to coarsened data, we consider the case where we observe a many-to-one function of Z for some of the individuals in the sample. Just as we allow that different subsets of the data Z may be missing for different individuals in the sample, we allow the possibility that different many-to-one functions may be observed for different individuals. Specifically, we will define a coarsening (missingness) variable $\mathcal{C}$ such that, when "$\mathcal{C} = r$," we only get to see a many-to-one function of the data, which we denote by $G_r(Z)$, and different r correspond to different

many-to-one functions. Therefore, $\mathcal{C}$ will be a single discrete random variable made up of positive integers $r = 1, \ldots, \ell$ and ∞, where ℓ denotes the number of different many-to-one functions considered. We reserve $\mathcal{C} = \infty$ to denote "no coarsening."

Example 1. A simple illustration of coarsened data is given by the following hypothetical example. An investigator is interested in studying the relationship between the concentration of some biological marker in some fixed volume of an individual's blood serum and some outcome. However, the investigator is also interested in determining the within-person variability in serum concentrations. Therefore, two blood samples of equal volume are drawn from each individual in a study. Denote by X_1 and X_2 the serum concentrations for these two samples and by Y the response variable. The full data for this scenario are given as (Y, X_1, X_2). To save on expense, the investigator measures the concentrations separately on the two samples from only a subset of the patients chosen at random. For the remaining patients, the two blood samples are combined and one measurement is made to obtain the concentration from the combined samples. Since the blood volumes are the same, the combined concentration would be $(X_1 + X_2)/2$. Hence, in this example, there are two levels of coarsening. When $\mathcal{C} = \infty$, we observe the full data (Y, X_1, X_2) (i.e., $G_\infty(Y, X_1, X_2) = (Y, X_1, X_2)$) whereas when $\mathcal{C} = 1$, we observe the coarsened data $\{Y, (X_1 + X_2)/2\}$ (i.e., $G_1(Y, X_1, X_2) = \{Y, (X_1 + X_2)/2\}$). □

We now illustrate how missing data is just a special case of the more general concept of coarsening.

Missing Data as a Special Case of Coarsening

Suppose the full data Z for a single individual is made up of an l-dimensional vector of random variables, say

$$Z = \left(Z^{(1)}, \ldots, Z^{(l)}\right)^T.$$

Having missing data is equivalent to the case where a subset of the elements of $(Z^{(1)}, \ldots, Z^{(l)})$ are observed and the remaining elements are missing. This can be represented using the coarsening notation $\{\mathcal{C}, G_{\mathcal{C}}(Z)\}$, where, the many-to-one function $G_r(Z)$ maps the vector Z to a subset of elements of this vector whenever $\mathcal{C} = r$.

For example, let $Z = (Z^{(1)}, Z^{(2)})^T$ be a vector of two random variables. Define

$\mathcal{C}$	$G_{\mathcal{C}}(Z)$
1	$Z^{(1)}$
2	$Z^{(2)}$
∞	$(Z^{(1)}, Z^{(2)})^T$

That is, if $\mathcal{C} = 1$, then we only observe $Z^{(1)}$, and $Z^{(2)}$ is missing; if $\mathcal{C} = 2$, then we observe $Z^{(2)}$, and $Z^{(1)}$ is missing; and if $(\mathcal{C} = \infty)$, then there are no missing data and we observe $Z = (Z^{(1)}, Z^{(2)})^T$.

Remark 2. If we were dealing only with missing data, say, where different subsets of an l-dimensional random vector may be missing, it may be more convenient to define the missingness variable to be an l-dimensional vector of 1's and 0's to denote which element of the vector is observed or missing. If it is convenient to switch to such notation, we will use R to denote such missingness indicators. This was the notation used to represent missing data, for example, in Chapter 6. □

The theory developed in this book will apply to missing and coarsened data problems where it is assumed that there is a positive probability of observing the full data. That is,

$$P(\mathcal{C} = \infty | Z = z) \geq \varepsilon > 0 \quad \text{for } z \text{ a.e.}$$

Therefore, some problems that may be thought of as missing-data problems but where no complete data are ever observed would not be part of the theory we will consider. For example, measurement error problems that do not include a validation set (i.e., when the true underlying covariate is never observed for any individual in our sample but instead only some mismeasured version of the covariate is available) cannot be covered by the theory developed in this book.

Coarsened-Data Mechanisms

In problems where the data are coarsened or missing, it is assumed that we get to observe the coarsening variable $\mathcal{C}$ and the corresponding coarsened data $G_{\mathcal{C}}(Z)$. Thus, the *observed data* are realizations of the iid random quantities

$$\{\mathcal{C}_i, G_{\mathcal{C}_i}(Z_i)\}, i = 1, \ldots, n.$$

To specify models for coarsened data, we must specify the probability distribution of the coarsening process together with the probability model for the full data (i.e., the data had there not been any coarsening). As with missingness, we can define different coarsening mechanisms that can be categorized as coarsening completely at random (CCAR), coarsening at random (CAR), and noncoarsening at random (NCAR).

These are defined as follows,

1. *Coarsening completely at random*:

$$P(\mathcal{C} = r \big| Z) = \varpi(r) \quad \text{for all} \quad r, Z; \text{ i.e., } \mathcal{C} \perp\!\!\!\perp Z.$$

2. *Coarsening at random*:

$$P(\mathcal{C} = r \big| Z) = \varpi\{r, G_r(Z)\}.$$

The probability of coarsening depends on Z only as a function of the observed data.

3. *Noncoarsening at random*:
Noncoarsening at random (NCAR) corresponds to models where coarsening at random fails to hold. That is, the probability of coarsening depends on Z, possibly as a function of unobserved parts of Z; i.e., if there exists z_1, z_2 such that

$$G_r(z_1) = G_r(z_2)$$

for some r and

$$P(\mathcal{C} = r|z_1) \neq P(\mathcal{C} = r|z_2),$$

then the coarsening is NCAR.

As with nonmissing at random (NMAR), when coarsening is NCAR, we run into nonidentifiability problems. Therefore, we focus our attention on models where the coarsening mechanism is either CCAR or CAR.

When working with coarsened data, we distinguish among three types of data, namely *full data*, *observed data*, and *complete data*, which we now define:

- *Full data* are the data $Z_1, \ldots, Z_n$ that are iid with density $p_Z(z, \beta, \eta)$ and that we would like to observe. That is, full data are the data that we would observe had there been no coarsening. With such data we could make inference on the parameter β using standard statistical techniques developed for such parametric or semiparametric models.
- Because of coarsening, full data are not observed; instead, the *observed data* are denoted by iid random quantities

$$[\{\mathcal{C}_1, G_{\mathcal{C}_1}(Z_1)\}, \ldots, \{\mathcal{C}_n, G_{\mathcal{C}_n}(Z_n)\}],$$

where $\mathcal{C}_i$ denotes the coarsening variable and $G_{\mathcal{C}_i}(Z_i)$ denotes the corresponding coarsened data for the i-th individual in the sample. It is observed data that are available to us for making inference on the parameter β.
- Finally, when $\mathcal{C}_i = \infty$, then the data for the i-th individual are not coarsened (i.e., when $\mathcal{C}_i = \infty$, we observe the data Z_i). Therefore, we denote by *complete data* the data only for individuals i in the sample such that $\mathcal{C}_i = \infty$ (i.e., complete data are $\{Z_i : \text{ for all } i \text{ such that } \mathcal{C}_i = \infty\}$). Complete data are often used for statistical analysis in many software packages when there are missing data.

7.2 The Density and Likelihood of Coarsened Data

In order to find observed (coarsened)-data estimators of the parameter of interest β using likelihood methods, or, for that matter, in order to derive the underlying semiparametric theory, we need to derive the likelihood of the observed data in terms of the parameter β and other nuisance parameters. To derive the density of the observed data, we first consider the unobservable random vectors

$$\{(\mathcal{C}_i, Z_i), i = 1, \ldots, n\} \quad \text{assumed iid.}$$

We emphasize that these data are unobservable because, when $\mathcal{C}_i = r,\ r \neq \infty$, we only get to observe the many-to-one transformation $G_r(Z_i)$ and not Z_i itself. Nonetheless, working with such data will make the assumptions necessary to obtain the likelihood of the observed data transparent. It is sometimes convenient to denote the data $\{(\mathcal{C}_i, Z_i), i = 1, \ldots, n\}$ as the full data, whereas previously we said that the full data will be defined as $\{Z_i,\ i = 1, \ldots, n\}$. Therefore, in what follows, we will sometimes refer to full data by $\{Z_i,\ i = 1, \ldots, n\}$ and other times by $\{(\mathcal{C}_i, Z_i), i = 1, \ldots, n\}$, and the distinction should be clear by the context.

Since the observed data $\{\mathcal{C}, G_{\mathcal{C}}(Z)\}$ are a known function of the full data $(\mathcal{C}, Z)$, this means that the density of the observed data is induced by the density of the full data. The density of the full data and the corresponding likelihood, in terms of the parameters β, η, and ψ, are given by

$$p_{\mathcal{C},Z}(r, z, \psi, \beta, \eta) = P(\mathcal{C} = r | Z = z, \psi) p_Z(z, \beta, \eta).$$

That is, the density and likelihood of the full data are deduced from the density and likelihood of Z, $p_Z(z, \beta, \eta)$, and the density and likelihood for the coarsening mechanism (i.e., the probability of coarsening given Z). We emphasize that the density for the coarsening mechanism may also be from a model that is described through the parameter ψ.

Remark 3. Since the coarsening variable $\mathcal{C}$ is discrete, the dominating measure for $\mathcal{C}$ is the counting measure that puts unit mass on each of the finite values that $\mathcal{C}$ can take including $\mathcal{C} = \infty$. The dominating measure for Z is, as before, defined to be ν_Z (generally the Lebesgue measure for a continuous random variable, the counting measure for a discrete random variable, or a combination when Z is a random vector of continuous and discrete random variables). Consequently, the dominating measure for the densities of $(\mathcal{C}, Z)$ is just the product of the counting measure for $\mathcal{C}$ by ν_Z. □

Discrete Data

For simplicity, we will first consider the case when Z itself is a discrete random vector. Consequently, the dominating measure is the counting measure over the discrete combinations of $\mathcal{C}$ and Z, and integrals with respect to such a

dominating measure are just sums. Although this is overly simplistic, it will be instructive in describing the probabilistic structure of the problem. We will also indicate how this can be generalized to continuous Z as well.

Thus, with discrete data, the probability density of the observed data $\{\mathcal{C}, G_{\mathcal{C}}(Z)\}$ is derived as

$$\begin{aligned} P\{\mathcal{C} = r, G_{\mathcal{C}}(Z) = g_r\} &= \sum_{\{z:G_r(z)=g_r\}} P(\mathcal{C} = r, Z = z) \\ &= \sum_{\{z:G_r(z)=g_r\}} P(\mathcal{C} = r|Z = z)P(Z = z). \end{aligned}$$

Remark 4. Rather than developing one set of notation for discrete Z and another set of notation (using integrals) for continuous Z, we will, from now on, use integrals with respect to the appropriate dominating measure. Therefore, when we have discrete Z, and ν_Z is the corresponding counting measure, we will denote

$$P(Z \in A) = \sum_{z \in A} P(Z = z)$$

as

$$\int_{z \in A} p_Z(z) d\nu_Z(z). \; \square$$

With this convention in mind, we write the density and likelihood of the observed data, when Z is discrete, as

$$\begin{aligned} p_{\mathcal{C},G_{\mathcal{C}}(Z)}(r, g_r, \psi, \beta, \eta) &= \int_{\{z:G_r(z)=g_r\}} p_{\mathcal{C},Z}(r, z, \psi, \beta, \eta) d\nu_Z(z) \\ &= \int_{\{z:G_r(z)=g_r\}} P(\mathcal{C} = r|Z = z, \psi) p_Z(z, \beta, \eta) d\nu_Z(z). \end{aligned} \tag{7.1}$$

Continuous Data

It will be instructive to indicate how the density of the observed data would be derived if Z was a continuous random vector and the relationship of this density to (7.1). For example, consider the case where $Z = (Z_1, \ldots, Z_l)^T$ is continuous. Generally, $G_r(Z)$ is a dimensional-reduction transformation, unless $r = \infty$. This is certainly the case for missing-data mechanisms.

Let $G_r(z)$ be l_r-dimensional, $l_r < l$ for $r \neq \infty$, and assume there exists a function $V_r(z)$ that is $(l - l_r)$-dimensional so that the mapping

$$z \leftrightarrow \{G_r^T(z), V_r^T(z)\}^T$$

is one-to-one for all r. Define the inverse transformation by

$$z = H_r(g_r, v_r).$$

Then, by the standard formula for change of variables, the density

$$p_{G_r,V_r}(g_r, v_r) = p_Z\{H_r(g_r, v_r)\}J(g_r, v_r), \tag{7.2}$$

where J is the Jacobian (more precisely, the Jacobian determinant) of H_r with respect to (g_r, v_r). If we want to find the density of the observed data, namely $p_{\mathcal{C},G_\mathcal{C}}(r, g_r)$, we can use

$$p_{\mathcal{C},G_\mathcal{C}}(r, g_r) = \int p_{\mathcal{C},G_\mathcal{C},V_\mathcal{C}}(r, g_r, v_r)dv_r, \tag{7.3}$$

where

$$p_{\mathcal{C},G_\mathcal{C},V_\mathcal{C}}(r, g_r, v_r) = P(\mathcal{C} = r|G_r = g_r, V_r = v_r)p_{G_r,V_r}(g_r, v_r).$$

Note that

$$P(\mathcal{C} = r|G_r = g_r, V_r = v_r) = P\{\mathcal{C} = r|Z = H_r(g_r, v_r)\}. \tag{7.4}$$

Consequently, using (7.2) and (7.4), we can write (7.3), including the parameters ψ, β, and η, as

$$\begin{aligned} &p_{\mathcal{C},G_\mathcal{C}}(r, g_r, \psi, \beta, \eta) \\ &= \int P\{\mathcal{C} = r|Z = H_r(g_r, v_r), \psi\}p_Z\{H_r(g_r, v_r), \beta, \eta\}J(g_r, v_r)dv_r. \end{aligned} \tag{7.5}$$

Therefore, the only difference between (7.1) for discrete Z and (7.5) for continuous Z is the Jacobian that appears in (7.5). Since the Jacobian does not involve parameters in the model, it will not have an effect on the subsequent likelihood.

Likelihood when Data Are Coarsened at Random

The likelihood we derived in (7.1) was general, as it allowed for any coarsening mechanism, including NCAR. As we indicated earlier, we will restrict attention to coarsening at random mechanisms (CAR), where $P(\mathcal{C} = r|Z = z) = \varpi\{r, G_r(z)\}$ for all r, z. Coarsening completely at random (CCAR) is just a special case of this. We remind the reader that another key assumption being made throughout is that $P(\mathcal{C} = \infty|Z = z) \geq \varepsilon > 0$ for all r, z.

Now that we have shown how to derive the likelihood of the observed data from the marginal density of the desired data Z and the coarsening mechanism, $P(\mathcal{C} = r|Z = z)$, we can now derive the likelihood of the observed data when coarsening is CAR. To do so, we need to consider a model for the coarsening density in terms of parameters. For the time being, we will be very general and denote such a model by

$$P(\mathcal{C} = r|Z = z) = \varpi\{r, G_r(z), \psi\},$$

where ψ is an unknown parameter that is functionally independent of (β, η), the parameters for the full-data model.

Remark 5. The coarsening or missingness of the data can be by design where the probability $\varpi\{r, G_r(z)\}$ is known to the investigator. For such problems, additional parameters ψ are not necessary. In other cases, where coarsening or missingness occur by happenstance, we may introduce parametric models with a finite-dimensional parameter ψ or semiparametric models with infinite-dimensional parameter ψ, where ψ needs to be estimated. The exercise of finding reasonable and coherent models for $\varpi\{r, G_r(z), \psi\}$, even under the MAR or CAR assumption, may not be straightforward and may require special considerations. Examples of such models will be given throughout the remainder of the book but, for the time being, it will be assumed that the model for $\varpi\{r, G_r(z), \psi\}$, as a function of ψ, is known and has been correctly specified. □

We now see that the observed data can be viewed as realizations of the iid random quantities $\{\mathcal{C}_i, G_{\mathcal{R}_i}(Z_i)\}$, $i = 1, \ldots, n$, with density from a statistical model described through the parameter of interest β and nuisance parameters η and ψ. The CAR assumption will allow simplification of the likelihood, as we now demonstrate.

When data are CAR, the likelihood of the observed data for a single observation given by (7.1) for discrete Z (now considered as functions of (ψ, β, η)) is

$$\begin{aligned} p_{\mathcal{C},G_{\mathcal{C}}(Z)}(r, g_r, \psi, \beta, \eta) &= \int\limits_{\{z:G_r(z)=g_r\}} P(\mathcal{C} = r|Z = z, \psi) p_Z(z, \beta, \eta) d\nu_Z(z) \\ &= \int\limits_{\{z:G_r(z)=g_r\}} \varpi(r, g_r, \psi) p_Z(z, \beta, \eta) d\nu_Z(z) \\ &= \varpi(r, g_r, \psi) \int\limits_{\{z:G_r(z)=g_r\}} p_Z(z, \beta, \eta) d\nu_Z(z). \end{aligned} \tag{7.6}$$

Notice that the parameter ψ for the coarsening process separates from the parameters (β, η) that describe the full-data model. Also notice that if Z were continuous and we used formula (7.5) to derive the density, then, under CAR, we obtain

$$p_{\mathcal{C},G_{\mathcal{C}}(Z)}(r,g_r,\psi,\beta,\eta) = \int \varpi(r,g_r,\psi)p_Z\{H_r(g_r,v_r),\beta,\eta\}\underbrace{J(g_r,v_r)}_{\Downarrow}dv_r$$

The Jacobian does not involve any of the parameters.

$$= \varpi(r,g_r,\psi)\int p_Z\{H_r(g_r,v_r),\beta,\eta\}J(g_r,v_r)dv_r. \quad (7.7)$$

In both (7.6) and (7.7),

$$p_{\mathcal{C},G_{\mathcal{C}}(Z)}(r,g_r,\psi,\beta,\eta) = \varpi(r,g_r,\psi)p_{G_r(Z)}(g_r,\beta,\eta). \quad (7.8)$$

Brief Remark on Likelihood Methods

Because the parameter ψ describing the coarsening mechanism separates from the parameters (β,η) describing the distribution of Z in the observed data likelihood, when the data are coarsened at random, likelihood methods are often proposed for estimating the parameters β and η.

That is, suppose we posit a parametric model for full data (i.e., $p_Z(z,\beta,\eta)$) and the aim is to estimate the parameter β using coarsened data

$$\{\mathcal{C}_i, G_{\mathcal{C}_i}(Z_i)\},\ i = 1,\ldots,n.$$

The likelihood for a realization of such data, $(r_i,g_{r_i}), i = 1,\ldots,n$, as a function of the parameters, when the coarsening mechanism is CAR, is (by (7.8)) equal to

$$\left\{\prod_{i=1}^n \varpi(r_i,g_{r_i},\psi)\right\}\left\{\prod_{i=1}^n p_{G_{r_i}(Z)}(g_{r_i},\beta,\eta)\right\}. \quad (7.9)$$

Therefore, the MLE for (β,η) only involves maximizing the function

$$\prod_{i=1}^n p_{G_{r_i}(Z)}(g_{r_i},\beta,\eta), \quad (7.10)$$

where

$$p_{G_r(Z)}(g_r,\beta,\eta) = \int\limits_{\{z:G_r(z)=g_r\}} p_Z(z,\beta,\eta)d\nu_Z(z).$$

Therefore, as long as we believe the CAR assumption, we can find the MLE for β and η without having to specify a model for the coarsening process. If the parameter space for (β,η) is finite-dimensional, this is especially attractive, as the MLE for β, under suitable regularity conditions, is an efficient estimator. Moreover, the coarsening probabilities, subject to the CAR assumption, play no role in either the estimation of β (or η for that matter) or the efficiency

of such an estimator. This has a great deal of appeal, as it frees the analyst from making modeling assumptions for the coarsening probabilities.

Maximizing functions such as (7.10) to obtain the MLE may involve integrals and complicated expressions that may not be easy to implement. Nevertheless, there has been a great deal of progress in developing optimization techniques involving quadrature or Monte Carlo methods, as well as other maximization routines such as the EM algorithm, which may be useful for this purpose. Since likelihood methods for missing (coarsened) data have been studied in great detail by others, there will be relatively little discussion of these methods in this book. We refer the reader to Allison (2002), Little and Rubin (1987), Schafer (1997), and Verbeke and Molenberghs (2000) for more details on likelihood methods for missing data.

Examples of Coarsened-Data Likelihoods

Return to Example 1

Let us return to Example 1 of Section 7.1. Recall that in this example two blood samples of equal volume were taken from each of n individuals in a study that measured the blood concentration of some biological marker. Some of the individuals, chosen at random, had concentration measurements made on both samples. These are denoted as X_1 and X_2. The remaining individuals had their blood samples combined and only one concentration measurement was made, equaling $(X_1+X_2)/2$. Although concentrations must be positive, let us, for simplicity, assume that the distribution of these blood concentrations is well approximated by a normal distribution. To assess the variability of these concentrations between and within individuals, we assume that $X_j = \alpha + e_j$, where α is normally distributed with mean μ_α and variance σ^2_α, and $e_j, j = 1, 2$ are independently normally distributed with mean zero and variance σ^2_e, also independent of α. With this representation, σ^2_α represents the variation between individuals and σ^2_e the variation within an individual. From this model, we deduce that $Z = (X_1, X_2)^T$ follows a bivariate normal distribution with common mean μ_α and common variance $\sigma^2_\alpha + \sigma^2_e$ and covariance σ^2_α.

Since the individuals chosen to have their blood samples combined were chosen at random, this is an example of coarsening completely at random (CCAR). Thus $P(\mathcal{C} = 1|Z) = \varpi$, where ϖ is the probability of being selected in the subsample and $P(\mathcal{C} = \infty|Z) = 1 - \varpi$.

The data for this example can be represented as realizations of the iid random vectors $\{\mathcal{C}_i, G_{\mathcal{C}_i}(Z_i)\}, i = 1, \ldots, n$, where, if $\mathcal{C}_i = \infty$, then we observe $G_\infty(Z_i) = (X_{i1}, X_{i2})$, whereas if $\mathcal{C}_i = 1$, then we observe $G_1(Z_i) = (X_{i1} + X_{i2})/2$. Under the assumptions of the model,

$$\begin{pmatrix} X_{i1} \\ X_{i2} \end{pmatrix} \sim N\left(\begin{pmatrix} \mu_\alpha \\ \mu_\alpha \end{pmatrix}, \Sigma \right), \tag{7.11}$$

where

$$\Sigma = \begin{pmatrix} \sigma_\alpha^2 + \sigma_e^2 \, , & \sigma_\alpha^2 \\ \sigma_\alpha^2 & , \, \sigma_\alpha^2 + \sigma_e^2 \end{pmatrix}.$$

It is also straightforward to show that

$$(X_{i1} + X_{i2})/2 \sim N\Big(\mu_\alpha, \sigma_\alpha^2 + \sigma_e^2/2\Big).$$

Consequently, the MLE for $(\mu_\alpha, \sigma_\alpha^2, \sigma_e^2)$ is obtained by maximizing the coarsened-data likelihood (7.10), which, for this example, is given by

$$\prod_{i=1}^{n} \Bigg\{ \Bigg(|\Sigma|^{-1/2} \exp\Bigg[-\frac{1}{2} \{(X_{i1} - \mu_\alpha, X_{i2} - \mu_\alpha)^T \Sigma^{-1} (X_{i1} - \mu_\alpha, X_{i2} - \mu_\alpha)\} \Bigg] \Bigg)^{I(\mathcal{C}_i = \infty)} \times \Bigg((\sigma_\alpha^2 + \sigma_e^2/2)^{-1/2} \exp\left[-\frac{\{(X_{i1} + X_{i2})/2 - \mu_\alpha\}^2}{2(\sigma_\alpha^2 + \sigma_e^2/2)} \right] \Bigg)^{I(\mathcal{C}_i = 1)} \Bigg\}. \quad (7.12)$$

We leave the calculation of the MLE for this example as an exercise at the end of the chapter.

Although maximizing the likelihood is the preferred method for obtaining estimators for the parameters in finite-dimensional parametric models of the full data Z, it may not be a feasible approach for semiparametric models when the parameter space of the full data is infinite-dimensional. We illustrate some of the difficulties through an example where the parameter of interest is easily estimated using likelihood techniques if the data are not coarsened but where likelihood methods become difficult when the data are coarsened.

The logistic regression model

Let Y be a binary response variable $\{0, 1\}$, and let X be a vector of covariates. A popular model for modeling the probability of response as a function of the covariates X is the logistic regression model where

$$P(Y = 1|X) = \frac{\exp(\beta^T X^*)}{1 + \exp(\beta^T X^*)},$$

where $X^* = (1, X^T)^T$, allowing us to consider an intercept term. We make no assumptions on X. With full data, $(Y_i, X_i), i = 1, \ldots, n$, the likelihood of a single observation is

$$p_{Y|X}(y|x) p_X(x) = \left[\frac{\exp\{(\beta^T x^*) y\}}{1 + \exp(\beta^T x^*)} \right] p_X\{x, \eta(\cdot)\}, \quad (7.13)$$

where the parameter $\eta(\cdot)$ is the infinite-dimensional nuisance function allowing all nonparametric densities for the marginal distribution of X.

If we use maximum likelihood to estimate β with full data, then because the parameters β and $\eta(\cdot)$ separate in (7.13), it suffices to maximize

$$\prod_{i=1}^{n} \left[\frac{\exp\{(\beta^T X_i^*) Y_i\}}{1 + \exp(\beta^T X_i^*)} \right] \tag{7.14}$$

without any regard to the nuisance parameter η. Equivalently, we can derive the maximum likelihood estimator for β by solving the score equations

$$\sum_{i=1}^{n} X_i^* \left\{ Y_i - \frac{\exp(\beta^T X_i^*)}{1 + \exp(\beta^T X_i^*)} \right\} = 0. \tag{7.15}$$

This was also derived in (4.65). This indeed is the standard analytic strategy for obtaining estimators for β in a logistic regression model implemented in most statistical software packages.

If, however, we had coarsened data (CAR), then the likelihood contribution for the part of the likelihood that involves β for a single observation is

$$\int_{\{(y,x): G_r(y,x) = g_r\}} \left\{ \frac{\exp(\beta^T x^*) y}{1 + \exp(\beta^T x^*)} \right\} p_X\{x, \eta(\cdot)\} d\nu_{Y,X}(y, x). \tag{7.16}$$

Whereas maximizing the likelihood in β in (7.14) for noncoarsened data involved only the parameter β, finding the MLE for β with coarsened data now involves maximizing a function in both β and the infinite-dimensional parameter $\eta(\cdot)$ in a likelihood that involves a product over i of terms like (7.16). Such maximization may be much more difficult if not impossible. We will return to this example later.

Consequently, it is important to consider alternatives to likelihood methods for estimating parameters with coarsened data. Before providing such alternatives, it is useful to go back to first principles and study the geometry of influence functions of estimators for the parameter β with coarsened data when the coarsening is CAR.

7.3 The Geometry of Semiparametric Coarsened-Data Models

The key to deriving the class of influence functions of RAL estimators for the parameter β and the corresponding geometry with coarsened data is to build on the corresponding theory of influence functions of estimators for β and its geometry had the data not been coarsened (i.e., with full data). In so doing, we need to distinguish between the geometry of full-data Hilbert spaces and that of observed-data Hilbert spaces.

We denote the full-data Hilbert space of all q-dimensional, mean-zero measurable functions of Z with finite variance equipped with the covariance inner

product by $\mathcal{H}^F$. This is contrasted with the observed-data Hilbert space of all q-dimensional, mean-zero, finite variance, measurable functions of $\{\mathcal{C}, G_{\mathcal{C}}(Z)\}$ equipped with the covariance inner product, which we denote by $\mathcal{H}$ (without the superscript F). In some cases, we may also consider the Hilbert space $\mathcal{H}^{\mathcal{C}Z}$ of all q-dimensional, mean-zero, finite variance, measurable functions of $(\mathcal{C}, Z)$ equipped with the covariance inner product. We note that $\mathcal{H}^F$ and $\mathcal{H}$ are both linear subspaces of $\mathcal{H}^{\mathcal{C}Z}$.

Because influence functions lie in a subspace of $\mathcal{H}$ orthogonal to the nuisance tangent space, the key in identifying influence functions is first to find the nuisance tangent space and its orthogonal complement. We remind the reader that

- The full-data nuisance tangent space is the mean-square closure of all full-data parametric submodel nuisance tangent spaces. The full-data nuisance tangent space is denoted by Λ^F.
- For a full-data parametric submodel $p_Z(z, \beta, \gamma)$, where β is q-dimensional and γ is r-dimensional, the nuisance score vector is

$$S_\gamma^F(Z) = \frac{\partial \log p_Z(Z, \beta_0, \gamma_0)}{\partial \gamma},$$

  and the full-data parametric submodel nuisance tangent space is the space spanned by S_γ^F; namely,

$$\left\{B^{q\times r} S_\gamma^F(Z) \text{ for all } q \times r \text{ matrices } B\right\}.$$

- The class of full-data influence functions are the elements $\varphi^F(Z) \in \mathcal{H}^F$ such that
  (i) $\varphi^F(Z) \in \Lambda^{F\perp}$ (i.e., orthogonal to the full-data nuisance tangent space)
  (ii) $E\{\varphi^F(Z) S_\beta^{F^T}(Z)\} = I^{q\times q}$ (identity matrix), where

$$S_\beta^F = \frac{\partial \log p_Z(Z, \beta_o, \eta_o)}{\partial \beta}.$$

- The efficient full-data score is

$$S_{\text{eff}}^F(Z) = S_\beta^F(Z) - \Pi\{S_\beta^F(Z) | \Lambda^F\},$$

  and the efficient full-data influence function is

$$\varphi_{\text{eff}}^F(Z) = \left[E\left\{S_{\text{eff}}^F(Z) S_{\text{eff}}(Z)^{F^T}\right\}\right]^{-1} S_{\text{eff}}^F(Z).$$

When considering missing or coarsened data, two issues need to be addressed:

(i) What is the class of observed-data influence functions, and how are they related to full-data influence functions?

(ii) How can we characterize the most efficient influence function and the semiparametric efficiency bound for coarsened-data semiparametric models?

Both of these, as well as many other issues regarding semiparametric estimators with coarsened data, will be studied carefully over the next several chapters.

We remind the reader that observed-data influence functions of RAL estimators for β must be orthogonal to the observed-data nuisance tangent space, which we denote by Λ (without the superscript F). Therefore, we will demonstrate how to derive the observed-data nuisance tangent space and its orthogonal complement.

When the data are discrete and coarsening is CAR, the observed-data likelihood for a single observation is given by (7.6); namely,

$$p_{\mathcal{C},G_{\mathcal{C}}(Z)}(r, g_r, \psi, \beta, \eta) = \varpi(r, g_r, \psi) \int\limits_{\{G_r(z)=g_r\}} p_Z(z, \beta, \eta) d\nu_Z(z).$$

A similar expression for continuous variables, involving Jacobians, was given by (7.7). From here on, we will use the representation of likelihood for discrete data, realizing that these results can be easily generalized to continuous variables using Jacobians.

The log-likelihood for a single observation is given by

$$\log \varpi(r, g_r, \psi) + \log \int\limits_{\{G_r(z)=g_r\}} p_Z(z, \beta, \eta) d\nu_Z(z). \tag{7.17}$$

The coarsened-data likelihood and log-likelihood are functions of the parameters β, η, and ψ, where β is the parameter of interest and hence η and ψ are nuisance parameters. Since the parameters η and ψ separate out in the likelihood for the observed data, we would expect that the nuisance tangent space will be the direct sum of two orthogonal spaces, one involving the space generated by the score vector with respect to ψ, which we denote by Λ_ψ, and the other space generated by the score vector with respect to η, which we denote by Λ_η. That is,

$$\Lambda = \Lambda_\psi \oplus \Lambda_\eta, \quad \Lambda_\psi \perp \Lambda_\eta. \tag{7.18}$$

We will give a formal proof of (7.18) later.

For the remainder of this chapter, we will only consider the space Λ_η and its complement. When the coarsening of the data is by design, where the coarsening probabilities $\varpi\{r, G_r(z)\}$ are known to the investigator, then there is no need to introduce an additional parameter ψ or the space Λ_ψ. In that case, the observed-data nuisance tangent space Λ is the same as Λ_η. Examples where the data are missing by design will be given in Section 7.4. We restrict ourselves to this situation for the time being. In the next chapter, we will

discuss what to do when the coarsening probabilities are not known to us by design and models for $\varpi\{r, G_r(z), \psi\}$, as a function of the parameter ψ, have to be developed.

The Nuisance Tangent Space Associated with the Full-Data Nuisance Parameter and Its Orthogonal Complement

The nuisance tangent space

The space Λ_η is defined as the mean-square closure of parametric submodel tangent spaces associated with the nuisance parameter η. Therefore, we begin by first considering the parametric submodel for the full-data Z given by $p_Z(z, \beta^{q\times 1}, \gamma^{r\times 1})$ and compute the corresponding observed-data score vector.

Lemma 7.1. The parametric submodel observed-data score vector with respect to γ is given by

$$S_\gamma(r, g_r) = E\left\{S_\gamma^F(Z)|G_r(Z) = g_r\right\}. \tag{7.19}$$

Proof. The log-likelihood for the observed data (at least the part that involves γ), given by (7.17), is

$$\log\left\{\int\limits_{\{G_r(z)=g_r\}} p_Z(z, \beta, \gamma)d\nu_Z(z)\right\}.$$

The score vector with respect to γ is

$$S_\gamma(r, g_r) = \frac{\partial}{\partial\gamma}\left[\log\left\{\int\limits_{\{G_r(z)=g_r\}} p_Z(z, \beta, \gamma)d\nu_Z(z)\right\}\right]\Bigg|_{\underbrace{\substack{\beta=\beta_0\\ \gamma=\gamma_0}}_{\substack{\| \\ \text{same as } \eta=\eta_0}}}$$

$$= \frac{\displaystyle\int\limits_{\{G_r(z)=g_r\}} \frac{\partial}{\partial\gamma} p_Z(z, \beta_0, \gamma_0)d\nu_Z(z)}{\displaystyle\int\limits_{\{G_r(z)=g_r\}} p_Z(z, \beta_0, \gamma_0)d\nu_Z(z)}. \tag{7.20}$$

Dividing and multiplying by $p_Z(z, \beta_0, \gamma_0)$ in the integral of the numerator of (7.20) yields

$$\frac{\displaystyle\int\limits_{\{G_r(z)=g_r\}} S_\gamma^F(z, \beta_0, \gamma_0)p_Z(z, \beta_0, \gamma_0)d\nu_Z(z)}{\displaystyle\int\limits_{\{G_r(z)=g_r\}} p_Z(z, \beta_0, \gamma_0)d\nu_Z(z)}.$$

Hence,

$$S_\gamma(r, g_r) = E\left\{S_\gamma^F(Z)|G_r(Z) = g_r\right\}. \square$$

Remark 6. Equation (7.19) is a conditional expectation that only involves the conditional probability distribution of Z given $G_r(Z)$ for a fixed value of r. It will be important for the subsequent theoretical development that we compare and contrast (7.19) with the conditional expectation

$$E\left\{S_\gamma^F(Z)|\mathcal{C} = r, G_\mathcal{C}(Z) = g_r\right\}. \tag{7.21}$$

In general, (7.21) will not equal (7.19); however, as we will now show, these are equal under the assumption of CAR. $\square$

Lemma 7.2. When the coarsening mechanism is CAR, then

$$S_\gamma(r, g_r) = E\left\{S_\gamma^F(Z)|G_r(Z) = g_r\right\} = E\left\{S_\gamma^F(Z)|\mathcal{C} = r, G_\mathcal{C}(Z) = g_r\right\}. \tag{7.22}$$

Proof. Equation (7.22) will follow if we can prove that

$$p_{Z|\mathcal{C},G_\mathcal{C}(Z)}(z|r, g_r) = p_{Z|G_r(Z)}(z|g_r), \tag{7.23}$$

which is true because when $G_r(z) = g_r$

$$\begin{aligned}
p_{Z|\mathcal{C},G_\mathcal{C}(Z)}(z|r, g_r) &= \frac{p_{\mathcal{C},Z}(r, z)}{\displaystyle\int_{\{G_r(u)=g_r\}} p_{\mathcal{C},Z}(r, u)d\nu_Z(u)} \\
&= \frac{p_{\mathcal{C}|Z}(r|z)p_Z(z)}{\displaystyle\int_{\{G_r(u)=g_r\}} p_{\mathcal{C}|Z}(r|u)p_Z(u)d\nu_Z(u)} \\
&= \frac{\varpi(r, g_r)p_Z(z)}{\varpi(r, g_r)\displaystyle\int_{\{G_r(u)=g_r\}} p_Z(u)d\nu_Z(u)} \qquad (7.24) \\
&= \frac{p_Z(z)}{\displaystyle\int_{\{G_r(v)=g_r\}} p_Z(u)d\nu_Z(u)} = p_{Z|G_r(Z)}(z|g_r). \square \qquad (7.25)
\end{aligned}$$

Remark 7. In order to prove (7.23), it was necessary that $\varpi(r, g_r)$ cancel in the numerator and denominator of (7.24), which is only true because of CAR. $\square$

Consequently, when the coarsening mechanism is CAR, the corresponding observed-data nuisance score vector for the parametric submodel $p_Z(z, \beta, \gamma)$ is

$$S_\gamma\{\mathcal{C}, G_\mathcal{C}(Z)\} = E\left\{S_\gamma^F(Z)|\mathcal{C}, G_\mathcal{C}(Z)\right\}. \tag{7.26}$$

We are now in a position to define the observed-data nuisance tangent space associated with the full-data nuisance parameter η.

Theorem 7.1. The space Λ_η (i.e., the mean square closure of parametric submodel nuisance tangent spaces spanned by $S_\gamma\{\mathcal{C}, G_\mathcal{C}(Z)\}$) is the space of elements

$$\Lambda_\eta = \left[E\left\{\alpha^F(Z)|\mathcal{C}, G_\mathcal{C}(Z)\right\} \text{ for all } \alpha^F \in \Lambda^F\right], \tag{7.27}$$

where Λ^F denotes the full-data nuisance tangent space. We will also denote this space by the shorthand notation

$$\Lambda_\eta = E\left\{\Lambda^F|\mathcal{C}, G_\mathcal{C}(Z)\right\}.$$

Proof. Using (7.26), we note that the linear subspace, within $\mathcal{H}$, spanned by the parametric submodel score vector $S_\gamma\{\mathcal{C}, G_\mathcal{C}(Z)\}$ is

$$\begin{aligned}&\left[B^{q\times r}E\left\{S_\gamma^F(Z)|\mathcal{C}, G_\mathcal{C}(Z)\right\} \text{ for all } B^{q\times r}\right]\\ &= \left[E\left\{B^{q\times r}S_\gamma^F(Z)|\mathcal{C}, G_\mathcal{C}(Z)\right\} \text{ for all } B^{q\times r}\right].\end{aligned}$$

The linear subspace Λ_η consisting of elements $B^{q\times r}E\{S_\gamma^F(Z)|\mathcal{C}, G_\mathcal{C}(Z)\}$ for some parametric submodel or a limit (as $n \to \infty$) of elements

$$B_n^{q\times r_n}E\{S_{\gamma n}^F(Z)|\mathcal{C}, G_\mathcal{C}(Z)\}$$

for a sequence of parametric submodels and conformable matrices. This is the same as the space consisting of elements

$$E\left\{B^{q\times r}S_\gamma^F(Z)|\mathcal{C}, G_\mathcal{C}(Z)\right\}$$

or limits of elements

$$E\left\{B_n^{q\times r_n}S_{\gamma n}^F(Z)|\mathcal{C}, G_\mathcal{C}(Z)\right\}.$$

But the space of elements $B^{q\times r}S_\gamma^F(Z)$ or limits of elements $B_n^{q\times r_n}S_{\gamma n}^F(Z)$ for parametric submodels is precisely the definition of the full-data nuisance tangent space Λ^F. Consequently, the space Λ_η can be characterized as the space of elements

$$\Lambda_\eta = \left[E\left\{\alpha^F(Z)|\mathcal{C}, G_\mathcal{C}(Z)\right\} \text{ for all } \alpha^F \in \Lambda^F\right]. \square$$

In the special case where data are missing or coarsened by design (i.e., when there are no additional parameters ψ necessary to define a model for the coarsening probabilities), then the observed-data nuisance tangent space is $\Lambda = \Lambda_\eta$. We know that an influence function of an observed-data RAL estimator for β must be orthogonal to Λ. Toward that end, we now characterize the space orthogonal to Λ_η (i.e., $\Lambda_\eta^\perp$).

The orthogonal complement of the nuisance tangent space

Lemma 7.3. The space $\Lambda_\eta^\perp$ consists of all elements $h^{q\times 1}\{\mathcal{C}, G_{\mathcal{C}}(Z)\} \in \mathcal{H}$ such that

$$E[h\{\mathcal{C}, G_{\mathcal{C}}(Z)\}|Z] \in \Lambda^{F\perp},$$

where $\Lambda^{F\perp}$ is the space orthogonal to the full-data nuisance tangent space.

Proof. The space $\Lambda_\eta^\perp$ consists of all elements $h^{q\times 1}\{\mathcal{C}, G_{\mathcal{C}}(Z)\} \in \mathcal{H}$ that are orthogonal to Λ_η. By Theorem 7.1, this corresponds to the set of elements $h(\cdot) \in \mathcal{H}$ such that

$$\begin{aligned} &E\left[h^T\{\mathcal{C}, G_{\mathcal{C}}(Z)\}E\{\alpha^F(Z)|\mathcal{C}, G_{\mathcal{C}}(Z)\}\right] = 0 \\ &\text{for all} \quad \alpha^F(Z) \in \Lambda^F. \end{aligned} \tag{7.28}$$

Using the law of iterated conditional expectations repeatedly, we obtain the following relationship for (7.28):

$$\begin{aligned} 0 &= E\left(E[h^T\{\mathcal{C}, G_{\mathcal{C}}(Z)\}\alpha^F(Z)|\mathcal{C}, G_{\mathcal{C}}(Z)]\right) \\ &= E[h^T\{\mathcal{C}, G_{\mathcal{C}}(Z)\}\alpha^F(Z)] \\ &= E\left(E[h^T\{\mathcal{C}, G_{\mathcal{C}}(Z)\}\alpha^F(Z)|Z]\right) \\ &= E\left(E[h^T\{\mathcal{C}, G_{\mathcal{C}}(Z)\}|Z]\ \alpha^F(Z)\right) \\ &\qquad \text{for all } \alpha^F(Z) \in \Lambda^F. \end{aligned} \tag{7.29}$$

Thus (7.29) implies that $h\{\mathcal{C}, G_{\mathcal{R}}(Z)\} \in \mathcal{H}$ belongs to $\Lambda_\eta^\perp$ if and only if $E[h\{\mathcal{C}, G_{\mathcal{C}}(Z)\}|Z]$ is orthogonal to every element $\alpha^F(Z) \in \Lambda^F$; i.e., that

$$E[h\{\mathcal{C}, G_{\mathcal{C}}(Z)\}|Z] \in \Lambda^{F\perp}. \ \square$$

To explore this relationship further, it will be convenient to introduce the notion of a mapping, and more specifically a linear mapping, from one Hilbert space to another Hilbert space.

Definition 1. A mapping, also sometimes referred to as an operator, $\mathcal{K}$, is a function that maps each element of some linear space to an element of another linear space. In all of our applications, the linear spaces will be well-defined Hilbert spaces. So, for example, if $\mathcal{H}^{(1)}$ and $\mathcal{H}^{(2)}$ denote two Hilbert spaces, then the mapping $\mathcal{K} : \mathcal{H}^{(1)} \to \mathcal{H}^{(2)}$ means that for any $h \in \mathcal{H}^{(1)}$, $\mathcal{K}(h) \in \mathcal{H}^{(2)}$. A linear mapping also has the property that $\mathcal{K}(ah_1 + bh_2) = a\mathcal{K}(h_1) + b\mathcal{K}(h_2)$ for any two elements $h_1, h_2 \in \mathcal{H}^{(1)}$ and any scalar constants a and b. A many-to-one mapping means that more than one element $h \in \mathcal{H}^{(1)}$ will map to the same element in $\mathcal{H}^{(2)}$. For more details regarding linear operators, we refer the reader to Chapter 6 of Luenberger (1969). $\square$

Define the many-to-one mapping

$$\mathcal{K} : \mathcal{H} \to \mathcal{H}^F$$

to be

$$\mathcal{K}(h) = E[h\{\mathcal{C}, G_{\mathcal{C}}(Z)\}|Z] \tag{7.30}$$

for $h \in \mathcal{H}$. Because of the linear properties of conditional expectations, the mapping $\mathcal{K}$, given by (7.30), is a linear mapping or linear operator.

By Lemma 7.3, the space $\Lambda_\eta^\perp$ can be defined as

$$\Lambda_\eta^\perp = \mathcal{K}^{-1}(\Lambda^{F\perp}),$$

where $\mathcal{K}^{-1}$ is the inverse operator.

Definition 2. *Inverse operator*
For any element $h^F \in \mathcal{H}^F$, $\mathcal{K}^{-1}(h^F)$ corresponds to the set of all elements (assuming at least one exists) $h \in \mathcal{H}$ such that $\mathcal{K}(h) = h^F$. Similarly, the space $\mathcal{K}^{-1}(\Lambda^{F\perp})$ corresponds to all elements of $h \in \mathcal{H}$ such that $\mathcal{K}(h) \in \Lambda^{F\perp}$. □

Since $\mathcal{K}$ is a linear operator and $\Lambda^{F\perp}$ is a linear subspace of $\mathcal{H}^F$, it is easy to show that $\mathcal{K}^{-1}(\Lambda^{F\perp})$ is a linear subspace of $\mathcal{H}$.

Let us consider the construction of the space $\Lambda_\eta^\perp = \mathcal{K}^{-1}(\Lambda^{F\perp})$ element by element. Consider a single element $\varphi^{*F}(Z) \in \Lambda^{F\perp}$. In the following theorem, we show how the inverse $\mathcal{K}^{-1}(\varphi^{*F})$ is computed.

Remark 8. Notation convention
When we refer to elements of the space $\Lambda^{F\perp}$, we will use the notation $\varphi^{*F}(Z)$. This is in contrast to the notation $\varphi^F(Z)$ (without the $*$), which we use to denote a full-data influence function. The space perpendicular to the full-data nuisance tangent space $\Lambda^{F\perp}$ is the space in which the class of full-data influence functions belongs. In order to be a full-data influence function, an element of $\Lambda^{F\perp}$ must also satisfy the property that $E\{\varphi^F(Z)S_{\text{eff}}^{F^T}(Z)\} = I^{q\times q}$. We remind the reader that, for any $\varphi^{*F}(Z) \in \Lambda^{F\perp}$, we can construct an influence function that equals $\varphi^{*F}(Z)$, up to a multiplicative constant, by taking

$$\varphi^F(Z) = \left[E\{\varphi^{*F}(Z)S_{\text{eff}}^{F^T}(Z)\}\right]^{-1}\varphi^{*F}(Z). \ \square$$

Lemma 7.4. For any $\varphi^{*F}(Z) \in \Lambda^{F\perp}$, let $\mathcal{K}^{-1}\{\varphi^{*F}(Z)\}$ denote the space of elements $\tilde{h}\{\mathcal{C}, G_{\mathcal{C}}(Z)\} \in \mathcal{H}$ such that

$$\mathcal{K}[\tilde{h}\{\mathcal{C}, G_{\mathcal{C}}(Z)\}] = E[\tilde{h}\{\mathcal{C}, G_{\mathcal{C}}(Z)\}|Z] = \varphi^{*F}(Z).$$

If we could identify any element $h\{\mathcal{C}, G_{\mathcal{C}}(Z)\}$ such that

$$\mathcal{K}(h) = \varphi^{*F}(Z),$$

then

$$\mathcal{K}^{-1}\{\varphi^{*F}(Z)\} = h\{\mathcal{C}, G_{\mathcal{C}}(Z)\} + \Lambda_2,$$

where Λ_2 is the linear subspace in $\mathcal{H}$ consisting of elements $L_2\{\mathcal{C}, G_{\mathcal{C}}(Z)\}$ such that

$$E[L_2\{\mathcal{C}, G_{\mathcal{C}}(Z)\}|Z] = 0;$$

that is, $\Lambda_2 = \mathcal{K}^{-1}(0)$.

Proof. The proof is straightforward. If $\tilde{h}\{\mathcal{C}, G_{\mathcal{R}}(Z)\}$ is an element of the space $h\{\mathcal{C}, G_{\mathcal{R}}(Z)\}+\Lambda_2$, then $\tilde{h}\{\mathcal{C}, G_{\mathcal{R}}(Z)\} = h\{\mathcal{C}, G_{\mathcal{C}}(Z)\}+L_2\{\mathcal{R}, G_{\mathcal{C}}(Z)\}$ for some $L_2\{\mathcal{C}, G_{\mathcal{R}}(Z)\} \in \Lambda_2$, in which case

$$\begin{aligned} E[\tilde{h}\{\mathcal{C}, G_{\mathcal{C}}(Z)\}|Z] &= E[h\{\mathcal{C}, G_{\mathcal{C}}(Z)\}|Z] + E[L_2\{\mathcal{C}, G_{\mathcal{C}}(Z)\}|Z] \\ &= \varphi^{*F}(Z) + 0 = \varphi^{*F}(Z). \end{aligned}$$

Conversely, if $E[\tilde{h}\{\mathcal{C}, G_{\mathcal{C}}(Z)\}|Z] = \varphi^{*F}(Z)$, then

$$\tilde{h}\{\mathcal{C}, G_{\mathcal{C}}(Z)\} = h\{\mathcal{C}, G_{\mathcal{C}}(Z)\} + [\tilde{h}\{\mathcal{C}, G_{\mathcal{C}}(Z)\} - h\{\mathcal{C}, G_{\mathcal{C}}(Z)\}],$$

where clearly $[\tilde{h}\{\mathcal{C}, G_{\mathcal{C}}(Z)\} - h\{\mathcal{C}, G_{\mathcal{C}}(Z)\}] \in \Lambda_2$. □

Therefore, in order to construct $\Lambda_\eta^\perp = \mathcal{K}^{-1}(\Lambda^{F\perp})$, we must, for each $\varphi^{*F}(Z) \in \Lambda^{F\perp}$,

(i) identify one element $h\{\mathcal{C}, G_{\mathcal{C}}(Z)\}$ such that

$$E[h\{\mathcal{C}, G_{\mathcal{C}}(Z)\}|Z] = \varphi^{*F}(Z),$$

and

(ii) find $\Lambda_2 = \mathcal{K}^{-1}(0)$.

We now derive the space $\Lambda_\eta^\perp$ in the following theorem.

Theorem 7.2. Under the assumption that

$$E\{I(\mathcal{C} = \infty)|Z\} = \varpi(\infty, Z) > 0 \quad \text{for all } Z \text{ (a.e.)}, \tag{7.31}$$

the space $\Lambda_\eta^\perp$ consists of all elements that can be written as

$$\begin{aligned} &\frac{I(\mathcal{C} = \infty)\varphi^{*F}(Z)}{\varpi(\infty, Z)} + \\ &\frac{I(\mathcal{C} = \infty)}{\varpi(\infty, Z)} \left[\sum_{r \neq \infty} \varpi\{r, G_r(Z)\} L_{2r}\{G_r(Z)\} \right] - \sum_{r \neq \infty} I(\mathcal{C} = r) L_{2r}\{G_r(Z)\}, \end{aligned} \tag{7.32}$$

where, for $r \neq \infty$, $L_{2r}\{G_r(Z)\}$ is an arbitrary $q \times 1$ measurable function of $G_r(Z)$ and $\varphi^{*F}(Z)$ is an arbitrary element of $\Lambda^{F\perp}$.

Proof. In accordance with the proof of Lemma 7.4, we begin by:

(i) Identifying h such that $E[h\{\mathcal{C}, G_{\mathcal{C}}(Z)\}|Z] = \varphi^{*F}(Z)$

A single solution to the equation

$$E[h\{\mathcal{C}, G_{\mathcal{C}}(Z)\}|Z] = \varphi^{*F}(Z)$$

is motivated by the idea of an inverse probability weighted complete-case estimator, which was first introduced in Section 6.4. Recall that $\mathcal{C} = \infty$ denotes the case when the data Z are completely observed and $\varpi(\infty, Z) = P(\mathcal{C} = \infty|Z)$. Now consider $h\{\mathcal{C}, G_{\mathcal{C}}(Z)\}$ to be

$$\frac{I(\mathcal{C} = \infty)\varphi^{*F}(Z)}{\varpi(\infty, Z)}.$$

This is clearly a function of the observed data. Moreover,

$$E\left\{\frac{I(\mathcal{C} = \infty)\varphi^{*F}(Z)}{\varpi(\infty, Z)}\middle| Z\right\} = \frac{\varphi^{*F}(Z)}{\varpi(\infty, Z)}E\{I(\mathcal{C} = \infty)|Z\} = \varphi^{*F}(Z),$$

where, in order for the equation above to hold, we must make sure we are not dividing 0 by 0; hence the need for assumption (7.31).

Consequently, $\Lambda_\eta^\perp = \mathcal{K}^{-1}(\Lambda^{F\perp})$ can be written as the direct sum of two linear subspaces; namely,

$$\Lambda_\eta^\perp = \frac{I(\mathcal{C} = \infty)\Lambda^{F\perp}}{\varpi(\infty, Z)} \oplus \Lambda_2, \tag{7.33}$$

which is the linear subspace of $\mathcal{H}$ with elements

$$\Lambda_\eta^\perp = \left\{ \frac{I(\mathcal{C} = \infty)\varphi^{*F}(Z)}{\varpi(\infty, Z)} + L_2\{\mathcal{C}, G_{\mathcal{C}}(Z)\} : \varphi^{*F}(Z) \in \Lambda^{F\perp}, \right.$$
$$\left. L_2\{\mathcal{C}, G_{\mathcal{C}}(Z)\} \in \Lambda_2; \quad \text{i.e., } E[L_2\{\mathcal{C}, G_{\mathcal{C}}(Z)\}|Z] = 0 \right\}. \tag{7.34}$$

To complete the proof, we need to derive the linear space Λ_2.

(ii) The space $\Lambda_2 = \mathcal{K}^{-1}(0)$

Because we are assuming that the coarsening variable $\mathcal{C}$ is discrete, we can express any function $h\{\mathcal{C}, G_{\mathcal{R}}(Z)\} \in \mathcal{H}$ as

$$I(\mathcal{C} = \infty)h_\infty(Z) + \sum_{r \neq \infty} I(\mathcal{C} = r)h_r\{G_r(Z)\}, \tag{7.35}$$

where $h_\infty(Z)$ denotes an arbitrary $q \times 1$ function of Z and $h_r\{G_r(Z)\}$ denotes an arbitrary $q \times 1$ function of $G_r(Z)$. The space of functions $L_2\{\mathcal{C}, G_{\mathcal{C}}(Z)\} \in \Lambda_2 \subset \mathcal{H}$ must satisfy

$$E[L_2\{\mathcal{C}, G_{\mathcal{C}}(Z)\}|Z] = 0;$$

that is,

$$E\left[I(\mathcal{C}=\infty)L_{2\infty}(Z) + \sum_{r\neq\infty} I(\mathcal{C}=r)L_{2r}\{G_r(Z)\}\Big|Z\right] = 0,$$

or

$$\varpi(\infty, Z)L_{2\infty}(Z) + \sum_{r\neq\infty} \varpi\{r, G_r(Z)\}L_{2r}\{G_r(Z)\} = 0. \tag{7.36}$$

Consequently, to obtain an arbitrary element of $L_2\{\mathcal{C}, G_{\mathcal{C}}(Z)\} \in \Lambda_2$, we can define any set of q-dimensional measurable functions $L_{2r}\{G_r(Z)\},\ r \neq \infty$, and, for any such set of functions, (7.36) will be satisfied by taking

$$L_{2\infty}(Z) = -\{\varpi(\infty, Z)\}^{-1} \sum_{r\neq\infty} \varpi\{r, G_r(Z)\}L_{2r}\{G_r(Z)\},$$

where, again, assumption (7.31) is needed to guarantee that we are not dividing by zero. Hence, for any $L_{2r}\{G_r(Z)\},\ r \neq \infty$, we can define a typical element of Λ_2 as

$$\frac{I(\mathcal{C}=\infty)}{\varpi(\infty, Z)}\left[\sum_{r\neq\infty} \varpi\{r, G_r(Z)\}L_{2r}\{G_r(Z)\}\right] - \sum_{r\neq\infty} I(\mathcal{C}=r)L_{2r}\{G_r(Z)\}. \tag{7.37}$$

Combining the results from (7.34) and (7.37), we are now able to explicitly define the linear space $\Lambda_\eta^\perp$ to be that consisting of all elements given by (7.32). □

Identifying the space $\Lambda^\perp$ will often guide us in deriving semiparametric estimators. When data are coarsened by design, $\Lambda^\perp = \Lambda_\eta^\perp$.

Remark 9. The representation of $\Lambda_\eta^\perp$ given by(7.33) as a direct sum of two linear spaces will give us insight on how to construct estimating equations whose solution will yield semiparametric RAL estimators for β with coarsened data.

In Chapters 4 and 5, we showed how to use elements of $\Lambda^{F\perp}$ (i.e., the space orthogonal to the full-data nuisance tangent space) to construct estimating equations whose solution resulted in full-data RAL estimators for β. Since the first space in the direct sum (7.33), namely $\frac{I(\mathcal{C}=\infty)\Lambda^{F\perp}}{\varpi(\infty,Z)}$, consists of the inverse probability weighted complete-case elements of $\Lambda^{F\perp}$, this suggests that observed-data estimators for β can be constructed by using inverse probability weighted complete-case (IPWCC) full-data estimating equations. This would lead to what are called IPWCC estimators. We gave a simple example of this in Section 6.4.

The second space, Λ_2 in (7.33) will be referred to as the augmentation space. Estimators for β that include elements of Λ_2 as part of the estimator will be referred to as augmented inverse probability weighted complete-case (AIPWCC) estimators. In Section 6.5, we introduced such an estimator and showed how the augmentation term can help us gain efficiency and, in some cases, leads to estimators with the property of double robustness. □

Therefore, we will formally define the two linear subspaces as follows.

Definition 3. The linear subspace contained in $\mathcal{H}$ consisting of elements

$$\left\{ \frac{I(\mathcal{C}=\infty)\varphi^{*F}(Z)}{\varpi(\infty, Z)}; \text{ for all } \varphi^{*F}(Z) \in \Lambda^{F\perp} \right\},$$

also denoted as $\frac{I(\mathcal{C}=\infty)\Lambda^{F\perp}}{\varpi(\infty,Z)}$, will be defined to be the IPWCC space. □

Definition 4. The linear space $\Lambda_2 \subset \mathcal{H}$ will be defined to be the augmentation space. □

Before continuing to more complicated situations, it will be instructive to see how the geometry we have developed so far will aid us in constructing estimators in several examples when data are missing at random by design.

7.4 Example: Restricted Moment Model with Missing Data by Design

Consider the semiparametric restricted moment model that assumes that

$$E(Y|X) = \mu(X, \beta),$$

where Y is the response variable and X is a vector of covariates. Here, $Z = (Y, X)$ denotes full data. We studied the semiparametric properties of this model in great detail in Sections 4.5 and 4.6, where we also showed, in (4.48), that a typical element of $\Lambda^{F\perp}$ is given as

$$A(X)\{Y - \mu(X, \beta_0)\}.$$

This motivates the generalized estimating equation (GEE), or m-estimator, which is the solution to

$$\sum_{i=1}^{n} A(X_i)\{Y_i - \mu(X_i, \beta)\} = 0 \tag{7.38}$$

using a sample of data $(Y_i, X_i), i = 1, \ldots, n$.

Suppose, by design, we coarsen the data at random. For example, let the vector of covariates X for a single observation be partitioned into two sets of

variables, $X = ({X^{(1)}}^T, {X^{(2)}}^T)^T$, where $X^{(1)}$ are variables that are relatively inexpensive to collect, whereas $X^{(2)}$ are expensive to collect. For example, $X^{(2)}$ may be genetic markers that are expensive to process, whereas $X^{(1)}$ may be descriptive variables such as age, race, sex, etc. In such a case, we might decide to collect the response variable Y and the inexpensive covariates $X^{(1)}$ on everyone in the sample but collect the expensive covariates $X^{(2)}$ only on a subset of individuals. Moreover, we let the probability of collecting $X^{(2)}$ depend on the values of Y and $X^{(1)}$. This might be the case if, say, we want to overrepresent some values of Y and $X^{(1)}$ in the subset where all the data are collected. This is an example of missing data by design. That is, the full data are denoted by $Z_i = (Y_i, X_i^{(1)}, X_i^{(2)}), i = 1, \ldots, n$. Y_i and $X_i^{(1)}$ are observed on everyone, whereas $X_i^{(2)}$ may be missing for some individuals. To implement such a design, we would collect the data $(Y_i, X_i^{(1)})$ for all patients $i = 1, \ldots, n$, as well as blood samples that could be used to obtain the expensive genetic markers. For patient i we then choose, at random, the complete-case binary indicator R_i taking the value 1 or 0 with probability $\pi(Y_i, X_i^{(1)})$ and $1 - \pi(Y_i, X_i^{(1)})$ respectively, where the function $0 < \pi(y, x^{(1)}) \leq 1$ is a known function of the response $Y = y$ and the covariates $X^{(1)} = x^{(1)}$ chosen by the investigator. If $R_i = 1$, then we process the blood sample and obtain the genetic markers $X_i^{(2)}$; otherwise, we let that data be missing.

Since there are only two levels of coarsening in this problem, it is convenient to work with the binary indicator R to denote whether the observation was complete or whether some of the data ($X^{(2)}$ in this case) were missing. The relationship to the notation we have been using is as follows: $R = (0, 1)$, where R is not scripted, is equivalent to the coarsening variable $\mathcal{C} = (1, \infty)$, $G_1(Z) = (Y, X^{(1)})$, $G_\infty(Z) = Z = (Y, X^{(1)}, X^{(2)})$, $P(\mathcal{C} = 1|Z) = \varpi\{1, G_1(Z)\} = 1 - \pi(Y, X^{(1)})$, and $P(\mathcal{C} = \infty|Z) = \varpi\{\infty, G_\infty(Z)\} = \pi(Y, X^{(1)})$.

Note 1. On notation for missingness probabilities
In keeping with much of the notation in the literature, we denote the probability of a complete case by $P(R = 1|Y, X^{(1)})$ as $\pi(Y, X^{(1)})$. This should not be confused with coarsening probabilities, which are denoted as $P(\mathcal{C} = r|Z) = \varpi\{r, G_r(Z)\}$. □

Since the missingness probabilities are known by design for this example, the nuisance tangent space for the observed data is Λ_η, and the space orthogonal to the nuisance tangent space, $\Lambda_\eta^\perp$, derived in (7.32) of Theorem 7.2, is

$$\left\{ \frac{R\varphi^{*F}(Z)}{\pi(Y, X^{(1)})} + L_2\{\mathcal{C}, G_\mathcal{C}(Z)\} : \varphi^{*F}(Z) \in \Lambda^{F\perp}, L_2\{\mathcal{C}, G_\mathcal{C}(Z)\} \in \Lambda_2 \right\}. \quad (7.39)$$

Because $L_{21}\{G_1(Z)\}$ is an arbitrary $q \times 1$ function of $(Y, X^{(1)})$, which we denote by $L(Y, X^{(1)})$, we can use formula (7.37) to show, after some algebra, that any element $L_2\{\mathcal{C}, G_\mathcal{C}(Z)\} \in \Lambda_2$ can be expressed as

$$\left\{\frac{R-\pi(Y,X^{(1)})}{\pi(Y,X^{(1)})}\right\}L(Y,X^{(1)}). \tag{7.40}$$

Since a typical element $\varphi^{*F}(Z) \in \Lambda^{F\perp}$ for the restricted moment model is

$$A(X)\{Y-\mu(X,\beta_0)\},$$

for arbitrary $A(X)$, then by (7.39) and (7.40), a typical element of $\Lambda_\eta^\perp$ is

$$\frac{R[A(X)\{Y-\mu(X,\beta_0)\}]}{\pi(Y,X^{(1)})}+\left\{\frac{R-\pi(Y,X^{(1)})}{\pi(Y,X^{(1)})}\right\}L(Y,X^{(1)})$$

for arbitrary $A(X)$ and $L(Y,X^{(1)})$.

We have shown that identifying elements orthogonal to the nuisance tangent space and using these as estimating functions (i.e., functions of the data and the parameter) may guide us in constructing estimating equations whose solution would yield a consistent, asymptotically normal estimator for β. Therefore, for this problem, we might consider estimating β with a sample of coarsened data

$$(R_i, Y_i, X_i^{(1)}, R_i X_i^{(2)}), \quad i=1,\ldots,n,$$

by using the m-estimator that solves

$$\sum_{i=1}^{n}\left[\frac{R_i}{\pi(Y_i,X_i^{(1)})}A(X_i)\{Y_i-\mu(X_i,\beta)\}\right.$$
$$\left.+\left\{\frac{R_i-\pi(Y_i,X_i^{(1)})}{\pi(Y_i,X_i^{(1)})}\right\}L(Y_i,X_i^{(1)})\right]=0. \tag{7.41}$$

If this estimator is to be consistent, at the least we would need that, at the truth,

$$E\left[\frac{R}{\pi(Y,X^{(1)})}A(X)\{Y-\mu(X,\beta_0)\}+\left\{\frac{R-\pi(Y,X^{(1)})}{\pi(Y,X^{(1)})}\right\}L(Y,X^{(1)})\right]=0.$$

Using the law of iterated conditioning, where we first condition on Y, X, we obtain

$$E\left[\frac{A(X)\{Y-\mu(X,\beta_0)\}}{\pi(Y,X^{(1)})}E(R|Y,X)\right.$$
$$\left.+\left\{\frac{E(R|Y,X)-\pi(Y,X^{(1)})}{\pi(Y,X^{(1)})}\right\}L(Y,X^{(1)})\right]. \tag{7.42}$$

Since

$$E(R|Y,X)=P(R=1|Y,X)=P(R=1|Y,X^{(1)},X^{(2)}),$$

which, by design, equals $\pi(Y,X^{(1)})$, we obtain that (7.42) is equal to

$$E[A(X)\{Y - \mu(X, \beta_0)\} + 0] = 0. \tag{7.43}$$

Also, the usual expansion of m-estimators can be used to derive asymptotic normality. That is,

$$\begin{aligned}
&0 \\
&= \sum_{i=1}^{n} \left[\frac{R_i}{\pi(Y_i, X_i^{(1)})} A(X_i)\{Y_i - \mu(X_i, \hat{\beta}_n)\} + \left\{ \frac{R_i - \pi(Y_i, X_i^{(1)})}{\pi(Y_i, X_i^{(1)})} \right\} L(Y_i, X_i^{(1)}) \right] \\
&= \sum_{i=1}^{n} \left[\frac{R_i}{\pi(Y_i, X_i^{(1)})} A(X_i)\{Y_i - \mu(X_i, \beta_0)\} + \left\{ \frac{R_i - \pi(Y_i, X_i^{(1)})}{\pi(Y_i, X_i^{(1)})} \right\} L(Y_i, X_i^{(1)}) \right] \\
&\quad - \left[\sum_{i=1}^{n} \frac{R_i}{\pi(Y_i, X_i^{(1)})} A(X_i) D(X_i, \beta_n^*) \right] (\hat{\beta}_n - \beta_0),
\end{aligned}$$

where $D(X, \beta) = \partial \mu(X, \beta)/\partial \beta^T$ and β_n^* is an intermediate value between $\hat{\beta}_n$ and β_0. Therefore,

$$\begin{aligned}
n^{1/2}(\hat{\beta}_n - \beta_0) = &\left[n^{-1} \sum_{i=1}^{n} \frac{R_i}{\pi(Y_i, X_i^{(1)})} A(X_i) D(X_i, \beta_n^*) \right]^{-1} \\
&\times n^{-1/2} \sum_{i=1}^{n} \left[\frac{R_i}{\pi(Y_i, X_i^{(1)})} A(X_i)\{Y_i - \mu(X_i, \beta_0)\} \right. \\
&\left. + \left\{ \frac{R_i - \pi(Y_i, X_i^{(1)})}{\pi(Y_i, X_i^{(1)})} \right\} L(Y_i, X_i^{(1)}) \right].
\end{aligned}$$

Under suitable regularity conditions,

$$n^{-1} \sum \left\{ \frac{R_i}{\pi(Y_i, X_i^{(1)})} A(X_i) D(X_i, \beta_n^*) \right\} \xrightarrow{P} E\left\{ \frac{R}{\pi(Y, X^{(1)})} A(X) D(X, \beta_0) \right\}.$$

Using iterated conditioning, where first we condition on Y, X, we obtain

$$E\{A(X) D(X, \beta_0)\}.$$

Consequently,

$$\begin{aligned}
n^{1/2}(\hat{\beta}_n - \beta_0) = &\, n^{-1/2} \sum_{i=1}^{n} [E\{A(X) D(X, \beta_0)\}]^{-1} \\
&\left[\frac{R_i}{\pi(Y_i, X_i^{(1)})} A(X_i)\{Y_i - \mu(X_i, \beta_0)\} + \left\{ \frac{R_i - \pi(Y_i, X_i^{(1)})}{\pi(Y_i, X_i^{(1)})} \right\} L(Y_i, X_i^{(1)}) \right] \\
&+ o_p(1).
\end{aligned} \tag{7.44}$$

Therefore, the i-th influence function for $\hat{\beta}_n$ is

$$[E\{A(X)D(X,\beta_0)\}]^{-1}\left[\frac{R_i}{\pi(Y_i,X_i^{(1)})}A(X_i)\{Y_i-\mu(X_i,\beta_0)\}\right.$$
$$\left.+\left\{\frac{R_i-\pi(Y_i,X_i^{(1)})}{\pi(Y_i,X_i^{(1)})}\right\}L(Y_i,X_i^{(1)})\right],$$

which we demonstrated has mean zero, in (7.42) and (7.43), using iterated conditional expectations.

We note that this influence function is proportional to the element in $\Lambda_\eta^\perp$ that motivated the corresponding m-estimator. Also, because this estimator is asymptotically linear, as shown in (7.44), we immediately deduce that this estimator is asymptotically normal with asymptotic variance being the variance of its influence function. Other than regularity conditions, no assumptions were made on the distribution of (Y,X), beyond that of the restricted moment assumption, to obtain asymptotic normality. Therefore, the resulting estimator is a semiparametric estimator.

Standard methods using a sandwich variance can be used to derive an estimator for the asymptotic variance of $\hat{\beta}_n$, the solution to (7.41). Such a sandwich estimator was derived for the full-data GEE estimator in (4.9) of Section 4.1. We leave the details as an exercise at the end of the chapter.

Hence, for the restricted moment model with missing data that are missing by design, we have derived the space orthogonal to the nuisance tangent space (i.e., $\Lambda_\eta^\perp$) and have constructed an m-estimator with influence function proportional to any element of $\Lambda_\eta^\perp$. Since all influence functions of RAL estimators for β must belong to $\Lambda_\eta^\perp$, this means that any RAL estimator for β must be asymptotically equivalent to one of the estimators given by the solution to (7.41).

Remark 10. The estimator for β, given as the solution to (7.41), is referred to as an augmented inverse probability weighted complete-case (AIPWCC) estimator. If $L(Y,X^{(1)})$ is chosen to be identically equal to zero, then the estimating equation in (7.41) becomes

$$\sum_{i=1}^{n}\frac{R_i}{\pi(Y_i,X_i^{(1)})}A(X_i)\{Y_i-\mu(X_i,\beta)\}=0. \tag{7.45}$$

The solution to (7.45) is referred to as an inverse probability weighted complete-case (IPWCC) estimator. The second term in (7.41), which involves the arbitrary function $L(Y,X^{(1)})$, allows contributions from individuals with missing data into the estimating equation. Properly chosen augmentation will result in an estimator with greater efficiency. □

The choice of the influence function and hence the corresponding class of estimators depends on the arbitrary functions $A(X)$ and $L(Y,X^{(1)})$. With full data, the class of estimating equations is characterized by (7.38). This

requires us to choose the function $A(X)$. In Chapter 4, we proved that the optimal choice for $A(X)$ was $D^T(X)V^{-1}(X)$, where $V(X) = \text{var}\,(Y|X)$, and suggested adaptive strategies for finding locally efficient estimators for β in Section 4.6.

With missing data by design, we also want to find the optimal RAL estimator for β; i.e., the RAL estimator for β with the smallest asymptotic variance. This means that we must derive the functions $A(X)$ and $L(Y, X^{(1)})$, which yields an estimator in (7.41) with the smallest asymptotic variance. Finding the optimal estimator with coarsened data will require special considerations that will be the focus of later chapters. In general, the optimal choice of $A(X)$ with coarsened data is not necessarily the same as it is for full data. These issues will be studied more carefully.

The Logistic Regression Model

We gave an example in Section 7.2 where we argued that with coarsened data it was difficult to obtain estimators for β using likelihood methods. Specifically, we considered the logistic regression model for the probability of response $Y = 1$ as a function of covariates X, where Y denotes a binary response variable. Let us consider the likelihood for such a model if we had missing data by design as described above; that is, where $X = (X^{(1)^T}, X^{(2)^T})^T$ and where we always observe Y and $X^{(1)}$ on everyone in the sample but only observe $X^{(2)}$ on a subset chosen at random with probability $\pi(Y, X^{(1)})$ by design. Also, to allow for an intercept term in the logistic regression model, we define $X^* = (1, X^{(1)^T}, X^{(2)^T})^T$ and $X^{(1*)} = (1, X^{(1)^T})^T$. The density of the full data (Y, X) for this problem can be written as

$$p_{Y,X}(y, x, \beta, \eta_1, \eta_2) = p_{Y|X}(y|x, \beta)p_{X^{(2)}|X^{(1)}}(x^{(2)}|x^{(1)}, \eta_1)p_{X^{(1)}}(x^{(1)}, \eta_2)$$
$$= \left[\frac{\exp\{(\beta_1^T x^{(1*)} + \beta_2^T x^{(2)})y\}}{1 + \exp(\beta_1^T x^{(1*)} + \beta_2^T x^{(2)})}\right] p_{X^{(2)}|X^{(1)}}(x^{(2)}|x^{(1)}, \eta_1)p_{X^{(1)}}(x^{(1)}, \eta_2),$$

where β is partitioned as $\beta = (\beta_1^T, \beta_2^T)^T$, $p_{X^{(2)}|X^{(1)}}(x^{(2)}|x^{(1)}, \eta_1)$ denotes the conditional density of $X^{(2)}$ given $X^{(1)}$, specified through the parameter η_1, and $p_{X^{(1)}}(x^{(1)}, \eta_2)$ denotes the marginal density of $X^{(1)}$, specified through the parameter η_2. Because the parameter of interest β separates from the parameters η_1 and η_2 in the density above, finding the MLE for β with full data only involves maximizing the part of the likelihood above involving β and is easily implemented in most software packages.

In contrast, the density of the observed data $(R, Y, X^{(1)}, RX^{(2)})$ is given by

$$
\begin{aligned}
&\{p_{Y|X}(y|x,\beta)\}^r\{p_{X^{(2)}|X^{(1)}}(x^{(2)}|x^{(1)},\eta_1)\}^r \\
&\times\left\{\int p_{Y|X}(y|x,\beta)p_{X^{(2)}|X^{(1)}}(x^{(2)}|x^{(1)},\eta_1)d\nu_{X^{(2)}}(x^{(2)})\right\}^{1-r} p_{X^{(1)}}(x^{(1)},\eta_2) \\
&=\left[\frac{\exp\{(\beta_1^T x^{(1*)}+\beta_2^T x^{(2)})y\}}{1+\exp(\beta_1^T x^{(1*)}+\beta_2^T x^{(2)})}\right]^r \{p_{X^{(2)}|X^{(1)}}(x^{(2)}|x^{(1)},\eta_1)\}^r \qquad (7.46) \\
&\times\left\{\int\left[\frac{\exp\{(\beta_1^T x^{(1*)}+\beta_2^T x^{(2)})y\}}{1+\exp(\beta_1^T x^{(1*)}+\beta_2^T x^{(2)})}\right]\right. \\
&\qquad\left. p_{X^{(2)}|X^{(1)}}(x^{(2)}|x^{(1)},\eta_1)d\nu_{X^{(2)}}(x^{(2)})\right\}^{1-r} \qquad (7.47) \\
&\times p_{X^{(1)}}(x^{(1)},\eta_2).
\end{aligned}
$$

Because the parameters β and η_1 do not separate in the density above, deriving the MLE for β involves maximizing, as a function of β and η_1, the product (over $i=1,\ldots,n$) of terms (7.46) $\times$ (7.47). Even if we were willing to make simplifying parametric assumptions about the conditional distribution of $X^{(2)}$ given $X^{(1)}$ in terms of a finite number of parameters η_1, this would be a complicated maximization, but if we wanted to be semiparametric (i.e., put no restrictions on the conditional distribution of $X^{(2)}$ given $X^{(1)}$), then this problem would be impossible as it would suffer from the curse of dimensionality. Notice that in the likelihood formulation above, nowhere do the probabilities $\pi(Y,X^{(1)})$ come into play, even though they are known to us by design.

Since the logistic regression model is just a simple example of a restricted moment model, estimators for the parameter β for the semiparametric model, which puts no restrictions on the joint distribution of $(X^{(1)},X^{(2)})$, can be found easily by solving the estimating equation (7.41), where $\mu(X_i,\beta)=\exp(\beta^T X_i^*)/\{1+\exp(\beta^T X_i^*)\}$ and for some choice of $A(X)$ and $L(Y,X^{(1)})$.

With no missing data, we showed in (4.65) that the optimal choice for $A(X)$ is X^*. Consequently, one easy way of obtaining an estimator for β is by solving (7.41) using $A(X_i)=X_i^*$ and $L(Y_i,X_i^{(1)})=0$, leading to the estimating equation

$$
\sum_{i=1}^n \frac{R_i}{\pi(Y_i,X_i^{(1)})}X_i^*\left\{Y_i-\frac{\exp(\beta^T X_i^*)}{1+\exp(\beta^T X_i^*)}\right\}=0. \qquad (7.48)
$$

This estimator is an inverse probability weighted complete case (IPWCC) estimator for β. Although this estimator is a consistent, asymptotically normal semiparametric estimator for β, it is by no means efficient. Since this estimator only uses the complete cases (i.e., the data from individual $i:R_i=1$), it is intuitively clear that additional efficiency can be gained by using the data from individuals $i:R_i=0$, where only some of the data are missing. Therefore, it would be preferable to use an AIPWCC estimator given by (7.41); namely,

$$\sum_{i=1}^{n}\left[\frac{R_i}{\pi(Y_i,X_i^{(1)})}X_i^*\left\{Y_i-\frac{\exp(\beta^TX_i^*)}{1+\exp(\beta^TX_i^*)}\right\}\right.$$
$$\left.+\left\{\frac{R_i-\pi(Y_i,X_i^{(1)})}{\pi(Y_i,X_i^{(1)})}\right\}L(Y_i,X_i^{(1)})\right]=0, \tag{7.49}$$

with some properly chosen $L(Y,X^{(1)})$.

This illustrates the usefulness of understanding the semiparametric theory for missing and coarsened data. Of course, the choice of $A(X)$ and $L(Y,X^{(1)})$ that will result in efficient estimators for β still needs to be addressed.

7.5 Recap and Review of Notation

Before continuing, we believe it is worthwhile to review some of the basic ideas and notation that have been developed thus far.

Full data

- Full data are denoted by Z with density from a semiparametric model $p_Z(z,\beta,\eta)$, where β denotes the q-dimensional parameter of interest and η denotes the infinite-dimensional nuisance parameter.
- $\mathcal{H}^F$ denotes the full-data Hilbert space defined as all mean-zero, q-dimensional measurable functions of Z with finite variance equipped with the covariance inner product.
- Λ^F is the full-data nuisance tangent space spanned by the full-data nuisance score vectors for parametric submodels and their mean-square closure.
- $\Lambda^{F\perp}=\{$set of elements $\varphi^{*F}(Z)$ that are orthogonal to $\Lambda^F\}$. This is the space on which influence functions lie. Identifying this space helps motivate full-data m-estimators.

Observed (coarsened) data

- Coarsened data are denoted by $\{\mathcal{C},G_{\mathcal{C}}(Z)\}$, where the coarsening variable $\mathcal{C}$ is a discrete random variable taking on values $1,\ldots,\ell$ and ∞. When $\mathcal{C}=r,r=1,\ldots,\ell$, then we observe the many-to-one transformation $G_r(Z)$. $\mathcal{C}=\infty$ is reserved to denote complete data; i.e., $G_\infty(Z)=Z$.
- We distinguish among three types of coarsening mechanisms:
  - *Coarsening completely at random* (CCAR): The coarsening probabilities do not depend on the data.
  - *Coarsening at random* (CAR): The coarsening probabilities only depend on the data as a function of the observed data.
  - *Noncoarsening at random* (NCAR): The coarsening probabilities depend on the unobserved part of the data.

- When coarsening is CAR, we denote the coarsening probabilities by

$$P(\mathcal{C} = r|Z) = \varpi\{r, G_r(Z)\}.$$

- A key assumption is that there is a positive probability of observing complete data; that is,

$$P(\mathcal{C} = \infty|Z = z) = \varpi(\infty, Z) > \epsilon > 0 \text{ for all } z$$

  in the support of Z.
- $\mathcal{H}$ denotes the observed-data Hilbert space of q-dimensional, mean-zero, finite-variance, measurable functions of $\{\mathcal{C}, G_{\mathcal{C}}(Z)\}$ equipped with the covariance inner product.
- Because $\mathcal{C}$ takes on a finite set of values, a typical function $h\{\mathcal{C}, G_{\mathcal{C}}(Z)\}$ can be written as

$$h\{\mathcal{C}, G_{\mathcal{C}}(Z)\} = I(\mathcal{C} = \infty)h_\infty(Z) + \sum_{r \neq \infty} I(\mathcal{C} = r)h_r\{G_r(Z)\}.$$

- The observed-data nuisance tangent space

$$\Lambda = \Lambda_\psi \oplus \Lambda_\eta, \Lambda_\psi \perp \Lambda_\eta,$$

  where Λ_ψ is spanned by the score vector with respect to the parameter ψ used to describe the coarsening process and Λ_η is spanned by the observed-data nuisance score vectors for parametric submodels and their mean-square closures. Specifically,

$$\Lambda_\eta = \left\{ E\{\alpha^F(Z)|\mathcal{C}, G_{\mathcal{C}}(Z)\} : \alpha^F(Z) \in \Lambda^F \right\} = E\{\Lambda^F|\mathcal{C}, G_{\mathcal{C}}(Z)\}.$$

- In this chapter, we did not consider models for the coarsening probabilities; rather, we assumed they are known by design. Therefore, we didn't need to consider the space Λ_ψ, in which case the observed-data nuisance tangent space $\Lambda = \Lambda_\eta$.
- Observed-data estimating equations, when coarsening is by design, are motivated by considering elements in the space $\Lambda_\eta^\perp$, where

$$\Lambda_\eta^\perp = \left\{ \frac{I(\mathcal{C} = \infty)\Lambda^{F\perp}}{\varpi(\infty, Z)} \oplus \Lambda_2 \right\}$$

  and

$$\Lambda_2 = \left\{ L_2\{\mathcal{C}, G_{\mathcal{C}}(Z)\} : E[L_2\{\mathcal{C}, G_{\mathcal{C}}(Z)\}|Z] = 0 \right\}.$$

- The two linear spaces that make up $\Lambda_\eta^\perp$ are the IPWCC space

$$\frac{I(\mathcal{C}=\infty)\Lambda^{F\perp}}{\varpi(\infty,Z)}$$

and the augmentation space

$$\Lambda_2.$$

- In order to construct a typical element of Λ_2, for each $r \neq \infty$, choose an arbitrary function $L_{2r}\{G_r(Z)\}$. Then

$$\frac{I(\mathcal{C}=\infty)}{\varpi(\infty,Z)}\left[\sum_{r\neq\infty}\varpi\{r,G_r(Z)\}L_{2r}\{G_r(Z)\}\right]-\sum_{r\neq\infty}I(\mathcal{C}=r)L_{2r}\{G_r(Z)\}$$

is an element of Λ_2.

7.6 Exercises for Chapter 7

Returning to Example 1, introduced in Section 7.1, recall that two blood samples were taken from each individual in a study where one of the objectives was to assess the variability within and between persons in the concentration of some biological marker. As part of the design of this study, a subset of individuals was chosen at random with probability ϖ. These individuals had their two blood samples combined and the concentration was obtained on the pooled blood, whereas for the remaining individuals in the study, the concentration was obtained separately for each of the two blood samples. In Section 7.2, we introduced a bivariate normal model for the full data given by (7.11) in terms of parameters $(\mu_\alpha, \sigma_\alpha^2, \sigma_e^2)$, and in equation (7.12) we derived the likelihood of the observed coarsened data. The first three exercises below relate to this example.

1. Let us first consider only the full data for the time being.
   a) What is the likelihood of the full data $(X_{i1}, X_{i2}), i = 1, \ldots, n$?
   b) Find the MLE for the parameters $(\mu_\alpha, \sigma_\alpha^2, \sigma_e^2)$.
   c) Derive the influence function of the full-data MLE.
2. Return to the coarsened data whose likelihood for the parameters $(\mu_\alpha, \sigma_\alpha^2, \sigma_e^2)$ is given by (7.12).
   a) Derive the observed data MLE using the coarsened data.
   b) What is the relative efficiency between the coarsened-data MLE and the full-data MLE? Derive this separately for μ_α, σ_α^2, and σ_e^2.
3. We now consider AIPWCC estimators for this problem.
   a) Derive the augmentation space Λ_2.

b) Using the full-data influence function that was derived in 1(c) above (or, equivalently, using the full-data score vector), write out a set of AIPWCC estimating equations that can be used to obtain observed-data estimators for $(\mu_\alpha, \sigma^2_\alpha, \sigma^2_e)$.

4. Derive an estimator for the asymptotic variance of $\hat{\beta}_n$, the AIPWCC estimator for β for the restricted moment given by the solution to (7.41) where data were missing by design.

8

The Nuisance Tangent Space and Its Orthogonal Complement

8.1 Models for Coarsening and Missingness

Two Levels of Missingness

In the previous chapter, we gave an example where the missingness (coarsening) mechanism was known to us by design. For most missing-data problems, this is not the case, and we must consider models (either parametric or semiparametric) for the coarsening probabilities. For example, suppose the full-data Z is a random vector that can be partitioned as $Z = (Z_1^T, Z_2^T)^T$, where Z_1 is always observed but Z_2 may be missing on a subset of individuals. This scenario occurs frequently in practice where, say, one of the variables being collected is missing on some individuals or where a set of variables that are collected at the same time may be missing on some individuals. In this example, there are two levels of missingness; either all the data are available on an individual, which is denoted by letting the complete-case indicator $R = 1$ (unscripted), or only the data Z_1 are available, which is denoted by letting $R = 0$. Using the coarsening notation, this would correspond to $\mathcal{C} = \infty$ or $\mathcal{C} = 1$, respectively. If we assume that missingness is MAR, then this would imply that $P(R = 1|Z) = \{1 - P(R = 0|Z)\} = \{1 - P(R = 0|Z_1)\} = \pi(Z_1)$ and $P(R = 0|Z_1) = 1 - \pi(Z_1)$. We remind the reader that using the coarsening notation, the probability $\pi(Z_1) = \varpi\{\infty, G_\infty(Z)\}$ and $1 - \pi(Z_1) = \varpi\{1, G_1(Z)\}$, where $G_1(Z) = Z_1$ and $G_\infty(Z) = Z$. If the missingness was not by design, then the probability of a complete case $\pi(Z_1)$, as a function of Z_1, is unknown to us and must be estimated from the data. This is generally accomplished by positing a model in terms of parameters ψ. Since, in this simplest example of missing data, the complete-case indicator is a binary variable, a natural model is the logistic regression model, where

$$\pi(Z_1, \psi) = \frac{\exp(\psi_0 + \psi_1^T Z_1)}{1 + \exp(\psi_0 + \psi_1^T Z_1)}, \tag{8.1}$$

and the parameter $\psi = (\psi_0, \psi_1^T)^T$ needs to be estimated from the observed data. Although this illustration assumed a logistic regression model that was linear in Z_1, we could easily have considered more complex models where we include higher-order terms, interactions, regression splines, or whatever else the data analyst deems appropriate.

Monotone and Nonmonotone Coarsening for more than Two Levels

When there are more than two levels of missingness or coarsening of the data, we distinguish between monotone and nonmonotone coarsening.

A form of missingness that often occurs in practice is monotone missingness. Because of its importance, we now describe monotone missingness, or more generally monotone coarsening, in more detail and discuss methods for developing models for such monotone missingness mechanisms.

For some problems, we can order the coarsening variable $\mathcal{C}$ in such a way that the coarsened data $G_r(Z)$ when $\mathcal{C} = r$ is a coarsened version of $G_{r'}(Z)$ for all $r' > r$. In such a case, $G_r(Z)$ is a many-to-one function of $G_{r+1}(Z)$; that is,

$$G_r(z) = f_r\{G_{r+1}(z)\},$$

where $f_r(\cdot)$ denotes a many-to-one function that depends on r. In other words, $G_1(Z)$ is the most coarsened data, $G_2(Z)$ less so, and $G_\infty(Z) = Z$ is not coarsened at all. For example, with longitudinal data, suppose we intend to measure an individual at l different time points so that $Z = (Y_1, \ldots, Y_l)$, where Y_j denotes the measurement at the j-th time point, $j = 1, \ldots, l$. For such longitudinal studies, it is not uncommon for some individuals to drop out during the course of the study, in which case we would observe the data up to the time they dropped out and all subsequent measurements would be missing. This pattern of missingness is monotone and can be described by

$$\begin{array}{cc} \underline{r} & \underline{G_r(Z)} \\ 1 & (Y_1) \\ 2 & (Y_1, Y_2) \\ \vdots & \\ l-1 & (Y_1, \ldots, Y_{l-1}) \\ \infty & (Y_1, \ldots, Y_l) \end{array}$$

When data are CAR, we consider models for the coarsening probabilities, which, in general, are denoted by

$$P(\mathcal{C} = r | Z = z, \psi) = \varpi\{r, G_r(z), \psi\}$$

in terms of the unknown parameters ψ. However, with monotone coarsening, it is more convenient to consider models for the discrete hazard function, defined as

$$\begin{aligned}\lambda_r\{G_r(Z)\} &= P(\mathcal{C} = r|\mathcal{C} \geq r, Z),\ r \neq \infty \\ &= 1,\ r = \infty. \end{aligned} \tag{8.2}$$

That $\lambda_r(\cdot)$ is a function of $G_r(Z)$ follows by noting that the right-hand side of (8.2) equals

$$\frac{P(\mathcal{C} = r|Z)}{P(\mathcal{C} \geq r|Z)} = \frac{\varpi\{r, G_r(Z)\}}{1 - \sum_{r' \leq r-1} \varpi\{r', G_{r'}(Z)\}} \tag{8.3}$$

and by the definition of monotone coarsening, where $G_{r'}(Z)$ is a function of $G_r(Z)$ for all $r' < r$. We also define

$$K_r\{G_r(Z)\} = P(\mathcal{C} > r|Z) = \prod_{r'=1}^{r} [1 - \lambda_{r'}\{G_{r'}(Z)\}],\ r \neq \infty. \tag{8.4}$$

Consequently, we can equivalently express the coarsening probabilities in terms of the discrete hazard functions; namely,

$$\begin{aligned}\varpi\{r, G_r(Z)\} = K_{r-1}\{G_{r-1}(Z)\}\lambda_r\{G_r(Z)\} \text{ for } r > 1 \\ \text{and } \lambda_1\{G_1(Z)\} \text{ for } r = 1. \end{aligned} \tag{8.5}$$

Equations (8.3), (8.4), and (8.5) demonstrate that there is a one-to-one relationship between coarsening probabilities $\varpi\{r, G_r(Z)\}$ and discrete hazard functions $\lambda_r\{G_r(Z)\}$. Using discrete hazards, the probability of a complete case (i.e., $\mathcal{C} = \infty$) is given by

$$\varpi(\infty, Z) = \prod_{r \neq \infty} \Big[1 - \lambda_r\{G_r(Z)\}\Big]. \tag{8.6}$$

The use of discrete hazards provides a natural way of thinking about monotone coarsening. For example, suppose we were asked to design a longitudinal study with monotone missingness. We can proceed as follows. First, we would collect $G_1(Z) = Y_1$. Then, with probability $\lambda_1\{G_1(Z)\}$ (that is, with probability depending on Y_1), we would stop collecting additional data. However, with probability $1 - \lambda_1\{G_1(Z)\}$, we would collect Y_2, in which case we now have $G_2(Z) = (Y_1, Y_2)$. If we collected (Y_1, Y_2), then with probability $\lambda_2\{G_2(Z)\}$ we would stop collecting additional data, but with probability $1 - \lambda_2\{G_2(Z)\}$ we would collect Y_3, in which case we would have collected $G_3(Z) = (Y_1, Y_2, Y_3)$. We continue in this fashion, either stopping at stage r' after collecting $G_{r'}(Z) = (Y_1, \ldots, Y_{r'})$ or continuing with probability $\lambda_{r'}\{G_{r'}(Z)\}$ or $1 - \lambda_{r'}\{G_{r'}(Z)\}$, respectively. When monotone missingness is viewed in this fashion, it is clear that, conditional on having reached stage r', there are two choices: either stop or continue to the next stage with conditional probability $\lambda_{r'}\{G_{r'}(Z)\}$ or $1 - \lambda_{r'}\{G_{r'}(Z)\}$. Therefore, when we build models for the coarsening probabilities of monotone coarsened data, it is natural to consider individual models for each of the discrete hazards. Because

of the binary choice made at each stage, logistic regression models for the discrete hazards are often used. For example, for the longitudinal data given above, we may consider a model where it is assumed that

$$\lambda_r\{G_r(Z)\} = \frac{\exp(\psi_{0r} + \psi_{1r}Y_1 + \ldots + \psi_{rr}Y_r)}{1 + \exp(\psi_{0r} + \psi_{1r}Y_1 + \ldots + \psi_{rr}Y_r)},\ r \neq \infty. \tag{8.7}$$

Missing or coarsened data can also come about in a manner that is nonmonotone. For the longitudinal data example given above, suppose patients didn't necessarily drop out of the study but rather missed visits from time to time. In such a case, some of the longitudinal data might be missing but not necessarily in a monotone fashion. In the worst-case scenario, any of the $2^l - 1$ combinations of $(Y_1, \ldots, Y_l)$ might be missing for different patients in the study. Building coherent models for the missingness probabilities for such nonmonotone missing data, even under the assumption that missingness is MAR, is challenging. There have been some suggestions for developing nonmonotone missingness models given by Robins and Gill (1997) using what they call *randomized monotone missingness* (RMM) models. Because of the complexity of nonmonotone missingness models, we will not discuss such models specifically in this book. In what follows, we will develop the semiparametric theory assuming that coherent missingness or coarsening models were used. Specific examples with two levels of missingness or monotone missingness will be used to illustrate the results.

8.2 Estimating the Parameters in the Coarsening Model

Models for the coarsening probabilities are described through the parameter ψ. Specifically, it is assumed that $P(\mathcal{C} = r|Z = z, \psi) = \varpi\{r, G_r(z), \psi\}$, where ψ is often assumed to be a finite-dimensional parameter. Estimates for the parameter ψ can be obtained using maximum likelihood. We remind the reader that because of the factorization of the observed-data likelihood given by (7.6), the maximum likelihood estimator $\hat{\psi}_n$ for ψ is obtained by maximizing

$$\prod_{i=1}^{n} \varpi\{\mathcal{C}_i, G_{\mathcal{C}_i}(Z_i), \psi\}. \tag{8.8}$$

MLE for ψ with Two Levels of Missingness

With two levels of missingness, the likelihood (8.8) simplifies to

$$\prod_{i=1}^{n} \{\pi(Z_{1i}, \psi)\}^{R_i} \{1 - \pi(Z_{1i}, \psi)\}^{1-R_i}. \tag{8.9}$$

So, for example, if we entertained the logistic regression model (8.1), then the maximum likelihood estimator for $(\psi_0, \psi_1^T)^T$ would be obtained by maximizing (8.9) or, more specifically, by maximizing

$$\prod_{i=1}^{n} \left[\frac{\exp\{(\psi_0 + \psi_1^T Z_{1i})R_i\}}{1 + \exp(\psi_0 + \psi_1^T Z_{1i})} \right]. \tag{8.10}$$

This can be easily implemented in most available statistical software packages.

MLE for ψ with Monotone Coarsening

Because monotone missingness (or monotone coarsening) is prevalent in many studies, we now consider how to estimate the parameter ψ in this special case. As indicated in Section 8.1, it is more convenient to work with the discrete hazard function to describe monotone coarsening. The discrete hazard, denoted by $\lambda_r\{G_r(Z)\}$, was defined by (8.2). Therefore, it is natural to consider models for the discrete hazard in terms of parameters ψ, which we denote by $\lambda_r\{G_r(Z), \psi\}$. An example of such a model for monotone missing longitudinal data was given by (8.7). We also showed in Section 8.1 that there is a one-to-one relationship between the coarsening probabilities $\varpi\{r, G_r(Z)\}$ and the discrete hazard functions. Using (8.5), we see that the coarsening probability can be deduced through the discrete hazards leading to the model

$$\begin{aligned} \varpi\{r, G_r(Z), \psi\} &= \lambda_1\{G_1(Z), \psi\} \text{ for } r = 1, \\ \varpi\{r, G_r(Z), \psi\} &= \prod_{r'=1}^{r-1} [1 - \lambda_{r'}\{G_{r'}(Z), \psi\}] \lambda_r\{G_r(Z), \psi\} \text{ for } r > 1. \end{aligned} \tag{8.11}$$

Substituting the right-hand side of (8.11) for $\varpi(\cdot, \psi)$ into (8.8) and rearranging terms, we obtain that the likelihood for monotone coarsening can be expressed as

$$\prod_{r \neq \infty} \prod_{i:\mathcal{C}_i \geq r} \left[\lambda_r\{G_r(Z_i), \psi\} \right]^{I(\mathcal{C}_i = r)} \left[1 - \lambda_r\{G_r(Z_i), \psi\} \right]^{I(\mathcal{C}_i > r)}. \tag{8.12}$$

So, for example, if we consider the logistic regression models used to model the discrete hazards for the monotone missing longitudinal data given by (8.7), then the likelihood is given by

$$\prod_{r=1}^{l-1} \prod_{i:\mathcal{C}_i \geq r} \frac{\exp(\psi_{0r} + \psi_{1r} Y_{1i} + \ldots + \psi_{rr} Y_{ri}) I(\mathcal{C}_i = r)}{1 + \exp(\psi_{0r} + \psi_{1r} Y_{1i} + \ldots + \psi_{rr} Y_{ri})}. \tag{8.13}$$

Because the likelihood in (8.13) factors into a product of $l-1$ logistic regression likelihoods, standard logistic regression software can be used to maximize (8.13).

8.3 The Nuisance Tangent Space when Coarsening Probabilities Are Modeled

Our ultimate goal is to derive semiparametric estimators for the parameter β of a semiparametric model when the data are coarsened at random. As always, the key to deriving such estimators is to find elements orthogonal to the nuisance tangent space that in turn can be used to guide us in constructing estimating equations. Toward that end, we now return to the problem of finding the nuisance tangent space for semiparametric models with coarsened data when the coarsening probabilities are modeled using additional parameters ψ. We described in the previous sections how such coarsening probability models can be developed and estimated.

Therefore, as a starting point, we will assume that such a model for the coarsening probabilities has already been developed as a function of unknown parameters ψ and is denoted by $P(\mathcal{C} = r|Z) = \varpi\{r, G_r(Z), \psi\}$. When the observed data are CAR, we showed in (7.6) that the likelihood can be factored as

$$\varpi(r, g_r, \psi) \int_{\{z:G_r(z)=g_r\}} p_Z(z, \beta, \eta) d\nu_Z(z), \tag{8.14}$$

where the parameter ψ is finite-dimensional, say with dimension s.

As shown in (7.18), the observed-data nuisance tangent space can be written as a direct sum of two linear subspaces, namely

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

where Λ_ψ is the space associated with the coarsening model parameter ψ and Λ_η is the space associated with the infinite-dimensional nuisance parameter η. In Chapter 7, we derived the space Λ_η and its orthogonal complement. We now consider the space Λ_ψ and some of its properties. Because the space Λ_ψ will play an important role in deriving RAL estimators for β with coarsened data, when the coarsening probabilities are not known and must be modeled, we will denote this space as the coarsening model tangent space and give a formal definition as follows.

Definition 1. The space Λ_ψ, which we denote as the coarsening model tangent space, is defined as the linear subspace, within $\mathcal{H}$, spanned by the score vector with respect to ψ. That is,

$$\Lambda_\psi = \left[B^{q\times s} S_\psi^{s\times 1}\{\mathcal{C}, G_\mathcal{C}(Z), \psi_0\} \text{ for all } B^{q\times s} \right],$$

where

$$S_\psi^{s\times 1} = \frac{\partial \log \varpi\{\mathcal{C}, G_\mathcal{C}(Z), \psi_0\}}{\partial \psi}, \tag{8.15}$$

and ψ_0 denotes the true value of ψ. □

One of the properties of the space Λ_ψ is that it is contained in the augmentation space Λ_2, as we now prove.

Theorem 8.1. $\Lambda_\psi \subset \Lambda_2$

Proof. Since $\varpi\{r, G_r(z), \psi\} = P(\mathcal{C} = r|Z = z, \psi)$ is a conditional density, then this implies that

$$\sum_r \varpi\{r, G_r(z), \psi\} = 1 \quad \text{for all} \quad \psi, z.$$

Hence, for a fixed "z,"

$$\frac{\partial}{\partial \psi} \sum_r \varpi\{r, G_r(z), \psi\} = 0.$$

Taking the partial derivative inside the sum, dividing and multiplying by $\varpi\{r, G_r(z), \psi\}$, and setting $\psi = \psi_0$ yields

$$\sum_r S_\psi\{r, G_r(z), \psi_0\}\varpi\{r, G_r(z), \psi_0\} = 0 \quad \text{for all} \quad z,$$

or

$$E\left[S_\psi\{\mathcal{C}, G_\mathcal{C}(Z), \psi_0\}|Z\right] = 0.$$

Hence

$$E\left[\underbrace{B^{q\times s} S_\psi\{\mathcal{C}, G_\mathcal{C}(Z), \psi_0\}}_{\text{typical element of } \Lambda_\psi}|Z\right] = 0.$$

Consequently, if $h\{\mathcal{C}, G_\mathcal{C}(Z)\} \in \Lambda_\psi$, then $E[h\{\mathcal{C}, G_\mathcal{C}(Z)\}|Z] = 0$. This implies that $\Lambda_\psi \subset \Lambda_2$. □

Because the parameter ψ and the parameter η separate out in the likelihood (7.6), this should imply that the corresponding spaces Λ_ψ and Λ_η are orthogonal. We now prove this property more formally.

Theorem 8.2. $\Lambda_\psi \perp \Lambda_\eta$

Proof. Recall that the space Λ_η is given by

$$\Lambda_\eta = \left\{E\left\{\alpha^F(Z)|\mathcal{C}, G_\mathcal{C}(Z)\right\} \text{ for all } \alpha^F(Z) \in \Lambda^F\right\}.$$

We first demonstrate that $\Lambda_\eta \perp \Lambda_2$.

Choose an arbitrary element $h\{\mathcal{C}, G_\mathcal{C}(Z)\} \in \Lambda_2$ (i.e., $E(h|Z) = 0$) and an arbitrary element of Λ_η, say $E\{\alpha^F(Z)|\mathcal{C}, G_\mathcal{R}(Z)\}$, for some $\alpha^F(Z) \in \Lambda^F$. The inner product of these two elements is

$$
\begin{aligned}
& E\left[h^T\{\mathcal{C}, G_{\mathcal{C}}(Z)\} E\left\{\alpha^F(Z)|\mathcal{C}, G_{\mathcal{C}}(Z)\right\}\right] \\
& = E\left(E\left[h^T\{\mathcal{C}, G_{\mathcal{C}}(Z)\}\alpha^F(Z)|\mathcal{C}, G_{\mathcal{C}}(Z)\right]\right) \\
& = E\left[h^T\{\mathcal{C}, G_{\mathcal{C}}(Z)\}\alpha^F(Z)\right] \\
& = E\left(E\left[h^T\{\mathcal{C}, G_{\mathcal{C}}(Z)\}\alpha^F(Z)|Z\right]\right) \\
& = E\left(\underbrace{E\left[h^T\{\mathcal{C}, G_{\mathcal{C}}(Z)\}|Z\right]}_{\begin{array}{c}\| \\ 0 \text{ since } h \in \Lambda_2\end{array}} \alpha^F(Z)\right) \\
& = 0.
\end{aligned}
$$

Since Λ_ψ is contained in Λ_2, then this implies that Λ_ψ is orthogonal to Λ_η. □

We are now in a position to derive the space orthogonal to the nuisance tangent space.

8.4 The Space Orthogonal to the Nuisance Tangent Space

Since influence functions of RAL estimators for β belong to the space orthogonal to the nuisance tangent space, it is important to derive the space $\Lambda^\perp$, where $\Lambda = \Lambda_\psi \oplus \Lambda_\eta$ and $\Lambda_\psi \perp \Lambda_\eta$.

Because the nuisance tangent space Λ is the direct sum of two orthogonal spaces, we can show that the orthogonal complement

$$\Lambda^\perp = \Pi(\Lambda_\eta^\perp|\Lambda_\psi^\perp) = \Pi(\Lambda_\psi^\perp|\Lambda_\eta^\perp). \tag{8.16}$$

(We leave this as an exercise for the reader.) Using the first equality above, a typical element of $\Lambda^\perp$ can be found by taking an arbitrary element $h \in \Lambda_\eta^\perp$ and computing

$$h - \Pi(h|\Lambda_\psi) = \Pi(h|\Lambda_\psi^\perp).$$

In Chapter 7, we showed how to find elements orthogonal to Λ_η. In fact, in Theorem 7.2, we showed the important result that $\Lambda_\eta^\perp$ can be written as the direct sum of the IPWCC space and the augmentation space. That is,

$$\Lambda_\eta^\perp = \frac{I(\mathcal{C}=\infty)\Lambda^{F\perp}}{\varpi(\infty, Z)} \oplus \Lambda_2.$$

Specifically, a typical element of $\Lambda_\eta^\perp$ is given by formula (7.34); namely,

$$\left\{\frac{I(\mathcal{C}=\infty)\varphi^{*F}(Z)}{\varpi(\infty, Z, \psi_0)} + L_2\{\mathcal{C}, G_{\mathcal{C}}(Z)\} : \varphi^{*F}(Z) \in \Lambda^{F\perp} \right.$$
$$\left. \text{and } L_2\{\mathcal{C}, G_{\mathcal{C}}(Z)\} \in \Lambda_2\right\}.$$

Therefore, a typical element of $\Lambda^{\perp}$ is given by

$$\left\{ \frac{I(\mathcal{C}=\infty)\varphi^{*F}(Z)}{\varpi(\infty,Z,\psi_0)} + L_2\{\mathcal{C},G_{\mathcal{C}}(Z)\} \right.$$
$$- \Pi\left(\left[\frac{I(\mathcal{C}=\infty)\varphi^{*F}(Z)}{\varpi(\infty,Z,\psi_0)} + L_2\{\mathcal{C},G_{\mathcal{C}}(Z)\} \right] \middle| \Lambda_{\psi} \right)$$
$$\left. \text{for } \varphi^{*F}(Z) \in \Lambda^{F\perp} \text{ and } L_2\{\mathcal{C},G_{\mathcal{C}}(Z)\} \in \Lambda_2 \right\}. \tag{8.17}$$

Before discussing how these results can be used to derive RAL estimators for β when data are CAR and when the coarsening probabilities need to be modeled and estimated, which will be deferred to the next chapter, we close this chapter by defining the space of observed-data influence functions of RAL observed-data estimators for β.

8.5 Observed-Data Influence Functions

Because of condition (i) of Theorem 4.2, observed-data influence functions $\varphi\{\mathcal{C},G_{\mathcal{C}}(Z)\}$ not only must belong to the space $\Lambda^{\perp}$, but must satisfy

$$E\left[\varphi\{\mathcal{C},G_{\mathcal{C}}(Z)\}S_{\beta}^{T}\{\mathcal{C},G_{\mathcal{C}}(Z)\}\right] = I^{q\times q}, \tag{8.18}$$

where $S_{\beta}\{\mathcal{C},G_{\mathcal{C}}(Z)\}$ is the observed-data score vector with respect to β. For completeness, we will now define the space of observed-data influence functions.

Theorem 8.3. When data are coarsened at random (CAR) with coarsening probabilities $P(\mathcal{C}=r|Z)=\varpi\{r,G_r(Z),\psi\}$, where Λ_{ψ} is the space spanned by the score vector $S_{\psi}\{\mathcal{C},G_{\mathcal{C}}(Z)\}$ (i.e., the coarsening model tangent space), then the space of observed-data influence functions, also denoted by (IF), is the linear variety contained in $\mathcal{H}$, which consists of elements

$$\varphi\{\mathcal{C},G_{\mathcal{C}}(Z)\} = \left\{ \left[\frac{I(\mathcal{C}=\infty)\varphi^{F}(Z)}{\varpi(\infty,Z,\psi_0)} + L_2\{\mathcal{C},G_{\mathcal{C}}(Z)\} \right] - \Pi\{[\cdot]|\Lambda_{\psi}\}, \right. \tag{8.19}$$
$$\left. \text{where } \varphi^{F}(Z) \text{ is a full-data influence function and } L_2\{\mathcal{C},G_{\mathcal{C}}(Z)\} \in \Lambda_2 \right\}.$$

Proof. We first note that we can use the exact same arguments as used in lemmas 7.1 and 7.2 to show that the observed-data score vector with respect to β is

$$S_{\beta}\{\mathcal{C},G_{\mathcal{C}}(Z)\} = E\left\{S_{\beta}^{F}(Z)|\mathcal{C},G_{\mathcal{C}}(Z)\right\},$$

where $S_{\beta}^{F}(Z)$ is the full-data score vector with respect to β,

$$S_{\beta}^{F}(z) = \frac{\partial \log p_Z(z,\beta_0,\eta_0)}{\partial\beta}.$$

In order for an element of $\Lambda^{\perp}$, given by (8.17), to be an observed-data influence function, it must satisfy (8.18); that is,

$$
\begin{aligned}
&I^{q\times q} = \\
&E\Bigg(\Bigg[\left\{ \frac{I(\mathcal{C}=\infty)\varphi^{*F}(Z)}{\varpi(\infty, Z)} + L_2\{\mathcal{C}, G_{\mathcal{C}}(Z)\} \right\} \\
&- \Pi[\{\ \cdot\ \}|\Lambda_{\psi}] \Bigg] E\{S_{\beta}^{F^T}(Z)|\mathcal{C}, G_{\mathcal{C}}(Z)\} \Bigg),
\end{aligned}
$$

which, by the law of iterated conditional expectations, used repeatedly, is equal to

$$
\begin{aligned}
&= E[E\{[\ \cdot\ ]S_{\beta}^{F^T}(Z)|\mathcal{C}, G_{\mathcal{C}}(Z)\}] \\
&= E\{[\ \cdot\ ]S_{\beta}^{F^T}(Z)\} = E[E\{[\ \cdot\ ]S_{\beta}^{F^T}(Z)|Z\}] \\
&= E[E\{[\ \cdot\ ]|Z\}S_{\beta}^{F^T}(Z)] \\
&= E\left(E\left[\frac{I(\mathcal{C}=\infty)\varphi^{*F}(Z)}{\varpi(\infty, Z)} + L_2\{\mathcal{C}, G_{\mathcal{C}}(Z)\} - \Pi\left[\{\ \cdot\ \}|\Lambda_{\psi}\right]|Z \right] S_{\beta}^{F^T}(Z) \right).
\end{aligned}
$$

Because

$$
\begin{aligned}
&E\left\{ \frac{I(\mathcal{C}=\infty)\varphi^{*F}(Z)}{\varpi(\infty, Z)} \middle| Z \right\} = \varphi^{*F}(Z), \\
&E\left[L_2\{\mathcal{C}, G_{\mathcal{C}}(Z)\}|Z\right] = 0 \quad \text{since} \quad L_2 \in \Lambda_2,
\end{aligned}
$$

and

$$
E\left\{\Pi(\{\ \cdot\ \}|\Lambda_{\psi})|Z\right\} = 0 \quad \text{since} \quad \Lambda_{\psi} \subset \Lambda_2,
$$

this implies that

$$
E\left\{\varphi^{*F}(Z)S_{\beta}^{F^T}(Z)\right\} = I^{q\times q}. \tag{8.20}
$$

Equation (8.20) is precisely the condition necessary for a typical element $\varphi^{*F} \in \Lambda^{F\perp}$ to be a full-data influence function.

Therefore, the space of observed-data influence functions consists of elements

$$
\varphi\{\mathcal{C}, G_{\mathcal{C}}(Z)\} = \left\{ \left[\frac{I(\mathcal{C}=\infty)\varphi^{F}(Z)}{\varpi(\infty, Z, \psi_0)} + L_2\{\mathcal{C}, G_{\mathcal{C}}(Z)\} \right] - \Pi\{[\ \cdot\ ]|\Lambda_{\psi}\}, \right.
$$

$$
\left. \text{where } \varphi^{F}(Z) \text{ is a full-data influence function and } L_2\{\mathcal{C}, G_{\mathcal{C}}(Z)\} \in \Lambda_2 \right\}. \ \square
$$

When data are coarsened by design, then the coarsening probabilities $\varpi\{r, G_r(Z)\}$ are known to us. We will sometimes refer to this as the parameter $\psi = \psi_0$ being known. When this is the case, there is no need to introduce the space Λ_{ψ}. Therefore, we obtain the following simple corollary.

Corollary 1. When the coarsening probabilities $\varpi\{r, G_r(Z)\}$ are known to us by design, then the space of observed-data influence functions is the linear variety consisting of elements

$$\varphi\{\mathcal{C}, G_{\mathcal{C}}(Z)\} = \left\{ \left[\frac{I(\mathcal{C}=\infty)\varphi^F(Z)}{\varpi(\infty, Z, \psi_0)} + L_2\{\mathcal{C}, G_{\mathcal{C}}(Z)\} \right], \right. \tag{8.21}$$

$$\left. \text{where } \varphi^F(Z) \text{ is a full-data influence function and } L_2\{\mathcal{C}, G_{\mathcal{C}}(Z)\} \in \Lambda_2 \right\}.$$

Remark 1. Notational convention
The space of observed-data influence functions will be denoted by (IF) and the space of full-data influence functions will be denoted by $(IF)^F$. As a result of Corollary 1, when data are coarsened by design, we can write the space (linear variety) of observed-data influence functions, using shorthand notation, as

$$(IF) = \frac{I(\mathcal{C}=\infty)(IF)^F}{\varpi(\infty, Z)} + \Lambda_2, \tag{8.22}$$

and, by Theorem 8.3, when the coarsening probabilities have to be modeled, as

$$(IF) = \left\{ \frac{I(\mathcal{C}=\infty)(IF)^F}{\varpi(\infty, Z)} + \Lambda_2 \right\} - \Pi(\{\ \cdot\ \}|\Lambda_\psi). \ \square \tag{8.23}$$

8.6 Recap and Review of Notation

Monotone coarsening

- An important special case of coarsened data is when the coarsening is monotone; that is, the coarsening variable can be ordered in such a way that $G_r(Z)$ is a many-to-one function of $G_{r+1}(Z)$ (i.e., $G_r(Z) = f_r\{G_{r+1}(Z)\}$).
- When coarsening is monotone, it is convenient to denote coarsening probabilities through the discrete hazard function. The definition and the relationship to coarsening probabilities are
  $\lambda_r\{G_r(Z)\} = P(\mathcal{C}=r|\mathcal{C}\geq r, Z),\ r \neq \infty,$
  $K_r\{G_r(Z)\} = P(\mathcal{C}>r|Z) = \prod_{r'=1}^{r}[1-\lambda_{r'}\{G_{r'}(Z)\}],\ r \neq \infty,$
  $\varpi\{r, G_r(Z)\} = K_{r-1}\{G_{r-1}(Z)\}\lambda_r\{G_r(Z)\}$, and
  the probability of a complete case is

$$P(\mathcal{C}=\infty|Z) = \varpi(\infty, Z) = K_\ell\{G_\ell(Z)\}.$$

The geometry of semiparametric models with coarsened data

- $(IF)^F$ denotes the space of full-data influence functions where a typical element is denoted by $\varphi^F(Z)$. This space is a linear variety where
  (i) $\varphi^F(Z) \in \Lambda^{F\perp}$,
  (ii) $E\{\varphi^F(Z)S_\beta^{F^T}(Z)\} = I^{q\times q}$.
- The observed-data nuisance tangent space

$$\Lambda = \Lambda_\psi \oplus \Lambda_\eta, \Lambda_\psi \ \perp \ \Lambda_\eta,$$

  where Λ_ψ, denoted as the coarsening model tangent space, is spanned by the score vector with respect to the parameter ψ used to describe the coarsening process, and Λ_η is spanned by the observed-data nuisance score vectors for parametric submodels and their mean-square closures. Specifically,

$$\Lambda_\eta = E\{\Lambda^F|\mathcal{C}, G_{\mathcal{C}}(Z)\}.$$

- $\Lambda_\psi \subset \Lambda_2, \quad \Lambda_\eta \ \perp \ \Lambda_2$
-

$$\Lambda_\eta^\perp = \left\{ \frac{I(\mathcal{C}=\infty)\Lambda^{F\perp}}{\varpi(\infty, Z)} \oplus \Lambda_2 \right\}.$$

- When the coarsening probabilities are unknown to us and need to be modeled, then

$$\Lambda^\perp = \left\{ \frac{I(\mathcal{C}=\infty)\Lambda^{F\perp}}{\varpi(\infty, Z)} \oplus \Lambda_2 - \Pi\left[\frac{I(\mathcal{C}=\infty)\Lambda^{F\perp}}{\varpi(\infty, Z)} \oplus \Lambda_2 \middle| \Lambda_\psi \right] \right\}.$$

-

$$(IF) = \left\{ \text{space of observed-data influence functions } \varphi\{\mathcal{C}, G_{\mathcal{C}}(Z)\} \right\}.$$

  When the coarsening probabilities are unknown to us and need to be modeled, then

$$(IF) = \left\{ \frac{I(\mathcal{C}=\infty)(IF)^F}{\varpi(\infty, Z)} + \Lambda_2 - \Pi\left[\frac{I(\mathcal{C}=\infty)(IF)^F}{\varpi(\infty, Z)} + \Lambda_2 \middle| \Lambda_\psi \right] \right\}.$$

  When the coarsening probabilities are known to us by design, then

$$(IF) = \frac{I(\mathcal{C}=\infty)(IF)^F}{\varpi(\infty, Z)} + \Lambda_2.$$

8.7 Exercises for Chapter 8

1. In Section 8.1, we described how data may be monotonically missing using longitudinal data $(Y_1, \ldots, Y_l)$ as an illustration. We also described a

model for the coarsening process where we modeled the discrete hazard function using equation (8.7) and derived the likelihood contribution for the coarsening model in (8.13). Let ψ_r denote the vector of parameters $(\psi_{0r}, \ldots, \psi_{rr})^T$ for $r = 1, \ldots, l-1$, and let ψ denote the entire parameter space for the coarsening probabilities; that is, $\psi = (\psi_1^T, \ldots, \psi_{l-1}^T)^T$.

a) Derive the score vector

$$S_{\psi_r}\{\mathcal{C}, G_{\mathcal{C}}(Z)\} = \frac{\partial \log[\varpi\{\mathcal{C}, G_{\mathcal{C}}(Z), \psi\}]}{\partial \psi_r}, r = 1, \ldots, l-1.$$

b) Show that the coarsening model tangent space Λ_ψ is equal to the direct sum

$$\Lambda_{\psi_1} \oplus \ldots \oplus \Lambda_{\psi_{l-1}},$$

where Λ_{ψ_r} is the linear space spanned by the vector $S_{\psi_r}\{\mathcal{C}, G_{\mathcal{C}}(Z)\}$, $r = 1, \ldots, l-1$, and that these spaces are mutually orthogonal to each other.

2. Give a formal proof of (8.16).

9

Augmented Inverse Probability Weighted Complete-Case Estimators

9.1 Deriving Semiparametric Estimators for β

ψ known

We begin by assuming the parameter ψ_0 that defines the coarsening model is known to us by design. Semiparametric estimators for β can be obtained by deriving elements orthogonal to the nuisance tangent space and using these to motivate estimating functions that can be used to construct estimating equations whose solution will lead to semiparametric estimators for β. In Chapter 7, we showed that when the parameter ψ_0 is known, the space orthogonal to the nuisance tangent space is the direct sum of the IPWCC space and the augmentation space, namely

$$\Lambda^{\perp} = \frac{I(\mathcal{C} = \infty)\Lambda^{F\perp}}{\varpi(\infty, Z, \psi_0)} \oplus \Lambda_2,$$

where a typical element of this space is

$$\frac{I(\mathcal{C} = \infty)\varphi^{*F}(Z)}{\varpi(\infty, Z, \psi_0)} + L_2\{\mathcal{C}, G_{\mathcal{C}}(Z), \psi_0\}, \tag{9.1}$$

where $\varphi^{*F}(Z)$ is an arbitrary element orthogonal to the full-data nuisance tangent space (i.e., $\varphi^{*F}(Z) \in \Lambda^{F\perp}$), and $L_2\{\mathcal{C}, G_{\mathcal{C}}(Z), \psi_0\}$ is an arbitrary element of Λ_2 that can be constructed by taking arbitrary functions $L_{2r}\{G_r(Z)\}, r \neq \infty$, and then using (7.37).

In Section 7.4, we gave examples of how these results could be used to derive augmented inverse probability weighted complete-case (AIPWCC) estimators for the regression parameters in a restricted moment model with missing data by design when there were two levels of missingness. We now expand this discussion to more general coarsening by design mechanisms.

If we want to obtain a semiparametric estimator for β, we would proceed as follows. We start with a full-data estimating equation that yields a full-data RAL estimator for β. It will be assumed that we know how to construct

such estimating equations for full-data semiparametric models. For example, a full-data m-estimator could be derived by solving the estimating equation

$$\sum_{i=1}^{n} m(Z_i, \beta) = 0,$$

where the estimating function evaluated at the truth, $m(Z, \beta_0)$, was chosen so that $m(Z, \beta_0) = \varphi^{*F}(Z) \in \Lambda^{F\perp}$. For example, we take $m(Z, \beta) = A(X)\{Y - \mu(X, \beta)\}$ for the restricted moment model. The influence function of such a full-data estimator for β was derived in Chapter 3, formula (3.6), and is given by

$$-\left[E\left\{\frac{\partial m(Z_i, \beta_0)}{\partial \beta^T}\right\}\right]^{-1} m(Z_i, \beta_0) = \varphi^F(Z_i). \tag{9.2}$$

Remark 1. An estimating function $m(Z, \beta)$ is a function of the random variable Z and the parameter β being estimated, and the corresponding m-estimator is the solution to the estimating equation made up of a sum of iid quantities $\sum_{i=1}^{n} m(Z_i, \beta) = 0$. However, in many cases, elements orthogonal to the nuisance tangent are defined as $\varphi^{*F}(Z) = m(Z, \beta_0, \eta_0)$, where η_0 denotes the true value of the full-data nuisance parameter. Consequently, it may not be possible to define an estimating function $m^*(Z, \beta)$ that is only a function of Z and β that satisfies $E_{\beta,\eta}\{m^*(Z, \beta)\} = 0$ for all β and η, as would be necessary to obtain consistent asymptotically normal estimators for all β and η. However, if we could find a consistent estimator for η, say $\hat{\eta}_n$, then a natural strategy would be to derive an estimator for β that is the solution to the estimating equation

$$\sum_{i=1}^{n} m(Z_i, \beta, \hat{\eta}_n) = 0. \tag{9.3}$$

The estimating equation (9.3) is not a sum of iid quantities and hence the resulting estimator is not, strictly speaking, an m-estimator. However, in many situations, and certainly in all cases considered in this book, the estimator $\hat{\beta}_n$ that solves (9.3) will be asymptotically equivalent to the m-estimator $\hat{\beta}_n^*$ that solves the equation

$$\sum_{i=1}^{n} m(Z_i, \beta, \eta_0) = 0,$$

with η_0 known, in the sense that $n^{1/2}(\hat{\beta}_n - \hat{\beta}_n^*) \xrightarrow{P} 0$. Without going into detail, this asymptotic equivalence occurs because $m(Z, \beta_0, \eta_0) = \varphi^{*F}(Z)$ is orthogonal to the nuisance tangent space. We illustrated this asymptotic equivalence for parametric models in Section 3.3 using equation (3.30) (also see Remark 4 of this section). Therefore, from here on, with a slight abuse of notation, we will still refer to estimators such as those that solve (9.3) as m-estimators with estimating function $m(Z, \beta)$. □

With coarsened data by design, we use (9.1) to motivate the following estimating equation:

$$\sum_{i=1}^{n}\left[\frac{I(\mathcal{C}_i=\infty)m(Z_i,\beta)}{\varpi(\infty,Z_i,\psi_0)}+L_2\{\mathcal{C}_i,G_{\mathcal{C}_i}(Z_i),\psi_0\}\right]=0. \tag{9.4}$$

The estimator that is the solution to the estimating equation (9.4) is referred to as an AIPWCC estimator. If we take the element $L_2(\cdot)$ to be identically equal to zero, then the estimating equation becomes

$$\sum_{i=1}^{n}\left[\frac{I(\mathcal{C}_i=\infty)m(Z_i,\beta)}{\varpi(\infty,Z_i,\psi_0)}\right]=0,$$

and the resulting estimator is an IPWCC estimator since only complete cases are considered in the sum above (i.e., $\{i:\mathcal{C}_i=\infty\}$), weighted by the inverse probability of being a complete case. The term $L_2\{\mathcal{C}_i,G_{\mathcal{C}_i}(Z_i),\psi_0\}$ allows contributions to the sum by observations that are not complete (i.e., coarsened), and this is referred to as the augmented term.

Using standard Taylor series expansions for m-estimators (which we leave as an exercise for the reader), we can show that the influence function of the estimator, derived by solving (9.4), is equal to

$$\begin{aligned}&-\left[E\left\{\frac{\partial m(Z_i,\beta_0)}{\partial\beta^T}\right\}\right]^{-1}\left[\frac{I(\mathcal{C}_i=\infty)m(Z_i,\beta_0)}{\varpi(\infty,Z_i,\psi_0)}+L_2\{\mathcal{C}_i,G_{\mathcal{C}_i}(Z_i),\psi_0\}\right]\\&=\frac{I(\mathcal{C}_i=\infty)\varphi^F(Z_i)}{\varpi(\infty,Z_i,\psi_0)}+L_2^*\{\mathcal{C}_i,G_{\mathcal{C}_i}(Z_i),\psi_0\},\end{aligned} \tag{9.5}$$

where $\varphi^F(Z_i)$ was defined in (9.2) and

$$L_2^*=-\left[E\left\{\frac{\partial m(Z_i,\beta_0)}{\partial\beta^T}\right\}\right]^{-1}L_2\in\Lambda_2. \tag{9.6}$$

Therefore, we can now summarize these results. If coarsening of the data were by design with known coarsening probabilities $\varpi\{r,G_r(Z),\psi_0\}$, for all r, and we wanted to obtain an observed-data RAL estimator for β in a semiparametric model, we would proceed as follows.

1. Choose a full-data estimating function $m(Z,\beta)$.
2. Choose an element of the augmentation space Λ_2 as follows.
   a) For each $r\neq\infty$, choose a function $L_{2r}\{G_r(Z)\}$.
   b) Construct $L_2\{\mathcal{C},G_{\mathcal{C}}(Z),\psi_0\}$ to equal

$$\begin{aligned}&\frac{I(\mathcal{C}=\infty)}{\varpi(\infty,Z,\psi_0)}\left[\sum_{r\neq\infty}\varpi\{r,G_r(Z),\psi_0\}L_{2r}\{G_r(Z)\}\right]\\&-\sum_{r\neq\infty}I(\mathcal{C}=r)L_{2r}\{G_r(Z)\}.\end{aligned}$$

3. Obtain the estimator for β by solving equation (9.4).

The resulting estimator, $\hat{\beta}_n$, under suitable regularity conditions, will be a consistent, asymptotically normal RAL estimator for β with influence function given by (9.5). The asymptotic variance of this estimator is, of course, the variance of the influence function. Estimators for the asymptotic variance can be obtained using the sandwich variance estimator (3.10) derived in Chapter 3 specifically for m-estimators.

We see from this construction that the estimator depends on the choice of $m(Z, \beta)$ and the functions $L_{2r}\{G_r(Z)\}, r \neq \infty$. Of course, we would want to choose these functions so that the resulting estimator is as efficient as possible. This issue will be the focus of Chapters 10 and 11.

$\boldsymbol{\psi}$ *unknown*

The development above shows how we can take results regarding semiparametric estimators for the parameter β for full-data models and modify them to estimate the parameter β with coarsened data (CAR) when the the coarsening probabilities are known to us by design. In most problems, the coarsening probabilities are not known and must be modeled using the unknown parameter ψ. We discussed models for the coarsening process and estimators for the parameters in these models in Chapter 8. We also showed in Chapter 8 the impact that such models have on the observed-data nuisance tangent space, its orthogonal complement, and the space of observed-data influence functions.

If the parameter ψ is unknown, two issues emerge:

(i) The unknown parameter ψ must be estimated.
(ii) The influence function of an observed-data RAL estimator for β must be an element in the space defined by (7.37) (i.e., involving a projection onto the coarsening model tangent space Λ_ψ).

One obvious strategy for estimating β with coarsened data when ψ is unknown is to find a consistent estimator for ψ and substitute this estimator for ψ_0 in the estimating equation (9.4). A natural estimator for ψ is obtained by maximizing the coarsening model likelihood (8.8). The resulting MLE is denoted by $\hat{\psi}_n$. The influence function of the estimator for β, obtained by substituting the maximum likelihood estimator $\hat{\psi}_n$ for ψ_0 in equation (9.4), is given by the following important theorem.

Theorem 9.1. If the coarsening process follows a parametric model, and if ψ is estimated using the maximum likelihood estimator, say $\hat{\psi}_n$, or any efficient estimator of ψ, then the solution to the estimating equation

$$\sum_{i=1}^{n} \left[\frac{I(\mathcal{C}_i = \infty) m(Z_i, \beta)}{\varpi(\infty, Z_i, \hat{\psi}_n)} + L_2\{\mathcal{C}_i, G_{\mathcal{C}_i}(Z_i), \hat{\psi}_n\} \right] = 0 \tag{9.7}$$

will be an estimator whose influence function is

$$\frac{I(\mathcal{C}_i = \infty)\varphi^F(Z_i)}{\varpi(\infty, Z_i, \psi_0)} + L_2^*\{\mathcal{C}_i, G_{\mathcal{C}_i}(Z_i), \psi_0\}$$
$$- \Pi\left[\frac{I(\mathcal{C}_i = \infty)\varphi^F(Z_i)}{\varpi(\infty, Z_i, \psi_0)} + L_2^*\{\mathcal{C}_i, G_{\mathcal{C}_i}(Z_i), \psi_0\}\middle|\Lambda_\psi\right], \tag{9.8}$$

where $\varphi^F(\cdot)$ and $L_2^*(\cdot)$ are defined by (9.2) and (9.6). We note that such an influence function is indeed a member of the class of observed-data influence functions given by (8.19).

For notational convenience, we denote a typical influence function, if the parameter ψ is known, by

$$\tilde{\varphi}\{\mathcal{C}_i, G_{\mathcal{C}_i}(Z_i), \psi\} = \frac{I(\mathcal{C}_i = \infty)\varphi^F(Z_i)}{\varpi(\infty, Z_i, \psi)} + L_2^*\{\mathcal{C}_i, G_{\mathcal{C}_i}(Z_i), \psi\}, \tag{9.9}$$

and a typical influence function, if ψ is unknown, by

$$\varphi\{\mathcal{C}_i, G_{\mathcal{C}_i}(Z_i), \psi\} = \tilde{\varphi}\{\mathcal{C}_i, G_{\mathcal{C}_i}(Z_i), \psi\} - \Pi[\tilde{\varphi}\{\mathcal{C}_i, G_{\mathcal{C}_i}(Z_i), \psi\}|\Lambda_\psi].$$

Before giving the proof of the theorem above we present the following lemma.

Lemma 9.1.

$$E\left[\frac{\partial\tilde{\varphi}\{\mathcal{C}, G_{\mathcal{C}}(Z), \psi_0\}}{\partial\psi^T}\right] = -E\left[\tilde{\varphi}\{\mathcal{C}, G_{\mathcal{C}}(Z), \psi_0\}S_\psi^T\{\mathcal{C}, G_{\mathcal{C}}(Z), \psi_0\}\right]. \tag{9.10}$$

Proof. Lemma 9.1
We first note that the conditional expectation $E[h\{\mathcal{C}, G_{\mathcal{C}}(Z)\}|Z]$ for a typical function h, as given by (7.35), only depends on the parameter ψ. Namely, this conditional expectation equals

$$E_\psi(h|Z) = \sum_r h_r\{G_r(Z)\}\varpi\{r, G_r(Z), \psi\}. \tag{9.11}$$

Because of the definition of $\tilde{\varphi}\{\mathcal{C}, G_{\mathcal{C}}(Z), \psi\}$ given by (9.9) and the fact that $L_2^*\{\mathcal{C}, G_{\mathcal{C}}(Z), \psi\} \in \Lambda_2$, we obtain for any ψ, that

$$E_\psi\left\{\tilde{\varphi}\{\mathcal{C}, G_{\mathcal{C}}, \psi)\middle|Z\right\} = \varphi^F(Z),$$

which, by equation (9.11), equals

$$\sum_r \tilde{\varphi}\{r, G_r(z), \psi\}\varpi\{r, G_r(z), \psi\} = \varphi^F(z) \text{ for all } z, \psi,$$

where $\varphi^F(z)$ does not include the parameter ψ. Therefore

$$\frac{\partial}{\partial\psi^T}\sum_r \tilde{\varphi}\{r, G_r(z), \psi\}\varpi\{r, G_r(z), \psi\} = 0 \quad \text{for all } z, \psi. \tag{9.12}$$

Differentiating the product inside the sum (9.12) and setting $\psi = \psi_0$ yields

$$\sum_r \frac{\partial \tilde{\varphi}\{r, G_r(z), \psi_0\}}{\partial \psi^T} \varpi\{r, G_r(z), \psi_0\} + \sum_r \tilde{\varphi}\{r, G_r(z), \psi_0\} \frac{\partial \varpi\{r, G_r(z), \psi_0\}/\partial \psi^T}{\varpi\{r, G_r(z), \psi_0\}} \varpi\{r, G_r(z), \psi_0\} = 0,$$

or

$$E\left[\frac{\partial \tilde{\varphi}\{\mathcal{C}, G_{\mathcal{C}}(Z), \psi_0\}}{\partial \psi^T}\middle| Z\right] = -E\left[\tilde{\varphi}\{\mathcal{C}, G_{\mathcal{C}}(Z), \psi_0\} S_\psi^T\{\mathcal{C}, G_{\mathcal{C}}(Z), \psi_0\}|Z\right],$$

which, after taking unconditional expectations, implies

$$E\left[\frac{\partial \tilde{\varphi}\{\mathcal{C}, G_{\mathcal{C}}(Z), \psi_0\}}{\partial \psi^T}\right] = -E\left[\tilde{\varphi}\{\mathcal{C}, G_{\mathcal{C}}(Z), \psi_0\} S_\psi^T\{\mathcal{C}, G_{\mathcal{C}}(Z), \psi_0\}\right]. \quad \square$$

We are now in a position to prove Theorem 9.1.

Proof. Theorem 9.1
The usual expansion of (9.7) about β_0, but keeping $\hat{\psi}_n$ fixed, yields

$$n^{1/2}(\hat{\beta}_n - \beta_0) = n^{-1/2} \sum_{i=1}^n \left[\frac{I(\mathcal{C}_i = \infty)\varphi^F(Z_i)}{\varpi(\infty, Z_i, \hat{\psi}_n)} + L_2^*\{\mathcal{C}_i, G_{\mathcal{C}_i}(Z_i), \hat{\psi}_n\}\right] + o_p(1), \tag{9.13}$$

where $\varphi^F(Z_i)$ is given by (9.2) and $L_2^*\{\mathcal{C}_i, G_{\mathcal{C}_i}(Z_i), \psi\}$ is given by (9.6). Now we expand $\hat{\psi}_n$ about ψ_0 to obtain

$$n^{1/2}(\hat{\beta}_n - \beta_0) = n^{-1/2} \sum_{i=1}^n \tilde{\varphi}\{\mathcal{C}_i, G_{\mathcal{C}_i}(Z_i), \psi_0\} + \left[n^{-1} \sum_{i=1}^n \frac{\partial \tilde{\varphi}\{\mathcal{C}_i, G_{\mathcal{C}_i}(Z_i), \psi_n^*\}}{\partial \psi^T}\right] n^{1/2}(\hat{\psi}_n - \psi_0) + o_p(1), \tag{9.14}$$

where ψ_n^* is some intermediate value between $\hat{\psi}_n$ and ψ_0. Since under usual regularity conditions ψ_n^* converges in probability to ψ_0, we obtain

$$n^{-1} \sum_{i=1}^n \frac{\partial \tilde{\varphi}\{\mathcal{C}_i, G_{\mathcal{C}_i}(Z_i), \psi_n^*\}}{\partial \psi^T} \xrightarrow{P} E\left[\frac{\partial \tilde{\varphi}\{\mathcal{C}_i, G_{\mathcal{C}_i}(Z_i), \psi_0\}}{\partial \psi^T}\right]. \tag{9.15}$$

Standard results for finite-dimensional parametric models, as derived in Chapter 3, can be used to show that the influence function of the MLE $\hat{\psi}_n$ is given by

$$\left(E\left[S_\psi\{\mathcal{C}, G_{\mathcal{C}}(Z), \psi_0\} S_\psi^T\{\mathcal{C}, G_{\mathcal{C}}(Z), \psi_0\} \right] \right)^{-1} S_\psi\{\mathcal{C}, G_{\mathcal{C}}(Z), \psi_0\}, \qquad (9.16)$$

where $S_\psi\{\mathcal{C}, G_{\mathcal{C}}(Z), \psi_0\}$ is the score vector with respect to ψ defined in (8.15).

The influence function of $\hat{\psi}_n$ given by (9.16), together with (9.15) and (9.10) of Lemma 9.1, can be used to deduce that (9.14) is equal to

$$\begin{aligned} & n^{1/2}(\hat{\beta}_n - \beta_0) \\ & = n^{-1/2} \sum_{i=1}^{n} \Bigg\{ \tilde{\varphi}\{\mathcal{C}_i, G_{\mathcal{C}_i}(Z_i), \psi_0\} - E\left[\tilde{\varphi}\{\mathcal{C}_i, G_{\mathcal{C}_i}(Z_i), \psi_0\} S_\psi^T\{\mathcal{C}_i, G_{\mathcal{C}_i}, \psi_0\} \right] \\ & \left(E\left[S_\psi\{\mathcal{C}_i, G_{\mathcal{C}_i}(Z_i), \psi_0\} S_\psi^T\{\mathcal{C}_i, G_{\mathcal{C}_i}(Z_i), \psi_0\} \right] \right)^{-1} S_\psi\{\mathcal{C}_i, G_{\mathcal{C}_i}(Z_i), \psi_0\} \Bigg\} \\ & \quad + o_p(1). \end{aligned} \qquad (9.17)$$

The space Λ_ψ is the linear space spanned by $S_\psi\{\mathcal{C}, G_{\mathcal{C}}(Z), \psi_0\}$. Therefore, using results from Chapter 2 (equation (2.4)) for finding projections onto finite-dimensional linear subspaces, we obtain that

$$\Pi\left[\tilde{\varphi}\{\mathcal{C}, G_{\mathcal{C}}(Z)\} | \Lambda_\psi \right] = E(\tilde{\varphi} S_\psi^T) \left[E(S_\psi S_\psi^T) \right]^{-1} S_\psi\{\mathcal{C}, G_{\mathcal{C}}(Z)\}. \qquad (9.18)$$

Consequently, as a result of (9.17) and (9.18), the influence function of $\hat{\beta}_n$ is given by

$$\tilde{\varphi}\{\mathcal{C}, G_{\mathcal{C}}(Z), \psi_0\} - \Pi[\tilde{\varphi}\{\mathcal{C}, G_{\mathcal{C}}(Z), \psi_0\} | \Lambda_\psi],$$

where

$$\tilde{\varphi}\{\mathcal{C}, G_{\mathcal{C}}(Z), \psi_0\} = \frac{I(\mathcal{C} = \infty)\varphi^F(Z)}{\varpi(\infty, Z, \psi_0)} + L_2^*\{\mathcal{C}, G_{\mathcal{C}}(Z), \psi_0\}. \quad \square$$

Theorem 9.1 is important because it gives us a prescription for deriving observed-data RAL estimators for β when coarsening probabilities need to be modeled and estimated. Summarizing these results, if the data are coarsened at random with coarsening probabilities that need to be modeled, and we want to obtain an observed-data RAL estimator for β in a semiparametric model, we would proceed as follows.

1. Choose a model for the coarsening probabilities in terms of a parameter ψ, namely $\varpi\{r, G_r(Z), \psi\}$, for all r. In Section 8.1, we discussed how such models can be derived when there are two levels of coarsening or when the coarsening is monotone.
2. Using the observed data, estimate the parameter ψ using maximum likelihood (that is, by maximizing (8.8)), and denote the estimator by $\hat{\psi}_n$.
3. Choose a full-data estimating function $m(Z, \beta)$.
4. Choose an element of the augmentation space Λ_2 by
   a) for each $r \neq \infty$, choosing a function $L_{2r}\{G_r(Z)\}$ and

b) constructing $L_2\{\mathcal{C}, G_{\mathcal{C}}(Z), \hat{\psi}_n\}$ to equal

$$\frac{I(\mathcal{C}=\infty)}{\varpi(\infty, Z, \hat{\psi}_n)}\left[\sum_{r\neq\infty}\varpi\{r, G_r(Z), \hat{\psi}_n\}L_{2r}\{G_r(Z)\}\right] - \sum_{r\neq\infty} I(\mathcal{C}=r)L_{2r}\{G_r(Z)\}.$$

5. Obtain the estimator for β by solving equation (9.7).

The resulting estimator, $\hat{\beta}_n$, under suitable regularity conditions, will be a consistent, asymptotically normal RAL estimator for β with influence function given by (9.8).

Interesting Fact

If the parameter ψ_0 was known to us and we used equation (9.4) to derive an estimator for β, then, since this estimator is asymptotically linear, the asymptotic variance would be the variance of the influence function (9.5), namely

$$\text{var}\left[\frac{I(\mathcal{C}=\infty)\varphi^F(Z)}{\varpi(\infty, Z, \psi_0)} + L_2^*\{\mathcal{C}, G_{\mathcal{C}}(Z), \psi_0\}\right].$$

If, however, ψ was estimated using the MLE $\hat{\psi}_n$ and equation (9.7) was used to derive an estimator for β, then the asymptotic variance would be the variance of its influence function namely

$$\text{var}\left(\frac{I(\mathcal{C}=\infty)\varphi^F(Z)}{\varpi(\infty, Z, \psi_0)} + L_2^*\{\mathcal{C}, G_{\mathcal{C}}(Z), \psi_0\} - \Pi\left[\frac{I(\mathcal{C}=\infty)\varphi^F(Z)}{\varpi(\infty, Z, \psi_0)} + L_2^*\{\mathcal{C}, G_{\mathcal{R}}(Z), \psi_0\}\middle|\Lambda_\psi\right]\right).$$

By the Pythagorean theorem, the variance of the second estimator must be less than or equal to the variance of the first. This leads to the interesting and perhaps unintuitive result that, even if we know the coarsening probabilities (say we coarsen the data by design), then we would still obtain a more efficient estimator for β by estimating the parameter ψ in a model that contains the truth and substituting this estimator for ψ in equation (9.4) rather than using the true ψ_0 itself.

Estimating the Asymptotic Variance

The asymptotic variance of the RAL estimator $\hat{\beta}_n$ is the variance of the influence function (9.8), which we denote by Σ. An estimator for the asymptotic variance, $\hat{\Sigma}_n$, can be obtained using a sandwich variance estimator. For completeness, we now describe how to construct this estimator:

$$\hat{\Sigma}_n = \left[\hat{E}\left\{\frac{\partial m(Z,\hat{\beta}_n)}{\partial \beta^T}\right\}\right]^{-1} \times \left[n^{-1}\sum_{i=1}^{n} g\{\mathcal{C}_i, G_{\mathcal{C}_i}(Z_i), \hat{\psi}_n, \hat{\beta}_n\} g^T\{\mathcal{C}_i, G_{\mathcal{C}_i}(Z_i), \hat{\psi}_n, \hat{\beta}_n\}\right] \times \left[\hat{E}\left\{\frac{\partial m(Z,\hat{\beta}_n)}{\partial \beta^T}\right\}\right]^{-1^T}, \tag{9.19}$$

where

$$\hat{E}\left\{\frac{\partial m(Z,\hat{\beta}_n)}{\partial \beta^T}\right\} = n^{-1}\sum_{i=1}^{n}\left\{\frac{I(\mathcal{C}_i=\infty)\partial m(Z_i,\hat{\beta}_n)/\partial \beta^T}{\varpi(\infty, Z_i, \hat{\psi}_n)}\right\},$$

$$g\{\mathcal{C}_i, G_{\mathcal{C}_i}(Z_i), \hat{\psi}_n, \hat{\beta}_n\} = q\{\mathcal{C}_i, G_{\mathcal{C}_i}(Z_i), \hat{\psi}_n, \hat{\beta}_n\} - \hat{E}(qS_\psi^T)\{\hat{E}(S_\psi S_\psi^T)\}^{-1} S_\psi\{\mathcal{C}_i, G_{\mathcal{C}_i}(Z_i), \hat{\psi}_n\},$$

$$q\{\mathcal{C}_i, G_{\mathcal{C}_i}(Z_i), \hat{\psi}_n, \hat{\beta}_n\} = \frac{I(\mathcal{C}_i=\infty)m(Z_i,\hat{\beta}_n)}{\varpi(\infty, Z_i, \hat{\psi}_n)} + L_2\{\mathcal{C}_i, G_{\mathcal{C}_i}(Z_i), \hat{\psi}_n\},$$

$$\hat{E}(qS_\psi^T) = n^{-1}\sum_{i=1}^{n} q\{\mathcal{C}_i, G_{\mathcal{C}_i}(Z_i), \hat{\psi}_n, \hat{\beta}_n\} S_\psi^T\{\mathcal{C}_i, G_{\mathcal{C}_i}(Z_i), \hat{\psi}_n\},$$

and

$$\hat{E}(S_\psi S_\psi^T) = n^{-1}\sum_{i=1}^{n} S_\psi\{\mathcal{C}_i, G_{\mathcal{C}_i}(Z_i), \hat{\psi}_n\} S_\psi^T\{\mathcal{C}_i, G_{\mathcal{C}_i}(Z_i), \hat{\psi}_n\}.$$

9.2 Additional Results Regarding Monotone Coarsening

The Augmentation Space Λ_2 with Monotone Coarsening

In deriving arbitrary semiparametric estimators for β with coarsened data, whether the coarsening process was by design, ψ_0 known, or whether the parameter ψ describing the coarsening process had to be estimated, a key component of the estimating equation, either (9.4) or (9.7), was the augmentation term $L_2\{\mathcal{C}, G_{\mathcal{C}}(Z)\} \in \Lambda_2$. In (7.37), we gave a general representation for such an arbitrary element of Λ_2 when the data are coarsened at random. However, in the special case when we have monotone coarsening, it will be convenient to derive another equivalent representation for the elements in Λ_2. This representation uses discrete hazards as defined in (8.2) and is given in the following theorem.

Theorem 9.2. Under monotone coarsening, a typical element of Λ_2 can be written as

$$\sum_{r\neq\infty}\left[\frac{I(\mathcal{C}=r)-\lambda_r\{G_r(Z)\}I(\mathcal{C}\geq r)}{K_r\{G_r(Z)\}}\right]L_r\{G_r(Z)\}, \tag{9.20}$$

where $\lambda_r\{G_r(Z)\}$ and $K_r\{G_r(Z)\}$ are defined by (8.2) and (8.4), respectively, and $L_r\{G_r(Z)\}$ denotes an arbitrary function of $G_r(Z)$ for $r\neq\infty$.

Remark 2. Equation (9.20) is made up of a sum of mean-zero conditionally uncorrelated terms; i.e., it has a martingale structure. We will take advantage of this structure in the next chapter when we derive the more efficient double robust estimators. □

Proof. Using (7.37), we note that a typical element of Λ_2 can be written as

$$\sum_{r\neq\infty}\left[I(\mathcal{C}=r)-\frac{I(\mathcal{C}=\infty)\varpi\{r,G_r(Z)\}}{\varpi(\infty,Z)}\right]L_{2r}\{G_r(Z)\}, \tag{9.21}$$

where $L_{2r}\{G_r(Z)\}$ denotes an arbitrary function of $G_r(Z)$ for $r\neq\infty$. To simplify notation, let $\lambda_r=\lambda_r\{G_r(Z)\}$, $K_r=K_r\{G_r(Z)\}$, $L_r=L_r\{G_r(Z)\}$, and $L_{2r}=L_{2r}\{G_r(Z)\}$. We will prove that there is a one-to-one relationship between the elements in (9.20) and the elements of (9.21), specifically that

$$L_{2r}=\frac{1}{K_{r-1}}L_r-\sum_{j=1}^{r-1}\frac{\lambda_j}{K_j}L_j \text{ for all } r\neq\infty, \tag{9.22}$$

and conversely

$$L_r=K_{r-1}L_{2r}+\sum_{j=1}^{r-1}\varpi_jL_{2j} \text{ for all } r\neq\infty. \tag{9.23}$$

We prove (9.23) by induction. For $r=1$, notice that $K_0=1$, and hence by (9.22), $L_1=L_{21}$. Now suppose that (9.23) holds for any $i\leq r$. Then for $r+1$, we have from (9.22) that

$$L_{2(r+1)}=\frac{1}{K_r}L_{r+1}-\sum_{i=1}^{r}\frac{\lambda_i}{K_i}L_i.$$

Therefore,

$$\begin{aligned}
L_{r+1}&=K_rL_{2(r+1)}+\sum_{i=1}^{r}\frac{K_r\lambda_i}{K_i}L_i\\
&=K_rL_{2(r+1)}+\sum_{i=1}^{r}\frac{K_r\lambda_i}{K_i}\left(K_{i-1}L_{2i}+\sum_{j=1}^{i-1}\varpi_jL_{2j}\right)\\
&=K_rL_{2(r+1)}+\sum_{i=1}^{r}\frac{K_r\varpi_i}{K_i}L_{2i}+\sum_{i=1}^{r}\sum_{j=1}^{i-1}\frac{K_r\lambda_i\varpi_j}{K_i}L_{2j}.
\end{aligned}$$

Interchange the order of summation,

$$\begin{aligned} L_{r+1} &= K_r L_{2(r+1)} + \sum_{i=1}^{r} \frac{K_r \varpi_i}{K_i} L_{2i} + \sum_{j=1}^{r-1} K_r \varpi_j L_{2j} \sum_{i=j+1}^{r} \frac{\lambda_i}{K_i} \\ &= K_r L_{2(r+1)} + \sum_{j=1}^{r} K_r \varpi_j L_{2j} \left(\frac{1}{K_j} + \sum_{i=j+1}^{r} \frac{\lambda_i}{K_i} \right). \end{aligned}$$

Note that

$$\frac{1}{K_j} + \frac{\lambda_{j+1}}{K_{j+1}} = \frac{1 - \lambda_{j+1}}{K_{j+1}} + \frac{\lambda_{j+1}}{K_{j+1}} = \frac{1}{K_{j+1}},$$

and hence

$$\frac{1}{K_j} + \sum_{i=j+1}^{r} \frac{\lambda_i}{K_i} = \frac{1}{K_r}.$$

Therefore,

$$L_{r+1} = K_r L_{2(r+1)} + \sum_{j=1}^{r} \varpi_j L_{2j}.$$

Substituting (9.23) into (9.20) yields

$$\begin{aligned} &\sum_{r \neq \infty} \frac{I(\mathcal{C} = r) - \lambda_r I(\mathcal{C} \geq r)}{K_r} \left(K_{r-1} L_{2r} + \sum_{j=1}^{r-1} \varpi_j L_{2j} \right) \\ &= \sum_{r \neq \infty} \frac{I(\mathcal{C} = r) - \lambda_r I(\mathcal{C} \geq r)}{K_r} K_{r-1} L_{2r} \\ &\qquad + \sum_{j \neq \infty} \varpi_j L_{2j} \sum_{j+1 \leq r \neq \infty} \frac{I(\mathcal{C} = r) - \lambda_r I(\mathcal{C} \geq r)}{K_r} \\ &= \sum_{r \neq \infty} \frac{I(\mathcal{C} = r) - \lambda_r I(\mathcal{C} \geq r)}{K_r} K_{r-1} L_{2r} \\ &\qquad + \sum_{r \neq \infty} \varpi_r L_{2r} \sum_{r+1 \leq j \neq \infty} \frac{I(\mathcal{C} = j) - \lambda_j I(\mathcal{C} \geq j)}{K_j}. \end{aligned}$$

But since $I(\mathcal{C} = j) = I(\mathcal{C} \geq j) - I(\mathcal{C} \geq j + 1)$, we have that

$$\begin{aligned} &\sum_{r+1 \leq j \neq \infty} \frac{I(\mathcal{C} = j) - \lambda_j I(\mathcal{C} \geq j)}{K_j} \\ &= \sum_{r+1 \leq j \neq \infty} \left\{ \frac{I(\mathcal{C} \geq j)}{K_{j-1}} - \frac{I(\mathcal{C} \geq j + 1)}{K_j} \right\} \\ &= \frac{I(\mathcal{C} \geq r + 1)}{K_r} - \frac{I(\mathcal{C} = \infty)}{\varpi_\infty}. \end{aligned}$$

Therefore, (9.20) can be written as

$$\begin{aligned}
&\sum_{r\neq\infty}\left\{\frac{I(\mathcal{C}=r)K_{r-1}}{K_r}-\frac{I(\mathcal{C}\geq r)\varpi_r}{K_r}+\frac{I(\mathcal{C}\geq r+1)\varpi_r}{K_r}-\frac{I(\mathcal{C}=\infty)\varpi_r}{\varpi_\infty}\right\}L_{2r}\\
&=\sum_{r\neq\infty}\left\{\frac{I(\mathcal{C}=r)K_{r-1}}{K_r}-\frac{I(\mathcal{C}=r)\varpi_r}{K_r}-\frac{I(\mathcal{C}=\infty)\varpi_r}{\varpi_\infty}\right\}L_{2r}\\
&=\sum_{r\neq\infty}\left\{I(\mathcal{C}=r)-\frac{I(\mathcal{C}=\infty)\varpi_r}{\varpi_\infty}\right\}L_{2r},
\end{aligned}$$

which is exactly (9.21). Similarly, substituting (9.22) into (9.21) yields (9.20). □

Example 1. Longitudinal data with monotone missingness
Consider the following problem. A promising new drug for patients with HIV disease is to be evaluated against a control treatment in a randomized clinical trial. A sample of n patients with HIV disease are randomized with equal probability (.5) to receive either the new treatment ($X = 1$) or the control treatment ($X = 0$). The primary endpoint used to evaluate the treatments is change in CD4 counts measured over time. Specifically, CD4 counts are to be measured at time points $t_1 < \ldots < t_l$, where $t_1 = 0$ denotes the time of entry into the study when a baseline CD4 count measurement is taken and $t_2 < \ldots < t_l$ are the subsequent times after treatment when CD4 counts will be measured. For example, CD4 count measurements may be taken every six months, with a final measurement at two years. The data for individual i can be summarized as $Z_i = (Y_i^T, X_i)^T$, where $Y_i = (Y_{1i}, \ldots, Y_{li})^T$, with Y_{ji} denoting the CD4 count measurement for patient i at time t_j and X_i denoting the treatment indicator to which patient i was assigned.

Suppose that it is generally believed that, after treatment is initiated, CD4 counts will roughly follow a linear trajectory over time. Therefore, a linear model

$$E(Y_i^{l\times 1}|X_i) = H^{l\times 3}(X_i)\beta^{3\times 1}, \tag{9.24}$$

is used to describe the data, where the design matrix $H(X_i)$ is an $l \times 3$ matrix with elements $H_{jj'}(X_i), j = 1, \ldots, l,\ j' = 1, 2, 3$, and $H_{j1}(X_i) = 1$, $H_{j2}(X_i) = t_j$, and $H_{j3}(X_i) = X_i t_j$. This model implies that the mean CD4 count at time t_j is given by

$$E(Y_{ji}|X_i) = \beta_1 + \beta_2 t_j + \beta_3 X_i t_j.$$

Hence, if $X_i = 0$, then the expected CD4 count is $\beta_1 + \beta_2 t_j$, whereas if $X_i = 1$, then the expected CD4 count is $\beta_1 + (\beta_2 + \beta_3)t_j$. This reflects the belief that CD4 response follows a linear trajectory after treatment is initiated. Because of randomization, the mean baseline response at time $t_1 = 0$ equals β_1, the same for both treatments, but the slope of the trajectory may depend on

treatment. The parameter β_3 reflects the strength of the treatment effect, where $\beta_3 = 0$ corresponds to the null hypothesis of no treatment effect.

The model (9.24) is an example of a restricted moment model, where $E(Y|X) = \mu(X,\beta) = H(X)\beta$. If the data were collected on everyone (i.e., full data), then according to the results developed in Section 4.5, a full-data estimating function would in general be based on $A^{3\times l}(X)\{Y - \mu(X,\beta)\}$, and the optimal estimating function would be $D^T(X)V^{-1}(X)\{Y - \mu(X,\beta)\}$, where $D(X) = \frac{\partial \mu(X,\beta)}{\partial \beta^T}$ and $V(X) = \text{var}(Y|X)$. In this example, $D(X) = H(X)$. Although we would expect the longitudinal CD4 count measurements to be correlated, for simplicity we use a working variance function $V(X) = \sigma^2 I^{l\times l}$. This leads to the full-data estimating function $m(Z,\beta) = H^T(X)\{Y - H(X)\beta\}$, and the corresponding full-data estimator would be obtained by solving the estimating equation

$$\sum_{i=1}^{n} H^T(X_i)\{Y_i - H(X_i)\beta\} = 0.$$

In this study, however, some patients dropped out during the course of the study, in which case we would observe the data up to the time they dropped out but all subsequent CD4 count measurements would be missing. This is an example of monotone coarsening as described in Section 8.1, where there are $\ell = l - 1$ levels of coarsening and where $G_r(Z_i) = (X_i, Y_{1i}, \ldots, Y_{ri})^T, r = 1, \ldots, l-1$ and $G_\infty(Z_i) = (X_i, Y_{1i}, \ldots, Y_{li})^T$. Because dropout was not by design, we need to model the coarsening probabilities in order to derive AIPWCC estimators for β. Since the data are monotonically coarsened, it is more convenient to model the discrete hazard function. Assuming the coarsening is CAR, we consider a series of logistic regression models similar to (8.7), namely

$$\lambda_r\{G_r(Z)\} = \frac{\exp(\psi_{0r} + \psi_{1r}Y_1 + \ldots + \psi_{rr}Y_r + \psi_{(r+1)r}X)}{1 + \exp(\psi_{0r} + \psi_{1r}Y_1 + \ldots + \psi_{rr}Y_r + \psi_{(r+1)r}X)}. \tag{9.25}$$

Note 1. The only difference between equations (8.7) and (9.25) is the inclusion of the treatment indicator X. □

Therefore, the parameter ψ in this example is $\psi = (\psi_{0r}, \ldots, \psi_{(r+1)r},\ r = 1, \ldots, l-1)^T$. The MLE $\hat{\psi}_n$ is obtained by maximizing the likelihood (8.13).

We now have most of the components necessary to construct an AIPWCC estimator for β. We still need to define an element of the augmentation space Λ_2. In accordance with Theorem 9.2, for monotonically coarsened data, we must choose a function $L_r\{G_r(Z)\},\ r = 1, \ldots, l-1$ (that is, a function $L_r(X, Y_1, \ldots, Y_r)$) and then use (9.20) to construct an element $L_2\{\mathcal{C}, G_{\mathcal{C}}(Z)\} \in \Lambda_2$.

Putting all these different elements together, we can derive an observed-data RAL estimator for β by using the results of Theorem 9.1, equation (9.7), by solving the estimating equation

$$\sum_{i=1}^{n}\left[\frac{I(\mathcal{C}_i=\infty)H^T(X_i)\{Y_i-H(X_i)\beta\}}{\varpi(\infty,Z_i,\hat{\psi}_n)}\right. \tag{9.26}$$
$$\left.+\sum_{r=1}^{l-1}\left\{\frac{I(\mathcal{C}_i=r)-\lambda_r\{G_r(Z_i),\hat{\psi}_n\}I(\mathcal{C}_i\geq r)}{K_r\{G_r(Z_i),\hat{\psi}_n\}}\right\}L_r\{G_r(Z)\}\right]=0,$$

where

$$\lambda_r\{G_r(Z_i),\hat{\psi}_n\}=\frac{\exp(\hat{\psi}_{0r}+\hat{\psi}_{1r}Y_{1i}+\ldots+\hat{\psi}_{rr}Y_{ri}+\hat{\psi}_{(r+1)r}X_i)}{1+\exp(\hat{\psi}_{0r}+\hat{\psi}_{1r}Y_{1i}+\ldots+\hat{\psi}_{rr}Y_{ri}+\hat{\psi}_{(r+1)r}X_i)},$$

$$K_r\{G_r(Z_i),\hat{\psi}_n\}=$$
$$\prod_{r'=1}^{r}\frac{1}{1+\exp(\hat{\psi}_{0r'}+\hat{\psi}_{1r'}Y_{1i}+\ldots+\hat{\psi}_{r'r'}Y_{r'i}+\hat{\psi}_{(r'+1)r'}X_i)},$$

and

$$\varpi(\infty,Z_i,\hat{\psi}_n)=K_{l-1}\{G_{l-1}(Z_i),\hat{\psi}_n\}.$$

The estimator $\hat{\beta}_n$, the solution to (9.26), will be an observed-data RAL estimator for β that is consistent and asymptotically normal, assuming, of course, that the model for the discrete hazard functions were correctly specified. An estimator for the asymptotic variance of the estimator for $\hat{\beta}_n$ can be obtained by using the sandwich variance estimator given by (9.19).

The efficiency of the estimator will depend on the choice for $m(Z_i,\beta)$ and the choice for $L_r(X_i,Y_{1i},\ldots,Y_{ri}),\ r=1,\ldots,l-1$. For illustration, we chose $m(Z_i,\beta)=H^T(X_i)\{Y_i-H(X_i)\beta\}$, but this may or may not be a good choice. Also, we did not discuss choices for $L_r(X_i,Y_{1i},\ldots,Y_{ri}),\ r=1,\ldots,l-1$. If $L_r(\cdot)$ were set equal to zero, then the corresponding estimator would be the IPWCC estimator. A more detailed discussion on choices for these functions and how they affect the efficiency of the resulting estimator will be given in Chapters 10 and 11. □

Since we are on the topic of monotone coarsening, we take this opportunity to note that in *survival analysis* the survival data are often right censored. The notion of right censoring was introduced in Section 5.2, where we derived semiparametric estimators for the proportional hazards model. Right censoring can be viewed as a specific example of monotone coarsening. That is, if data in a survival analysis are right censored, then we don't observe any data subsequent to the censoring time. The main difference between right censoring for survival data and the monotone coarsening presented thus far is that the censoring time for survival data is continuous, taking on an uncountably infinite number of values, whereas we have only considered coarsening models with a finite number of coarsened configurations. Nonetheless, we can make the analogy from monotone coarsening to censoring in survival analysis by

considering continuous-time hazard rates instead of discrete hazard probabilities.

In the next section, we give some of the analogous results for censored data. We show how to cast a censored-data problem as a monotone coarsening problem and we derive a typical influence function of an observed-data (censored data) influence function for a parameter β in terms of the influence function of a full-data (uncensored) estimator for β. Much of the exposition that follows is motivated by the work in the landmark paper of Robins and Rotnitzky (1992). To follow the results in the next section, the reader must be familiar with counting processes and the corresponding martingale processes used in an advanced course in censored survival analysis. If the reader does not have this background, then this section can be skipped without having it affect the reading of the remainder of the book.

9.3 Censoring and Its Relationship to Monotone Coarsening

In survival analysis, full data for a single individual can be summarized as $\{T, \bar{X}(T)\}$, where T denotes the survival time, $X(u)$ denotes the value of covariates (possibly time-dependent) measured at time u, and $\bar{X}(T)$ is the history of time-dependent covariates $X(u), u \leq T$. As always, the primary focus is to estimate parameters β that characterize important aspects of the distribution of $\{T, \bar{X}(T)\}$. We will assume that we know how to find estimators for the parameter of interest if we had full data $\{T_i, \bar{X}_i(T_i)\},\ i = 1, \ldots, n$.

Example 2. As an example, consider the following problem. Suppose we are interested in estimating the mean medical costs for patients with some illness during the duration of their illness. For patient i in our sample, let T_i denote the duration of their illness and let $X_i(u)$ denote the accumulated hospital costs incurred for patient i at time u (measured as the time from the beginning of illness). Clearly, $X_i(u)$ is a nondecreasing function of u, and the total medical cost for patient i is $X_i(T_i)$. The parameter of interest is given by $\beta = E\{X(T)\}$. If we make no assumptions regarding the joint distribution of $\{T, \bar{X}(T)\}$, we showed in Theorem 4.4 that the tangent space for this nonparametric model is the entire Hilbert space. Consequently, using arguments in Section 5.3, where we derived the influence function for the mean of a random variable under a nonparametric model, there can be at most one influence function of an RAL estimator for β. With a sample of iid data $\{T_i, \bar{X}_i(T_i)\},\ i = 1, \ldots, n$, the obvious estimator for the mean medical costs is

$$\hat{\beta}_n = n^{-1} \sum_{i=1}^{n} X_i(T_i),$$

which has influence function $X(T) - E\{X(T)\}$. If, however, the duration of illness is right censored for some individuals, then this problem becomes more

difficult. We will use this example for illustration as we develop censored data estimators. □

In actuality, we often don't observe the full data because of censoring, possibly because of incomplete follow-up of the patients due to staggered entry and finite follow-up, or because the patients drop out of the study prematurely. To accommodate censoring, we introduce a censoring variable $\tilde{C}$ that corresponds to the time at that an individual would be censored from the study.

Remark 3. Notational convention
Since we have been using $\mathcal{C}$ (scripted C) to denote the different levels of coarsened data and because censoring is typically denoted by the variable C (unscripted), the difference may be difficult to discern. Therefore, we will use $\tilde{C}$ (i.e., C with a tilde over it) to denote the censoring variable from here on. □

We assume that underlying any problem in survival analysis there are unobservable latent random variables

$$\{\tilde{C}_i, T_i, \bar{X}_i(T_i)\},\ i = 1, \ldots, n.$$

The joint distribution of $p_{\tilde{C},T,\bar{X}(T)}\{c, t, \bar{x}(t)\}$ can be written as

$$p_{\tilde{C}|T,\bar{X}(T)}\{c|t, \bar{x}(t)\}p_{T,\bar{X}(T)}\{t, \bar{x}(t)\},$$

where $p_{T,\bar{X}(T)}\{t, \bar{x}(t)\}$ denotes the density of the full data had we been able to observe them.

The observed (coarsened) data for this problem are denoted as

$$\{U, \Delta, \bar{X}(U)\},$$

where $U = \min(T, \tilde{C})$ (observed time on study), $\Delta = I(T \leq \tilde{C})$ (censoring indicator), and $\bar{X}(U)$ is the history of the time-dependent covariates while on study. Coarsening of the data in this case is related to the fact that we don't observe the full data because of censoring. We will show that censoring can be mapped to a form of monotone coarsening. The coarsening variable $\mathcal{C}$, which we took previously to be discrete, is now a continuous variable because of continuous-time censoring. A complete case, $\mathcal{C} = \infty$, corresponds to $\Delta = 1$ or, equivalently, to $(T \leq \tilde{C})$.

To make the connection between censoring and the coarsening notation used previously, we define $(\mathcal{C} = r)$ to be $(\tilde{C} = r,\ T > r)$ and, when $\mathcal{C} = r$, we observe $G_r\{T, \bar{X}(T)\} = \left[\bar{X}\{\min(r, T)\}, TI(T \leq r)\right]$ for $r < \infty$ and $G_\infty\{T, \bar{X}(T)\} = \{T, \bar{X}(T)\}$. With this notation

$$G_r\{T, \bar{X}(T)\} = G_\infty\{T, \bar{X}(T)\} = \{T, \bar{X}(T)\} \text{ whenever } r \geq T.$$

Nonetheless, this still satisfies all the assumptions of monotone coarsening.

Therefore, the observed data can be expressed as

$$[\mathcal{C}, G_{\mathcal{C}}\{T, \bar{X}(T)\},$$

where

$$[\mathcal{C} = r, G_r\{T, \bar{X}(T)\}] = \{\tilde{C} = r, T > r, \ \bar{X}(r)\}$$

and

$$[\mathcal{C} = \infty, G_\infty\{T, \bar{X}(T)\}] = \{T \leq \tilde{C}, T, \bar{X}(T)\}.$$

With monotone coarsening, we argued that it is more convenient to work with hazard functions in describing the coarsening probabilities. With a slight abuse of notation, the coarsening hazard function is given as

$$\lambda_r\{T, \bar{X}(T)\} = P\{\mathcal{C} = r | \mathcal{C} \geq r, T, \bar{X}(T)\}, \quad r < \infty. \tag{9.27}$$

Remark 4. Because the censoring variable $\tilde{C}$ is a continuous random variable, so is the corresponding coarsening variable $\mathcal{C}$. Therefore, the hazard function given by (9.27), which strictly speaking is used for discrete coarsening, has to be defined in terms of a continuous-time hazard function; namely,

$$\lambda_r\{T, \bar{X}(T)\} = \lim_{h \to 0} h^{-1} P\{r \leq \mathcal{C} =< r + h | \mathcal{C} \geq r, T, \bar{X}(T)\}, \quad r < \infty.$$

However, unless we need further clarification, it will be convenient to continue with this abuse of notation. □

The event $\mathcal{C} \geq r$, which includes $\mathcal{C} = \infty$, is equal to

$$(\mathcal{C} \geq r) = (\tilde{C} \geq r, T > \tilde{C}) \cup (T < \tilde{C}). \tag{9.28}$$

Therefore,

$$\lambda_r\{T, \bar{X}(T)\} = P[\tilde{C} = r, T \geq r | \{(\tilde{C} \geq r, T > \tilde{C}) \cup (T < \tilde{C})\}, T, \bar{X}(T)]. \tag{9.29}$$

If $T < r$, then (9.29) must equal zero, whereas if $T \geq r$, then $\{(\tilde{C} \geq r, T > \tilde{C}) \cup (T < \tilde{C})\} \cap (T \geq r) = (\tilde{C} \geq r)$. Consequently,

$$\lambda_r\{T, \bar{X}(T)\} = \underbrace{P\{\tilde{C} = r | \tilde{C} \geq r, T, \bar{X}(T)\}}_{\begin{array}{c}|| \\ \text{this is the hazard function of} \\ \text{censoring at time } r \text{ given } T, \bar{X}(T).\end{array}} I(T \geq r).$$

If, in addition, we make the coarsening at random (CAR) assumption,

$$\lambda_r\{T, \bar{X}(T)\} = \lambda_r[G_r\{T, \bar{X}(T)\}],$$

then

$$\lambda_r[G_r\{T, \bar{X}(T)\}] = P\{\tilde{C} = r | \tilde{C} \geq r, T \geq r, \bar{X}(r)\} I(T \geq r).$$

Let us denote the hazard function for censoring by

$$\lambda_{\tilde{C}}\{r, \bar{X}(r)\} = P\{\tilde{C} = r | \tilde{C} \geq r, T \geq r, \bar{X}(r)\}.$$

Then

$$\lambda_r[G_r\{T, \bar{X}(T)\}] = \lambda_{\tilde{C}}\{r, \bar{X}(r)\} I(T \geq r). \tag{9.30}$$

In order to construct estimators for a full-data parameter using coarsened data, such as those given by (9.4), we need to compute the probability of a complete case $\varpi[\infty, G_\infty\{T, \bar{X}(T)\}] = P\{\mathcal{C} = \infty | T, \bar{X}(T)\} = P\{\Delta = 1 | T, \bar{X}(T)\}$ and a typical element of the augmentation space, Λ_2. We now show how these are computed with censored data using the hazard function for the censoring time and counting process notation.

Probability of a Complete Case with Censored Data

For discrete monotone coarsening, we showed in (8.6) how the probability of a complete case $\mathcal{C} = \infty$ can be written in terms of the discrete hazards. For a continuous-time hazard function, the analogous relationship is given by

$$\begin{aligned}
\varpi[\infty, G_\infty\{T, \bar{X}(T)\}] = P\{\Delta = 1 | T, \bar{X}(T)\} &= \prod_{r<\infty} \Big(1 - \lambda_r[G_r\{T, \bar{X}(T)\}]dr\Big) \\
&= \exp\left\{ -\int_0^\infty \lambda_r[G_r\{T, \bar{X}(T)\}]dr \right\} \\
&= \exp\left\{ -\int_0^\infty \lambda_{\tilde{C}}\{r, \bar{X}(r)\} I(T \geq r) dr \right\} \\
&= \exp\left\{ -\int_0^T \lambda_{\tilde{C}}\{r, \bar{X}(r)\} dr \right\}.
\end{aligned} \tag{9.31}$$

The Augmentation Space, Λ_2, with Censored Data

In equation (9.20) of Theorem 9.2, we showed that an arbitrary element of the augmentation space, Λ_2, with monotone coarsening can be written as

$$\sum_{r \neq \infty} \frac{I(\mathcal{C} = r) - \lambda_r[G_r\{T, \bar{X}(T)\}] I(\mathcal{C} \geq r)}{K_r[G_r\{T, \bar{X}(T)\}]} L_r[G_r\{T, \bar{X}(T)\}]. \tag{9.32}$$

Using counting process notation, we denote the counting process corresponding to the number of observed censored observations up to and including time r by $N_{\tilde{C}}(r) = I(U \leq r, \Delta = 0) = I(\tilde{C} \leq r, T > \tilde{C})$. Consequently,

$$I(\mathcal{C} = r) = I(\tilde{C} = r, T > r) = dN_{\tilde{C}}(r),$$

where $dN_{\tilde{C}}(r)$ denotes the increment of the counting process. Using (9.30), we obtain

$$\lambda_r[G_r\{T, \bar{X}(T)\}]I(\mathcal{C} \geq r) = \lambda_{\tilde{C}}\{r, \bar{X}(r)\}I(T \geq r)I(\tilde{C} \geq r).$$

Letting $Y(r)$ denote the at-risk indicator, $Y(r) = I(U \geq r) = I(T \geq r, \tilde{C} \geq r)$, we obtain

$$\lambda_r[G_r\{T, \bar{X}(T)\}]I(\mathcal{C} \geq r) = \lambda_{\tilde{C}}\{r, \bar{X}(r)\}Y(r).$$

Because the elements in the sum of (9.32) are nonzero only if $T \geq r$ and $\tilde{C} \geq r$, it suffices to define $L_r[G_r\{T, \bar{X}(T)\}]$ and $K_r[G_r\{T, \bar{X}(T)\}]$ as $L_r\{\bar{X}(r)\}$ and $K_r\{\bar{X}(r)\}$, respectively. Moreover,

$$K_r\{\bar{X}(r)\} = \prod_{u \leq r} \left(1 - \lambda_{\tilde{C}}\{u, \bar{X}(u)\}du\right) = \exp\left[-\int_0^r \lambda_{\tilde{C}}\{u, \bar{X}(u)\}du\right]. \tag{9.33}$$

Consequently, a typical element of Λ_2, given by (9.32), can be written using stochastic integrals of counting process martingales, namely

$$\int_0^\infty \frac{dM_{\tilde{C}}\{r, \bar{X}(r)\}}{K_r\{\bar{X}(r)\}} L_r\{\bar{X}(r)\},$$

where

$$dM_{\tilde{C}}\{r, \bar{X}(r)\} = dN_{\tilde{C}}(r) - \lambda_{\tilde{C}}\{r, \bar{X}(r)\}Y(r)dr$$

is the usual counting process martingale increment and $K_r\{\bar{X}(r)\}$ is given by (9.33).

Deriving Estimators with Censored Data

Suppose we want to estimate the parameter β in a survival problem with censored data $\{U_i, \Delta_i, \bar{X}_i(U_i)\}$, $i = 1, \ldots, n$. Let us assume that we know how to estimate β if we had full data $Z_i = \{T_i, \bar{X}_i(T_i)\}$, $i = 1, \ldots, n$; say, for example, we had an unbiased estimating function $m(Z_i, \beta)$ that we could use as a basis for deriving an m-estimator. In addition, let us assume that we knew the hazard function for censoring, $\lambda_{\tilde{C}}\{r, \bar{X}(r)\}$. Since censoring is a special case of coarsened data, we could use (9.4) to obtain estimators for β with censored data.

Specifically, using the notation developed above, an AIPWCC estimator for β can be derived by solving the estimating equation

$$\sum_{i=1}^n \left[\frac{\Delta_i}{K_{U_i}\{\bar{X}_i(U_i)\}} m(Z_i, \beta) + \int_0^\infty \frac{dM_{\tilde{C}_i}\{r, \bar{X}_i(r)\}}{K_r\{\bar{X}_i(r)\}} L_r\{\bar{X}_i(r)\}\right] = 0, \tag{9.34}$$

where $K_r\{\bar{X}(r)\}$ is given by (9.33).

We now return to Example 2, where the goal was to estimate the mean medical costs of patients with some illness during the course of their illness. We defined the parameter of interest as $\beta = E\{X(T)\}$, where $X(T)$ was the

accumulated medical costs of a patient during the time T of his or her illness. As noted previously, with full data there is only one influence function of RAL estimators for β, namely $X(T) - \beta$. Consequently, the obvious full-data estimating function for this problem is $m(Z_i, \beta) = X_i(T_i) - \beta$. With censored data, we use (9.34) to derive an arbitrary estimator for β as the solution to

$$\sum_{i=1}^{n} \left[\frac{\Delta_i}{K_{U_i}\{\bar{X}_i(U_i)\}} \{X_i(U_i) - \beta\} + \int_0^\infty \frac{dM_{\tilde{C}_i}\{r, \bar{X}_i(r)\}}{K_r\{\bar{X}_i(r)\}} L_r\{\bar{X}_i(r)\} \right] = 0.$$

This leads to the estimator

$$\hat{\beta}_n = \frac{\sum_{i=1}^{n} \left[\frac{\Delta_i}{K_{U_i}\{\bar{X}_i(U_i)\}} X_i(U_i) + \int_0^\infty \frac{dM_{\tilde{C}_i}\{r, \bar{X}_i(r)\}}{K_r\{\bar{X}_i(r)\}} L_r\{\bar{X}_i(r)\} \right]}{\sum_{i=1}^{n} \frac{\Delta_i}{K_{U_i}\{\bar{X}_i(U_i)\}}}. \tag{9.35}$$

Notice that, unlike the case where there was only one influence function with the full data for this problem, there are many influence functions with censored data. The observed (censored) data influence functions depend on the choice of $L_r\{\bar{X}(r)\}$ for $r \geq 0$, which, in turn, affects the asymptotic variance of the corresponding estimator in (9.35). Clearly, we want to choose such functions in order to minimize the asymptotic variance of the corresponding estimator. This issue will be studied carefully in the next chapter.

We also note that the estimator given above assumes that we know the hazard function for censoring, $\lambda_{\tilde{C}}\{r, \bar{X}(r)\}$. In practice, this will be unknown to us and must be estimated from the observed data. If the censoring time $\tilde{C}$ is assumed independent of $\{T, \bar{X}(T)\}$, then one can estimate K_r using the Kaplan-Meier estimator for the censoring time $\tilde{C}$; see Kaplan and Meier (1958). If the censoring time is related to the time-dependent covariates, then a model has to be developed. A popular model for this purpose is Cox's proportional hazards model (Cox, 1972, 1975).

If $L_r(\cdot)$ is taken to be identically equal to zero, then we obtain the IPWCC estimator. This estimator, referred to as the simple weighted estimator for estimating the mean medical cost with censored cost data, was studied by Bang and Tsiatis (2000), who also derived the large-sample properties. More efficient estimators with a judicious choice of $L_r(\cdot)$ were also proposed in that paper.

9.4 Recap and Review of Notation

Constructing AIPWCC estimators

- Let $m(Z, \beta)$ be a full-data estimating function, chosen so that $m(Z, \beta_0) \in \Lambda^{F\perp}$.

- If coarsening probabilities are known by design (i.e., ψ_0 known), then an observed-data RAL AIPWCC estimator for β is obtained as the solution to the estimating equation

$$\sum_{i=1}^{n} \left[\frac{I(\mathcal{C}_i = \infty) m(Z_i, \beta)}{\varpi(\infty, Z_i, \psi_0)} + L_2\{\mathcal{C}_i, G_{\mathcal{C}_i}(Z_i), \psi_0\} \right] = 0,$$

and the i-th influence function of the resulting estimator $\hat{\beta}_n$ is

$$\tilde{\varphi}\{\mathcal{C}_i, G_{\mathcal{C}_i}(Z_i), \psi_0\} = \\ - \left[E\left\{ \frac{\partial m(Z_i, \beta_0)}{\partial \beta^T} \right\} \right]^{-1} \left[\frac{I(\mathcal{C}_i = \infty) m(Z_i, \beta_0)}{\varpi(\infty, Z_i, \psi_0)} + L_2\{\mathcal{C}_i, G_{\mathcal{C}_i}(Z_i), \psi_0\} \right],$$

where $L_2(\cdot)$ is an element of the augmentation space given by

$$\frac{I(\mathcal{C} = \infty)}{\varpi(\infty, Z, \psi_0)} \left[\sum_{r \neq \infty} \varpi\{r, G_r(Z), \psi_0\} L_{2r}\{G_r(Z)\} \right] \\ - \sum_{r \neq \infty} I(\mathcal{C} = r) L_{2r}\{G_r(Z)\},$$

for arbitrarily chosen functions $L_{2r}\{G_r(Z)\}, r \neq \infty$.

- If the coarsening probabilities need to be modeled (i.e., ψ is unknown), then an observed-data RAL AIPWCC estimator for β is obtained as the solution to the estimating equation

$$\sum_{i=1}^{n} \left[\frac{I(\mathcal{C}_i = \infty) m(Z_i, \beta)}{\varpi(\infty, Z_i, \hat{\psi}_n)} + L_2\{\mathcal{C}_i, G_{\mathcal{C}_i}(Z_i), \hat{\psi}_n\} \right] = 0,$$

where $\hat{\psi}_n$ denotes the MLE for ψ. The i-th influence function of the resulting estimator is

$$\tilde{\varphi}\{\mathcal{C}_i, G_{\mathcal{C}_i}(Z_i), \psi_0\} - E(\tilde{\varphi} S_\psi^T)\{E(S_\psi S_\psi^T)\}^{-1} S_\psi\{\mathcal{C}_i, G_{\mathcal{C}_i}(Z_i), \psi_0\} \\ = \tilde{\varphi}\{\mathcal{C}_i, G_{\mathcal{C}_i}(Z_i), \psi_0\} - \Pi[\tilde{\varphi}\{\mathcal{C}_i, G_{\mathcal{C}_i}(Z_i), \psi_0\} | \Lambda_\psi].$$

The augmentation space with monotone coarsening

- When the observed data are monotonically coarsened, then it will prove to be convenient to express the elements of the augmentation space using discrete hazard functions, specifically an element $L_2\{\mathcal{C}, G_{\mathcal{C}}(Z), \psi\} \in \Lambda_2$,

$$\sum_{r \neq \infty} \left[\frac{I(\mathcal{C} = r) - \lambda_r\{G_r(Z), \psi\} I(\mathcal{C} \geq r)}{K_r\{G_r(Z), \psi\}} \right] L_r\{G_r(Z)\},$$

for arbitrarily chosen functions $L_r\{G_r(Z)\}$ $r \neq \infty$.

9.5 Exercises for Chapter 9

1. In equation (9.4), we proposed an m-estimator (AIPWCC) for β when the coarsening probabilities are known by design.
   a) Derive the influence function for this estimator and demonstrate that it equals (9.5).
   b) Derive a consistent estimator for the asymptotic variance of this estimator.

10

Improving Efficiency and Double Robustness with Coarsened Data

Thus far, we have described the class of observed-data influence functions when data are coarsened at random (CAR) by taking advantage of results obtained for a full-data semiparametric model. We also illustrated how these results can be used to derive estimators using augmented inverse probability weighted complete-case estimating equations. The results were geometric, relying on our ability to define the spaces $\Lambda^{F\perp} \subset \mathcal{H}^F$ and $\Lambda_2 \subset \mathcal{H}$. Ultimately, the goal is to derive as efficient an estimator for β as is possible using coarsened data.

As we will see, this exercise will be primarily theoretical, as it will most often be the case that we cannot feasibly construct the most efficient estimator. Nonetheless, the study of efficiency with coarsened data will aid us in constructing more efficient estimators even if we are not able to derive the most efficient one.

10.1 Optimal Observed-Data Influence Function Associated with Full-Data Influence Function

We have already shown in (8.19) that all observed-data influence functions of RAL estimators for β can be written as

$$\varphi\{\mathcal{C}, G_{\mathcal{C}}(Z)\} = \left\{ \left[\frac{I(\mathcal{C}=\infty)\varphi^F(Z)}{\varpi(\infty, Z, \psi_0)} + L_2\{\mathcal{C}, G_{\mathcal{C}}(Z)\} \right] - \Pi\{[\,\cdot\,]|\Lambda_\psi\}, \quad (10.1) \right.$$
$$\left. \text{where } \varphi^F(Z) \text{ is a full-data influence function and } L_2\{\mathcal{C}, G_{\mathcal{C}}(Z)\} \in \Lambda_2 \right\}.$$

We know that the asymptotic variance of an RAL estimator for β is the variance of its influence function. Therefore, we now consider how to derive the optimal observed-data influence function within the class of influence functions (10.1) for a fixed full-data influence function $\varphi^F(Z) \in (IF)^F$, where optimal refers to the element with the smallest variance matrix.

Theorem 10.1. The optimal observed-data influence function among the class of observed-data influence functions given by (10.1) for a fixed $\varphi^F(Z) \in (IF)^F$ is obtained by choosing $L_2\{\mathcal{C}, G_{\mathcal{C}}(Z)\} = -\Pi[\frac{I(\mathcal{C}=\infty)\varphi^F(Z)}{\varpi(\infty,Z,\psi_0)}|\Lambda_2]$, in which case the optimal influence function is given by

$$\frac{I(\mathcal{C}=\infty)\varphi^F(Z)}{\varpi(\infty,Z,\psi_0)} - \Pi\left[\frac{I(\mathcal{C}=\infty)\varphi^F(Z)}{\varpi(\infty,Z,\psi_0)}\middle|\Lambda_2\right]. \tag{10.2}$$

Proof. We begin by noting that the space of elements in (10.1), for a fixed $\varphi^F(Z)$, is a linear variety as defined by Definition 7 of Chapter 3 (i.e., a translation of a linear space away from the origin). Specifically, this space is given as $V = x_0 + M$, where the element

$$x_0 = \frac{I(\mathcal{C}=\infty)\varphi^F(Z)}{\varpi(\infty,Z,\psi_0)} - \Pi\left[\frac{I(\mathcal{C}=\infty)\varphi^F(Z)}{\varpi(\infty,Z,\psi_0)}\middle|\Lambda_\psi\right]$$

and the linear subspace

$$M = \Pi[\Lambda_2|\Lambda_\psi^\perp].$$

To prove that this space is a linear variety, we must show that $x_0 \notin M$. By Theorem 8.1, we know that $\Lambda_\psi \subset \Lambda_2$. Therefore, it suffices to show that $x_0 \notin \Lambda_2$. This follows because $E(x_0|Z) = \varphi^F(Z) \neq 0$.

It is also straightforward to show that the linear space M is a q-replicating linear space as defined by Definition 6 of Chapter 3. (We leave this as an exercise for the reader.) Consequently, as a result of Theorem 3.3, the element in this linear variety with the smallest variance matrix is given as

$$x_0 - \Pi[x_0|M].$$

The theorem will follow if we can prove that

$$\Pi[x_0|M] = \Pi\left[\frac{I(\mathcal{C}=\infty)\varphi^F(Z)}{\varpi(\infty,Z,\psi_0)}\middle|\Lambda_2\right] - \Pi\left[\frac{I(\mathcal{C}=\infty)\varphi^F(Z)}{\varpi(\infty,Z,\psi_0)}\middle|\Lambda_\psi\right]. \tag{10.3}$$

In order to prove that (10.3) is a projection, we must show that

(a) (10.3) is an element of M and
(b) $x_0 - \Pi[x_0|M]$ is orthogonal to M.

(a) follows because $\Pi\left[\frac{I(\mathcal{C}=\infty)\varphi^F(Z)}{\varpi(\infty,Z,\psi_0)}\middle|\Lambda_2\right] \in \Lambda_2$ and

$$\Pi\left[\frac{I(\mathcal{C}=\infty)\varphi^F(Z)}{\varpi(\infty,Z,\psi_0)}\middle|\Lambda_\psi\right] = \Pi\left(\Pi\left[\frac{I(\mathcal{C}=\infty)\varphi^F(Z)}{\varpi(\infty,Z,\psi_0)}\middle|\Lambda_2\right]\middle|\Lambda_\psi\right), \tag{10.4}$$

where (10.4) follows because $\Lambda_\psi \subset \Lambda_2$.

(b) follows because

$$x_0 - \Pi[x_0|M] = \frac{I(\mathcal{C}=\infty)\varphi^F(Z)}{\varpi(\infty,Z,\psi_0)} - \Pi\left[\frac{I(\mathcal{C}=\infty)\varphi^F(Z)}{\varpi(\infty,Z,\psi_0)}\middle|\Lambda_2\right],$$

which is an element orthogonal to Λ_2 and hence orthogonal to M because $\Lambda_\psi \subset \Lambda_2$. □

Remark 1. At the end of Section 9.1, we made the observation that the asymptotic variance of an observed-data estimator for β would be more efficient if the parameter ψ in a model for the coarsening probabilities were estimated even if in fact this parameter was known by design. This stemmed from the fact that estimating the parameter ψ resulted in an influence function that subtracted off the projection of the term in the square brackets of (10.1) onto Λ_ψ. However, we now realize that if we constructed the estimator for β as efficiently as possible when ψ is known (that is, if we chose $L_2\{\mathcal{C}, G_{\mathcal{C}}(Z)\} = -\Pi[\frac{I(\mathcal{C}=\infty)\varphi^F(Z)}{\varpi(\infty,Z,\psi_0)}|\Lambda_2]$), then the term in the square brackets of (10.1) would equal

$$\frac{I(\mathcal{C}=\infty)\varphi^F(Z)}{\varpi(\infty,Z,\psi_0)} - \Pi\left[\frac{I(\mathcal{C}=\infty)\varphi^F(Z)}{\varpi(\infty,Z,\psi_0)}\middle|\Lambda_2\right],$$

which is orthogonal to Λ_2 and hence orthogonal to Λ_ψ, in which case the additional projection onto Λ_ψ that comes from estimating ψ would equal zero and therefore would not result in any gain in efficiency. □

A linear operator was defined in Chapter 7 (see Definition 1). It will be convenient to define the mapping from a full-data influence function to the corresponding optimal observed-data influence function given by Theorem 10.1 using a linear operator.

Definition 1. The linear operator $\mathcal{J} : \mathcal{H}^F \to \mathcal{H}$ is defined so that for any element $h^F(Z) \in \mathcal{H}^F$,

$$\mathcal{J}(h^F) = \frac{I(\mathcal{C}=\infty)h^F(Z)}{\varpi(\infty,Z,\psi_0)} - \Pi\left[\frac{I(\mathcal{C}=\infty)h^F(Z)}{\varpi(\infty,Z,\psi_0)}\middle|\Lambda_2\right]. \quad \square \tag{10.5}$$

Using this definition, we note that the optimal observed-data influence function within the class (10.1), for a fixed $\varphi^F(Z) \in (IF)^F$, is given by $\mathcal{J}(\varphi^F)$. Since any observed-data influence function of an RAL estimator for β must be an element within the class (10.1) for some $\varphi^F(Z) \in (IF)^F$, if we want to find the efficient influence function, it suffices to restrict attention to the class of influence functions $\mathcal{J}(\varphi^F)$ for $\varphi^F(Z) \in (IF)^F$. We define the space of such influence functions as follows.

Definition 2. We denote the space $\mathcal{J}\{(IF)^F\}$, the space whose elements are

$$\left\{\mathcal{J}(\varphi^F) \text{ for all } \varphi^F(Z) \in (IF)^F\right\},$$

by $(IF)_{\text{DR}}$. □

The subscript "DR" is used to denote double robustness. Therefore, the space $(IF)_{\text{DR}} \subset (IF)$ is defined as the set of double-robust observed-data influence functions. The term double robust was first introduced in Section 6.5. Why we refer to these as double-robust influence functions will become clear later in the chapter. Since the space of full-data influence functions is a linear variety in $\mathcal{H}^F$ (see Theorem 4.3) $(IF)^F = \varphi^F(Z) + \mathcal{T}^{F\perp}$, where $\varphi^F(Z)$ is an arbitrary full-data influence function and $\mathcal{T}^F$ is the full-data tangent space, it is clear that $(IF)_{\text{DR}} = \mathcal{J}\{(IF)^F\} = \mathcal{J}(\varphi^F) + \mathcal{J}(\mathcal{T}^{F\perp})$ is a linear variety in $\mathcal{H}$.

Remark 2. As we have argued repeatedly, because an influence function of an observed-data RAL estimator for β can be obtained, up to a proportionality constant, from an element orthogonal to the observed-data nuisance tangent space, this has motivated us to define estimating functions $m\{\mathcal{C}, G_{\mathcal{C}}(Z), \beta\}$, where $m\{\mathcal{C}, G_{\mathcal{C}}(Z), \beta_0\} \in \Lambda^{\perp}$.

If, however, we are interested in deriving observed-data RAL estimators for β whose influence function belongs to $(IF)_{\text{DR}}$, then we should choose an estimating function $m\{\mathcal{C}, G_{\mathcal{C}}(Z), \beta\}$ so that $m\{\mathcal{C}, G_{\mathcal{C}}(Z), \beta_0\}$ is an element of the linear space $\mathcal{J}(\Lambda^{F\perp})$. This follows because it suffices to define estimating functions that at the truth are proportional to influence functions. Since any element $\varphi^{*F}(Z) \in \Lambda^{F\perp}$ properly normalized will lead to a full-data influence function $\varphi^F(Z) = A^{q\times q}\varphi^{*F}(Z)$, where $A = \{E(\varphi^{*F} S_\beta^T)\}^{-1}$, and because $\mathcal{J}(\cdot)$ is a linear operator, this implies that a typical element $\mathcal{J}(\varphi^F)$ of $(IF)_{\text{DR}}$ is equal to $\mathcal{J}(A\varphi^{*F}) = A\mathcal{J}(\varphi^{*F})$; i.e., it is proportional to an element $\mathcal{J}(\varphi^{*F}) \in \mathcal{J}(\Lambda^{F\perp})$. We define the space $\mathcal{J}(\Lambda^{F\perp})$ to be the DR linear space. □

Definition 3. The linear subspace $\mathcal{J}(\Lambda^{F\perp}) \subset \Lambda^{\perp} \subset \mathcal{H}$, the space that consists of elements

$$\left\{\mathcal{J}(\varphi^{*F}) : \varphi^{*F}(Z) \in \Lambda^{F\perp}\right\},$$

will be referred to as the DR linear space. □

This now gives us a prescription for how to find observed-data RAL estimators for β whose influence function belongs to $(IF)_{\text{DR}}$. We start by choosing a full-data estimating function $m(Z, \beta)$ so that $m(Z, \beta_0) = \varphi^{*F}(Z) \in \Lambda^{F\perp}$. We then construct the observed-data estimating function $m\{\mathcal{C}, G_{\mathcal{C}}(Z), \beta\} = \mathcal{J}\{m(Z, \beta)\}$, where

$$\mathcal{J}\{m(Z,\beta)\} = \frac{I(\mathcal{C}=\infty)m(Z,\beta)}{\varpi(\infty, Z, \psi_0)} - \Pi\left[\frac{I(\mathcal{C}=\infty)m(Z,\beta)}{\varpi(\infty, Z, \psi_0)}\middle|\Lambda_2\right].$$

If the observed data were coarsened by design (i.e., ψ_0 known), then we would derive an estimator for β by solving the estimating equation

$$\sum_{i=1}^{n}\left[\frac{I(\mathcal{C}_i=\infty)m(Z_i,\beta)}{\varpi(\infty, Z_i, \psi_0)} + L_2\{\mathcal{C}_i, G_{\mathcal{C}_i}(Z_i), \beta, \psi_0\}\right] = 0, \tag{10.6}$$

where

$$L_2\{\mathcal{C}, G_{\mathcal{C}}(Z), \beta, \psi\} = -\Pi\left[\frac{I(\mathcal{C}=\infty)m(Z,\beta)}{\varpi(\infty, Z, \psi)}\middle|\Lambda_2\right].$$

If the coarsening probabilities were not known and had to be modeled using the unknown parameter ψ, then we would derive an estimator for β by solving the estimating equation

$$\sum_{i=1}^{n}\left[\frac{I(\mathcal{C}_i=\infty)m(Z_i,\beta)}{\varpi(\infty, Z_i, \hat{\psi}_n)} + L_2\{\mathcal{C}_i, G_{\mathcal{C}_i}(Z_i), \beta, \hat{\psi}_n\}\right] = 0, \tag{10.7}$$

where $\hat{\psi}_n$ is the MLE for the parameter ψ and $L_2(\cdot)$ is defined as above.

Finding projections onto the augmentation space Λ_2 is not necessarily easy. Later we will discuss a general procedure for finding such projections that involves an iterative process. However, in the case when there are two levels of coarsening or when the coarsening is monotone, a closed-form solution for the projection onto Λ_2 exists. We will study these two scenarios more carefully. We start by illustrating how to find improved estimators using these results when there are two levels of missingness.

10.2 Improving Efficiency with Two Levels of Missingness

Let the data for a single observation Z be partitioned as $(Z_1^T, Z_2^T)^T$, where we always observe Z_1 but Z_2 may be missing for some individuals. For this problem, it is convenient to use R to denote the complete-case indicator; that is, if $R = 1$, we observe $Z = (Z_1^T, Z_2^T)^T$, whereas if $R = 0$ we only observe Z_1. The observed data are given by $O_i = (R_i, Z_{1i}, R_i Z_{2i}), i = 1, \ldots, n$.

The assumption of missing at random implies that $P(R = 0|Z_1, Z_2) = P(R = 0|Z_1)$, which, in turn, implies that $P(R = 1|Z_1, Z_2) = P(R = 1|Z_1)$, which we denote by $\pi(Z_1)$. Such complete-case probabilities may be known to us by design or may have to be estimated using a model for the missingness probabilities that includes the additional parameter ψ. In the latter case, since R is binary, a logistic regression model is often used; for example,

$$\pi(Z_1, \psi) = \frac{\exp(\psi_0 + \psi_1^T Z_1)}{1 + \exp(\psi_0 + \psi_1^T Z_1)}. \tag{10.8}$$

The parameter $\psi = (\psi_0, \psi_1^T)^T$ can be estimated using a maximum likelihood estimator; that is, by maximizing

$$\prod_{i=1}^{n}\frac{\exp\{(\psi_0 + \psi_1^T Z_{1i})R_i\}}{1 + \exp(\psi_0 + \psi_1^T Z_{1i})} \tag{10.9}$$

using the data $(R_i, Z_{1i}), i = 1, \ldots, n$, which are available on everyone. In equation (10.8), we used a logistic regression model that was linear in Z_1;

however, we can make the model as flexible as necessary to fit the data. For example, we can include higher-order polynomial terms, interaction terms, splines, etc.

As always, there is an underlying full-data model $Z \sim p(z, \beta, \eta) \in \mathscr{P}$, where β is the q-dimensional parameter of interest and η is the nuisance parameter (possibly infinite-dimensional), and our goal is to estimate β using the observed data $O_i = (R_i, Z_{1i}, R_i Z_{2i}), i = 1, \ldots, n$.

In order to use either estimating equation (10.6) or (10.7), when ψ is known or unknown, respectively, to obtain an observed-data RAL estimator for β whose influence function is an element of $(IF)_{\text{DR}}$, we must find the projection of $\frac{I(\mathcal{C}=\infty)\varphi^{*F}(Z)}{\varpi(\infty,Z)}$ onto the augmentation space Λ_2, where $\varphi^{*F}(Z) = m(Z, \beta_0) \in \Lambda^{F\perp}$. Using the notation for two levels of missingness, we now consider how to derive the projection of $\frac{R\varphi^{*F}(Z)}{\pi(Z_1)}$ onto the augmentation space. According to (7.40) of Chapter 7, we showed that, with two levels of missingness, Λ_2 consists of the set of elements

$$L_2(O) = \left\{ \frac{R - \pi(Z_1)}{\pi(Z_1)} \right\} h_2(Z_1),$$

where $h_2^{q\times 1}(Z_1)$ is an arbitrary q-dimensional function of Z_1.

Finding the Projection onto the Augmentation Space

Theorem 10.2. The projection of $\frac{R\varphi^{*F}(Z)}{\pi(Z_1)}$ onto the augmentation space Λ_2 is the unique element $\left\{ \frac{R-\pi(Z_1)}{\pi(Z_1)} \right\} h_2^0(Z_1) \in \Lambda_2$, where

$$h_2^0(Z_1) = E\{\varphi^{*F}(Z)|Z_1\}. \tag{10.10}$$

Proof. The projection of $\frac{R\varphi^{*F}(Z)}{\pi(Z_1)}$ onto the space Λ_2 is the unique element $\left\{ \frac{R-\pi(Z_1)}{\pi(Z_1)} \right\} h_2^0(Z_1) \in \Lambda_2$ such that the residual

$$\left[\frac{R\varphi^{*F}(Z)}{\pi(Z_1)} - \left\{ \frac{R - \pi(Z_1)}{\pi(Z_1)} \right\} h_2^0(Z_1) \right]$$

is orthogonal to every element in Λ_2; that is,

$$E\left(\left[\frac{R\varphi^{*F}(Z)}{\pi(Z_1)} - \left\{ \frac{R - \pi(Z_1)}{\pi(Z_1)} \right\} h_2^0(Z_1) \right]^T \left\{ \frac{R - \pi(Z_1)}{\pi(Z_1)} \right\} h_2(Z_1) \right) = 0 \tag{10.11}$$

for all functions $h_2(Z_1)$. We derive the expectation in (10.11) by using the law of iterated conditional expectations, where we first condition on $Z = (Z_1^T, Z_2^T)^T$ to obtain

$$E\Bigg\{\Bigg(E\Bigg[\frac{R}{\pi(Z_1)}\left\{\frac{R-\pi(Z_1)}{\pi(Z_1)}\right\}\Bigg|Z\Bigg]\varphi^{*F}(Z) - E\Bigg[\left\{\frac{R-\pi(Z_1)}{\pi(Z_1)}\right\}^2\Bigg|Z\Bigg]h_2^0(Z_1)\Bigg)^T h_2(Z_1)\Bigg\}. \tag{10.12}$$

Because

$$E\Bigg[\frac{R\{R-\pi(Z_1)\}}{\pi^2(Z_1)}\Bigg|Z\Bigg] = E\Bigg[\left\{\frac{R-\pi(Z_1)}{\pi(Z_1)}\right\}^2\Bigg|Z\Bigg] = \frac{1-\pi(Z_1)}{\pi(Z_1)},$$

we write (10.12) as

$$E\Bigg[\left\{\frac{1-\pi(Z_1)}{\pi(Z_1)}\right\}\left\{\varphi^{*F}(Z) - h_2^0(Z_1)\right\}^T h_2(Z_1)\Bigg]. \tag{10.13}$$

Therefore, we must find the function $h_2^0(Z_1)$ such that (10.13) is equal to zero for all $h_2(Z_1)$. We derive (10.13) by again using the law of iterated conditional expectations, where we first condition on Z_1 to obtain

$$E\Bigg(\left\{\frac{1-\pi(Z_1)}{\pi(Z_1)}\right\}\Big[E\{\varphi^{*F}(Z)|Z_1\} - h_2^0(Z_1)\Big]^T h_2(Z_1)\Bigg). \tag{10.14}$$

By assumption, $\pi(Z_1) > \epsilon$ for all Z_1, which implies that $\frac{1-\pi(Z_1)}{\pi(Z_1)}$ is bounded away from zero and ∞ for all Z_1. Therefore, equation (10.14) will be equal to zero for all $h_2(Z_1)$ if and only if

$$h_2^0(Z_1) = E\{\varphi^{*F}(Z)|Z_1\}. \ \square$$

Thus, among estimators whose estimating functions are based on elements of $\Lambda^{\perp}$,

$$\Bigg[\frac{R\varphi^{*F}(Z)}{\pi(Z_1)} - \left\{\frac{R-\pi(Z_1)}{\pi(Z_1)}\right\}h_2(Z_1)\Bigg] - \Pi\{[\cdot]|\Lambda_\psi\},$$

for a fixed $\varphi^{*F}(Z) \in \Lambda^{F\perp}$, the one that gives the optimal answer (i.e., the estimator with the smallest asymptotic variance) is obtained by choosing $h_2(Z_1) = E\{\varphi^{*F}(Z)|Z_1\}$.

This result, although interesting, is not of practical use since we don't know the conditional expectation $E\{\varphi^{*F}(Z)|Z_1\}$. The result does, however, suggest that we consider an adaptive approach where we use the data to estimate this conditional expectation, as we now illustrate.

Adaptive Estimation

Since $Z = (Z_1^T, Z_2^T)^T$, in order to compute the conditional expectation of $E\{\varphi^{*F}(Z)|Z_1\}$, we need to know the conditional density of Z_2 given Z_1. Of

course, this conditional distribution depends on the unknown full-data parameters β and η. As we have argued, estimating β and η using likelihood methods for semiparametric models (i.e., when the parameter η is infinite-dimensional) may be difficult, if not impossible, using coarsened data. This is the reason we introduced AIPWCC estimators in the first place. Another approach, which we now advocate, is to be adaptive. That is, we posit a working model where the density of Z is assumed to be within the model $p^*_{Z_1,Z_2}(z_1,z_2,\xi) \in \mathscr{P}_\xi \subset \mathscr{P}$, where the parameter ξ is finite-dimensional. The conditional density of Z_2 given Z_1 as a function of the parameter ξ would be

$$p^*_{Z_2|Z_1}(z_2|z_1,\xi) = \frac{p^*_{Z_1,Z_2}(z_1,z_2,\xi)}{\int p^*_{Z_1,Z_2}(z_1,u,\xi)d\nu_{Z_2}(u)}.$$

For such a model, the parameter ξ can be estimated by maximizing the observed-data likelihood (7.10), which, with two levels of missingness, can be written as

$$\prod_{i=1}^{n} p^*_{Z_1,Z_2}(Z_{1i},Z_{2i},\xi)^{R_i}\left\{\int p^*_{Z_1,Z_2}(Z_{1i},u,\xi)d\nu_{Z_2}(u)\right\}^{1-R_i}. \tag{10.15}$$

Since we only need the conditional distribution of Z_2 given Z_1 to derive the desired projection, an even simpler approach would be to posit a parametric model for the conditional density of Z_2 given Z_1 in terms of a parametric model with a finite number of parameters ξ. Let us denote such a posited model by $p^*_{Z_2|Z_1}(z_2|z_1,\xi)$. Because of the missing at random (MAR) assumption, R is conditionally independent of Z_2 given Z_1. This is denoted as $R \perp\!\!\!\perp Z_2|Z_1$. Consequently,

$$p_{Z_2|Z_1,R}(z_2|z_1,r) = p_{Z_2|Z_1,R=1}(z_2|z_1,r=1) = p_{Z_2|Z_1}(z_2|z_1).$$

Therefore, it suffices to consider only the complete cases $\{i : R_i = 1\}$ because the conditional distribution of Z_2 given Z_1 among the complete cases is the same as the conditional distribution of Z_2 given Z_1 in the population. A natural estimator for ξ would be to maximize the conditional likelihood of Z_2 given Z_1 in ξ among the complete cases. Namely, we would estimate ξ by maximizing

$$\prod_{i:R_i=1} p^*_{Z_2|Z_1}(z_{2i}|z_{1i},\xi). \tag{10.16}$$

The resulting estimator is denoted by $\hat{\xi}^*_n$.

Remark 3. Whether we posit a simpler model for the density of $Z = (Z_1^T, Z_2^T)^T$ or the conditional density of Z_2 given Z_1, we must keep in mind that this is a posited model and that the true conditional density $p_{0_{Z_2|Z_1}}(z_2|z_1)$ may not be contained in this model. Moreover, if we develop a model directly for the conditional density of Z_2 given Z_1, then we must be careful that such a model is consistent with the underlying semiparametric model. □

Nonetheless, with the adaptive approach, we proceed as if the posited model were correct and estimate ξ using the observed data. Under suitable regularity conditions, this estimator will converge in probability to some constant ξ^*; i.e., $\hat{\xi}_n^* \xrightarrow{P} \xi^*$ and $n^{1/2}(\hat{\xi}_n^* - \xi^*)$ will be bounded in probability. In general, the posited model will not contain the truth, in which case

$$p^*_{Z_2|Z_1}(z_2|z_1, \xi^*) \neq p_{0_{Z_2|Z_1}}(z_2|z_1).$$

If, however, our posited model did contain the truth, then we denote this by taking ξ^* to equal ξ_0, where

$$p^*_{Z_2|Z_1}(z_2|z_1, \xi_0) = p_{0_{Z_2|Z_1}}(z_2|z_1).$$

With such a posited model for the conditional density of Z_2 given Z_1 and an estimator $\hat{\xi}_n^*$, we are able to estimate $h_2^0(Z_1) = E\{\varphi^{*F}(Z)|Z_1\}$ by using

$$h_2^*(Z_1, \hat{\xi}_n^*) = \int \varphi^{*F}(Z_1, u) p^*_{Z_2|Z_1}(u|Z_1, \hat{\xi}_n^*) d\nu_{Z_2}(u).$$

Again, keep in mind that

$$h_2^*(Z_1, \hat{\xi}_n^*) \to h_2^*(Z_1, \xi^*) = \int \varphi^{*F}(Z_1, u) p^*_{Z_2|Z_1}(u|Z_1, \xi^*) d\nu_{Z_2}(u),$$

where $h_2^*(Z_1, \xi^*)$ is a function of Z_1 but not necessarily that $h_2^0(Z_1) = h_2^*(Z_1, \xi^*)$ unless the posited model for the conditional density of Z_2 given Z_1 was correct.

With this as background, we now give a step-by-step algorithm on how to derive an improved estimator. In so doing, we consider the scenario where the parameter ψ in our missingness model is unknown and must be estimated.

Algorithm for Finding Improved Estimators with Two Levels of Missingness

1. We first consider how the parameter β would be estimated if there were no missing data (i.e., the full-data problem). That is, we choose an estimating function, say $m(Z, \beta)$, where $m(Z, \beta_0) = \varphi^{*F}(Z) \in \Lambda^{F\perp}$. We might consider using an estimating function that leads to an efficient or locally efficient full-data estimator for β. However, we must keep in mind that the efficient full-data influence function may not be the one that leads to an efficient observed-data influence function. A detailed discussion of this issue will be deferred until the next chapter.
2. We posit a model for the complete-case (missingness) probabilities in terms of the parameter ψ, say $P(R = 1|Z) = \pi(Z_1, \psi)$, and using the data $(R_i, Z_{1i}), i = 1, \ldots, n$, which are available on the entire sample, we derive the maximum likelihood estimator $\hat{\psi}_n$ for ψ. For example, we might use the

logistic regression model in (10.8), in which case the estimator is obtained by maximizing (10.9). In general, however, we would maximize

$$\prod_{i=1}^{n}\{\pi(Z_{1i},\psi)\}^{R_i}\{1-\pi(Z_{1i},\psi)\}^{1-R_i}.$$

We denote the MLE by $\hat{\psi}_n$.

3. We posit a model for either the distribution of $Z = (Z_1^T, Z_2^T)^T$ or the conditional distribution of Z_2 given Z_1 in terms of the parameter ξ. Either way, this results in a model for $p^*_{Z_2|Z_1}(z_2|z_1,\xi)$ in terms of the parameter ξ. Using the observed data, we derive the MLE $\hat{\xi}^*_n$ for ξ by maximizing either (10.15) or (10.16).
4. The estimator for β is obtained by solving the estimating equation

$$\sum_{i=1}^{n}\left[\frac{R_i m(Z_i,\beta)}{\pi(Z_{1i},\hat{\psi}_n)} - \left\{\frac{R_i-\pi(Z_{1i},\hat{\psi}_n)}{\pi(Z_{1i},\hat{\psi}_n)}\right\}h_2^*(Z_{1i},\beta,\hat{\xi}^*_n)\right] = 0, \qquad (10.17)$$

where

$$\begin{aligned} h_2^*(Z_{1i},\beta,\xi) &= E\left\{m(Z_i,\beta)|Z_{1i},\xi\right\} \\ &= \int m(Z_{1i},u,\beta)p^*_{Z_2|Z_1}(u|Z_{1i},\xi)d\nu_{Z_2}(u). \end{aligned}$$

Remarks Regarding Adaptive Estimators

The semiparametric theory that we have developed for coarsened data implicitly assumes that the model for the coarsening probabilities is correctly specified. With two levels of missingness, this means that $P(R=1|Z_1)=\pi_0(Z_1)$ is contained within the model $\pi(Z_1,\psi)$, in which case, under suitable regularity conditions, $\pi(Z_1,\hat{\psi}_n)\to\pi_0(Z_1)$, where $\hat{\psi}_n$ denotes the MLE for ψ. The fact that we used the augmented term

$$-\left\{\frac{R_i-\pi(Z_{1i},\hat{\psi}_n)}{\pi(Z_{1i},\hat{\psi}_n)}\right\}h_2^*(Z_{1i},\beta,\hat{\xi}^*_n) \qquad (10.18)$$

in equation (10.17) was an attempt to gain efficiency from the data that are incomplete (i.e., $\{i : R_i = 0\}$). To get the greatest gain in efficiency, the augmented term must equal

$$-\left\{\frac{R_i-\pi(Z_{1i},\hat{\psi}_n)}{\pi(Z_{1i},\hat{\psi}_n)}\right\}h_2^0(Z_{1i},\beta),$$

where

$$h_2^0(Z_{1i}, \beta) = E_0\{m(Z_i, \beta)|Z_{1i}\}, \tag{10.19}$$

where the conditional expectation on the right-hand side of (10.19) is with respect to the true conditional density of Z_2 given Z_1.

Because we are using a posited model, the function $h_2^*(Z_{1i}, \beta, \hat{\xi}_n^*)$ will converge to $h_2^*(Z_{1i}, \beta, \xi^*)$, which is not equal to the desired $h_2^0(Z_{1i}, \beta)$ unless the posited model was correct. Nonetheless, $h_2^*(Z_{1i}, \beta, \xi^*)$ is a function of Z_{1i}, in which case

$$-\left\{\frac{R_i - \pi(Z_{1i}, \psi_0)}{\pi(Z_{1i}, \psi_0)}\right\} h_2^*(Z_{1i}, \beta_0, \xi^*) \in \Lambda_2.$$

By Theorem 9.1, the solution to the estimating equation

$$\sum_{i=1}^{n}\left[\frac{R_i m(Z_i, \beta)}{\pi(Z_{1i}, \hat{\psi}_n)} - \left\{\frac{R_i - \pi(Z_{1i}, \hat{\psi}_n)}{\pi(Z_{1i}, \hat{\psi}_n)}\right\} h_2^*(Z_{1i}, \beta_0, \xi^*)\right] = 0, \tag{10.20}$$

where notice that we take $\beta = \beta_0$ and fix ξ^* in $h_2^*(\cdot)$, is an AIPWCC estimator for β whose influence function is

$$-\left[E\left\{\frac{\partial m(Z, \beta_0)}{\partial \beta^T}\right\}\right]^{-1} \times$$
$$\left\{\left[\frac{R m(Z, \beta_0)}{\pi(Z_1, \psi_0)} - \left\{\frac{R - \pi(Z_1, \psi_0)}{\pi(Z_1, \psi_0)}\right\} h_2^*(Z_1, \beta_0, \xi^*)\right] - \Pi\Big([\,]|\Lambda_\psi\Big)\right\}. \tag{10.21}$$

Let us denote the estimator that solves (10.20) by $\hat{\beta}_n^*$.

In the following theorem and proof, we give a heuristic justification to show that estimating ξ using $\hat{\xi}_n^*$ will have a negligible effect on the resulting estimator. That is, $\hat{\beta}_n$, the solution to (10.17), is asymptotically equivalent to $\hat{\beta}_n^*$, the solution to (10.20). By so doing, we deduce that the adaptive estimator $\hat{\beta}_n$ is a consistent, asymptotically normal estimator for β whose influence function is given by (10.21).

Theorem 10.3.

$$n^{1/2}(\hat{\beta}_n - \hat{\beta}_n^*) \xrightarrow{P} 0.$$

Proof. If we return to the proof of Theorem 9.1, we note that the asymptotic approximation given by (9.13), applied to the estimator $\hat{\beta}_n$, the solution to (10.17) would yield

$$n^{1/2}(\hat{\beta}_n - \beta_0) = -\left[E\left\{\frac{\partial m(Z, \beta_0)}{\partial \beta^T}\right\}\right]^{-1} \times$$
$$n^{-1/2}\sum_{i=1}^{n}\left(\left[\frac{R_i m(Z_i, \beta_0)}{\pi(Z_{1i}, \hat{\psi}_n)} - \left\{\frac{R_i - \pi(Z_{1i}, \hat{\psi}_n)}{\pi(Z_{1i}, \hat{\psi}_n)}\right\} h_2^*(Z_{1i}, \hat{\beta}_n, \hat{\xi}_n^*)\right]\right) + o_p(1), \tag{10.22}$$

whereas, applied to $\hat{\beta}_n^*$, the solution to (10.20) would yield

$$n^{1/2}(\hat{\beta}_n^* - \beta_0) = -\left[E\left\{\frac{\partial m(Z,\beta_0)}{\partial \beta^T}\right\}\right]^{-1} \times$$
$$n^{-1/2}\sum_{i=1}^{n}\left(\left[\frac{R_i m(Z_i,\beta_0)}{\pi(Z_{1i},\hat{\psi}_n)} - \left\{\frac{R_i - \pi(Z_{1i},\hat{\psi}_n)}{\pi(Z_{1i},\hat{\psi}_n)}\right\} h_2^*(Z_{1i},\beta_0,\xi^*)\right]\right) + o_p(1). \tag{10.23}$$

Taking differences between (10.22) and (10.23), we obtain that

$$n^{1/2}(\hat{\beta}_n - \hat{\beta}_n^*) = \Bigg[n^{-1/2}\sum_{i=1}^{n}\left\{\frac{R_i - \pi(Z_{1i},\hat{\psi}_n)}{\pi(Z_{1i},\hat{\psi}_n)}\right\} h_2^*(Z_{1i},\beta_0,\xi^*)$$
$$- n^{-1/2}\sum_{i=1}^{n}\left\{\frac{R_i - \pi(Z_{1i},\hat{\psi}_n)}{\pi(Z_{1i},\hat{\psi}_n)}\right\} h_2^*(Z_{1i},\hat{\beta}_n,\hat{\xi}_n^*)\Bigg] + o_p(1).$$

The proof is complete if we can show that the term in the square brackets above converges in probability to zero. This follows by expanding

$$n^{-1/2}\sum_{i=1}^{n}\left\{\frac{R_i - \pi(Z_{1i},\hat{\psi}_n)}{\pi(Z_{1i},\hat{\psi}_n)}\right\} h_2^*(Z_{1i},\hat{\beta}_n,\hat{\xi}_n^*)$$

about β_0 and ξ^*, while keeping $\hat{\psi}_n$ fixed, to obtain

$$n^{-1/2}\sum_{i=1}^{n}\left\{\frac{R_i - \pi(Z_{1i},\hat{\psi}_n)}{\pi(Z_{1i},\hat{\psi}_n)}\right\} h_2^*(Z_{1i},\beta_0,\xi^*) \tag{10.24}$$
$$+\left[n^{-1}\sum_{i=1}^{n}\left\{\frac{R_i - \pi(Z_{1i},\hat{\psi}_n)}{\pi(Z_{1i},\hat{\psi}_n)}\right\}\left\{\frac{\partial h_2^*(Z_{1i},\beta_n^*,\xi_n^*)}{\partial \beta^T}\right\}\right] n^{1/2}(\hat{\beta}_n - \beta_0) \tag{10.25}$$
$$+\left[n^{-1}\sum_{i=1}^{n}\left\{\frac{R_i - \pi(Z_{1i},\hat{\psi}_n)}{\pi(Z_{1i},\hat{\psi}_n)}\right\}\left\{\frac{\partial h_2^*(Z_{1i},\beta_n^*,\xi_n^*)}{\partial {\xi^*}^T}\right\}\right] n^{1/2}(\hat{\xi}_n^* - \xi^*), \tag{10.26}$$

where β_n^* and ξ_n^* are intermediate values between $\hat{\beta}_n$ and β_0 and $\hat{\xi}_n^*$ and ξ^*, respectively. Let us consider (10.26). Since $\hat{\psi}_n \xrightarrow{P} \psi_0$, $\beta_n^* \xrightarrow{P} \beta_0$, and $\xi_n^* \xrightarrow{P} \xi^*$, then, under suitable regularity conditions, the sample average in equation (10.26) is

$$n^{-1}\sum_{i=1}^{n}\left\{\frac{R_i - \pi(Z_{1i},\hat{\psi}_n)}{\pi(Z_{1i},\hat{\psi}_n)}\right\}\left\{\frac{\partial h_2^*(Z_{1i},\beta_n^*,\xi_n^*)}{\partial {\xi^*}^T}\right\} \xrightarrow{P}$$
$$E\left[\left\{\frac{R - \pi(Z_1,\psi_0)}{\pi(Z_1,\psi_0)}\right\}\left\{\frac{\partial h_2^*(Z_1,\beta_0,\xi^*)}{\partial {\xi*}^T}\right\}\right].$$

This last expectation can be shown to equal zero by a simple conditioning argument where we first find the conditional expectation given Z_1. Since,

under suitable regularity conditions, $n^{1/2}(\hat{\xi}_n^* - \xi^*)$ is bounded in probability, then a simple application of Slutsky's theorem can be used to show that (10.26) converges in probability to zero. A similar argument can be used also to show that (10.25) converges in probability to zero. □

If, in addition, the model for the conditional distribution of Z_2 given Z_1 was correctly specified, then

$$\left\{\frac{R-\pi(Z_1,\psi_0)}{\pi(Z_1,\psi_0)}\right\} h_2^*(Z_1,\beta_0,\xi^*) = \left\{\frac{R-\pi(Z_1,\psi_0)}{\pi(Z_1,\psi_0)}\right\} h_2^0(Z_1,\beta_0)$$
$$= \Pi\left[\frac{Rm(Z,\beta_0)}{\pi(Z_1,\psi_0)}\middle|\Lambda_2\right].$$

This would then imply that the term inside the square brackets "$[\,\cdot\,]$" in (10.21) is an element orthogonal to Λ_2 (i.e., $[\,\cdot\,] \in \Lambda_2^{\perp}$), in which case $\Pi\Big([\,\cdot\,]\Big|\Lambda_\psi\Big)$ (i.e., the last term of (10.21)) is equal to zero because $\Lambda_\psi \subset \Lambda_2$. The resulting estimator would have influence function

$$\frac{R\varphi^F(Z)}{\pi(Z_1,\psi_0)} - \Pi\left[\left\{\frac{R\varphi^F(Z)}{\pi(Z_1,\psi_0)}\right\}\middle|\Lambda_2\right] = \mathcal{J}(\varphi^F) \in (IF)_{\mathrm{DR}}, \tag{10.27}$$

where $\varphi^F(Z) = -\left[E\left\{\frac{\partial m(Z,\beta_0)}{\partial\beta^T}\right\}\right]^{-1} m(Z,\beta_0)$, and this influence function is within the class of the so-called double-robust influence functions. The variance of this influence function represents the smallest asymptotic variance among observed-data estimators that used $m(Z,\beta_0) = \varphi^{*F}(Z)$ as the basis for an augmented inverse probability weighted complete-case estimating equation.

Estimating the Asymptotic Variance

To estimate the asymptotic variance of $\hat{\beta}_n$, where $\hat{\beta}_n$ is the solution to (10.17), we propose using the sandwich variance estimator given in Chapter 9, equation (9.19). For completeness, this estimator is given by

$$\hat{\Sigma}_n = \left[\hat{E}\left\{\frac{\partial m(Z,\hat{\beta}_n)}{\partial\beta^T}\right\}\right]^{-1}$$
$$\times \left[n^{-1}\sum_{i=1}^{n} g(O_i,\hat{\psi}_n,\hat{\beta}_n,\hat{\xi}_n^*) g^T(O_i,\hat{\psi}_n,\hat{\beta}_n,\hat{\xi}_n^*)\right]$$
$$\times \left[\hat{E}\left\{\frac{\partial m(Z,\hat{\beta}_n)}{\partial\beta^T}\right\}\right]^{-1^T}, \tag{10.28}$$

where

$$\hat{E}\left\{\frac{\partial m(Z,\hat{\beta}_n)}{\partial \beta^T}\right\} = n^{-1}\sum_{i=1}^{n}\left\{\frac{R_i \partial m(Z_i,\hat{\beta}_n)/\partial \beta^T}{\pi(Z_{1i},\hat{\psi}_n)}\right\},$$

$$g(O_i,\hat{\psi}_n,\hat{\beta}_n,\hat{\xi}_n^*) =$$
$$q(O_i,\hat{\psi}_n,\hat{\beta}_n,\hat{\xi}_n^*) - \hat{E}(qS_\psi^T)\{\hat{E}(S_\psi S_\psi^T)\}^{-1}S_\psi(O_i,\hat{\psi}_n),$$

$$q(O_i,\hat{\psi}_n,\hat{\beta}_n,\hat{\xi}_n^*) = \frac{R_i m(Z_i,\hat{\beta}_n)}{\pi(Z_{1i},\hat{\psi}_n)} - \left\{\frac{R_i - \pi(Z_{1i},\hat{\psi}_n)}{\pi(Z_{1i},\hat{\psi}_n)}\right\} h_2^*(Z_{1i},\hat{\beta}_n,\hat{\xi}_n^*),$$

$$\hat{E}(qS_\psi^T) = n^{-1}\sum_{i=1}^{n} q(O_i,\hat{\psi}_n,\hat{\beta}_n,\hat{\xi}_n^*)S_\psi^T(O_i,\hat{\psi}_n),$$

and

$$\hat{E}(S_\psi S_\psi^T) = n^{-1}\sum_{i=1}^{n} S_\psi(O_i,\hat{\psi}_n)S_\psi^T(O_i,\hat{\psi}_n).$$

Double Robustness with Two Levels of Missingness

Thus far, we have taken the point of view that the missingness model was correctly specified; that is, that $\pi_0(Z_1) = P(R = 1|Z_1)$ is contained in the model $\pi(Z_1,\psi)$ for some value of ψ. If this is the case, we denote the true value of ψ by ψ_0 and $\pi_0(Z_1) = \pi(Z_1,\psi_0)$. However, unless the missingness was by design, we generally don't know the true missingness model, and therefore the possibility exists that we have misspecified this model. Even if the missingness model is misspecified, under suitable regularity conditions, the maximum likelihood estimator $\hat{\psi}_n$ will converge in probability to some constant ψ^*, but $\pi(Z_1,\psi^*) \neq \pi_0(Z_1)$. That is,

$$\pi(Z_1,\hat{\psi}_n) \to \pi(Z_1,\psi^*) \neq \pi_0(Z_1).$$

As we will demonstrate, the attempt to gain efficiency by positing a model for the conditional distribution of Z_2 given Z_1 and estimating

$$h_2^*(Z_1,\beta,\hat{\xi}_n^*) = E\{m(Z,\beta)|Z_1,\xi\}\Big|_{\xi=\hat{\xi}_n^*}$$

actually gives us the extra protection of double robustness, which was briefly introduced in Section 6.5. We now explore this issue further.

Using standard asymptotic approximations, we can show that the estimator $\hat{\beta}_n$, which is the solution to the estimating equation (10.17), will be consistent and asymptotically normal if

$$E\left[\frac{Rm(Z,\beta_0)}{\pi(Z_1,\psi^*)} - \left\{\frac{R-\pi(Z_1,\psi^*)}{\pi(Z_1,\psi^*)}\right\} h_2^*(Z_1,\beta_0,\xi^*)\right] = 0, \tag{10.29}$$

where $\hat{\psi}_n \xrightarrow{P} \psi^*$ and $\hat{\xi}_n^* \xrightarrow{P} \xi^*$.

In developing our estimator for β, we considered two models, one for the missingness probabilities $\pi(Z_1,\psi)$ and another for the conditional density of Z_2 given Z_1 $p^*_{Z_2|Z_1}(z_2|z_1,\xi)$. If the missingness model is correctly specified, then

$$\pi(Z_1,\psi^*) = \pi(Z_1,\psi_0) = P(R=1|Z_1). \tag{10.30}$$

If the model for the conditional density of Z_2 given Z_1 is correctly specified, then

$$h_2^*(Z_1,\beta_0,\xi^*) = E_0\{m(Z,\beta_0)|Z_1\}. \tag{10.31}$$

We now show the so-called double-robustness property; that is, the estimator $\hat{\beta}_n$, the solution to (10.17), is consistent and asymptotically normal (a result that follows under suitable regularity conditions when (10.29) is satisfied) if either (10.30) or (10.31) is true.

After adding and subtracting similar terms, we write the expectation in (10.29) as

$$\begin{aligned} &E\left[m(Z,\beta_0) + \left\{\frac{R-\pi(Z_1,\psi^*)}{\pi(Z_1,\psi^*)}\right\}\left\{m(Z,\beta_0) - h_2^*(Z_1,\beta_0,\xi^*)\right\}\right] \\ &= 0 + E\left[\left\{\frac{R-\pi(Z_1,\psi^*)}{\pi(Z_1,\psi^*)}\right\}\left\{m(Z,\beta_0) - h_2^*(Z_1,\beta_0,\xi^*)\right\}\right]. \end{aligned} \tag{10.32}$$

If (10.30) is true, whether (10.31) holds or not, we write (10.32) as

$$E\left[\left\{\frac{R-P(R=1|Z_1)}{P(R=1|Z_1)}\right\}\left\{m(Z,\beta_0) - h_2^*(Z_1,\beta_0,\xi^*)\right\}\right]. \tag{10.33}$$

We derive the expectation of (10.33) by using the law of conditional iterated expectations, where we first condition on $Z=(Z_1,Z_2)$ to obtain

$$E\left[\left\{\frac{E(R|Z_1,Z_2)-P(R=1|Z_1)}{P(R=1|Z_1)}\right\}\left\{m(Z,\beta_0) - h_2^*(Z_1,\beta_0,\xi^*)\right\}\right]. \tag{10.34}$$

Because of MAR, $R \perp\!\!\!\perp Z_2|Z_1$, which implies that $E(R|Z_1,Z_2) = E(R|Z_1) = P(R=1|Z_1)$. This then implies that (10.34) equals zero, which, in turn, implies that $\hat{\beta}_n$ is consistent when (10.30) is true.

If (10.31) is true, whether (10.30) holds or not, we write (10.32) as

$$E\left(\left\{\frac{R-\pi(Z_1,\psi^*)}{\pi(Z_1,\psi^*)}\right\}\left[m(Z,\beta_0) - E_0\{m(Z,\beta_0)|Z_1\}\right]\right). \tag{10.35}$$

Again, we evaluate (10.35) by using the law of conditional iterated expectations, where we first condition on (R,Z_1) to obtain

$$E\left(\left\{ \frac{R - \pi(Z_1, \psi^*)}{\pi(Z_1, \psi^*)} \right\} \left[E\{m(Z, \beta_0)|R, Z_1\} - E_0\{m(Z, \beta_0)|Z_1\} \right] \right). \quad (10.36)$$

Because of MAR, $R \perp\!\!\!\perp Z_2|Z_1$, which implies that $E\{m(Z, \beta_0)|R, Z_1\} = E_0\{m(Z, \beta_0)|Z_1\}$. This then implies that (10.36) equals zero, which, in turn, implies that $\hat{\beta}_n$ is consistent when (10.31) is true.

Remarks Regarding Double-Robust Estimators

In developing the adaptive strategy that led to double-robust estimators, we had to posit a simplifying model for $p^*_Z(z, \xi)$ or for $p^*_{Z_2|Z_1}(z_2|z_1, \xi)$ and then estimate the parameter ξ. We originally took the point of view that the model for the coarsening (missingness) probabilities $\pi(Z_1, \psi)$ was correctly specified and hence the posited model for the distribution of Z was used to compute projections onto the augmentation space, which, in turn, gained us efficiency while still leading to a consistent and asymptotically normal estimator for β even if the posited model was misspecified. To estimate the parameter ξ, we suggested using likelihood methods such as maximizing (10.15) or (10.16). However, any estimator that would lead to a consistent estimator of ξ (assuming the posited model was correct) could have been used. In fact, we might use an IPWCC or AIPWCC estimator for ξ since, in some cases, these may be easier to implement. For example, if a full-data estimating function for ξ was easily obtained (i.e., $m(Z, \xi)$ such that $E_\xi\{m(Z, \xi)\} = 0$), then a simple IPWCC estimator for ξ, using observed data, could be obtained by solving

$$\sum_{i=1}^{n} \frac{R_i m(Z_i, \xi)}{\pi(Z_{1i}, \hat{\psi}_n)} = 0. \quad (10.37)$$

Such a strategy is perfectly reasonable as long as we believe that the model for the coarsening probabilities is correctly specified because this is necessary to guarantee that the solution to (10.37) would lead to a consistent estimator of ξ if the posited model for the distribution of Z was correct.

However, if our goal is to obtain a double-robust estimator for β, then we must make sure that the estimator for ξ is a consistent estimator if the posited model for the distribution of Z was correctly specified, regardless of whether the model for the coarsening probabilities was correctly specified or not. This means that we could use likelihood methods such as maximizing (10.15) or (10.16), as such estimators do not involve the coarsening probabilities, but we could not use IPWCC or AIPWCC estimators such as (10.37).

Logistic Regression Example Revisited

In Chapter 7, we gave an example where we were interested in estimating the parameter β in a logistic regression model used to model the relationship of a binary outcome variable Y as a function of covariates $X = (X_1^T, X_2)^T$,

where, for ease of illustration, we let X_2 be a single random variable. Also, for this example, we let all the covariates in X be continuous random variables. Specifically, we consider the model

$$P(Y = 1|X) = \frac{\exp(\beta^T X^*)}{1 + \exp(\beta^T X^*)},$$

where $X^* = (1, X_1^T, X_2)^T$ is introduced to allow for an intercept term. We also denote the full data $Z = (Y, X)$. We considered the case where there were two levels of missingness, specifically where (Y, X_1) was always observed but where the single variable X_2 was missing on some of the individuals. The observed data are denoted by $O_i = (R_i, Y_i, X_{1i}, R_i X_{2i}), i = 1, \ldots, n$.

In the example in Chapter 7, we assumed that missingness was by design, but here we will assume that the missing data were not by design and hence the probability of missingness has to be modeled. We assume that the data are missing at random (MAR) and the missingness probability, or more precisely the probability of a complete case, follows the logistic regression model (10.8).

$$\pi(Y, X_1, \psi) = \frac{\exp(\psi_0 + \psi_1 Y + \psi_2^T X_1)}{1 + \exp(\psi_0 + \psi_1 Y + \psi_2^T X_1)}. \tag{10.38}$$

Estimates of the parameter $\psi = (\psi_0, \psi_1, \psi_2^T)^T$ are obtained by maximizing the likelihood (10.9). This can be easily accomplished using standard statistical software. The resulting estimator is denoted by $\hat{\psi}_n$ and the estimator for the probability of a complete case is denoted by $\pi(Y_i, X_{1i}, \hat{\psi}_n)$.

We next consider the choice for the full-data estimating function $m(Z, \beta)$. With full data, the optimal choice for $m(Z, \beta)$ is

$$m(Y, X_1, X_2, \beta) = X^* \left\{ Y - \frac{\exp(\beta^T X^*)}{1 + \exp(\beta^T X^*)} \right\}.$$

Although this may not be the optimal choice with the introduction of missing data, for the time being we will use this to construct observed-data AIPWCC estimators. In Chapter 7, AIPWCC estimators were introduced for this problem using equation (7.49) when the data were missing by design. When the missingness probability is modeled using (10.38), then we would consider AIPWCC estimators that are the solution to

$$\sum_{i=1}^{n} \left[\frac{R_i}{\pi(Y_i, X_{1i}, \hat{\psi}_n)} X_i^* \left\{ Y_i - \frac{\exp(\beta^T X_i^*)}{1 + \exp(\beta^T X_i^*)} \right\} \right.$$
$$\left. - \left\{ \frac{R_i - \pi(Y_i, X_{1i}, \hat{\psi}_n)}{\pi(Y_i, X_{1i}, \hat{\psi}_n)} \right\} L(Y_i, X_{1i}) \right] = 0, \tag{10.39}$$

where $L(Y, X_1)$ is some arbitrary q-dimensional function of Y and X_1. We now realize that in order to obtain as efficient an estimator as possible

for β among the estimators in (10.39), we should choose $L(Y, X_1)$ to equal $E\{m(Y, X_1, X_2, \beta)|Y, X_1\}$; that is,

$$L_2^*(Y, X_1) = E\left[X^*\left\{Y - \frac{\exp(\beta^T X^*)}{1 + \exp(\beta^T X^*)}\right\}\middle| Y, X_1\right].$$

Because the function $L(Y, X_1)$ depends on the conditional distribution of X_2 given Y and X_1, which is unknown to us, we posit a model for this conditional distribution and use adaptive methods.

The model we consider is motivated by the realization that a logistic regression model for Y given X would be obtained if the conditional distribution of X given Y followed a multivariate normal distribution with a mean that depends on Y but with a variance matrix that is independent of Y; that is, the conditional distribution of X given $Y = 1$ would be $MVN(\mu_1, \Sigma)$ and would be $MVN(\mu_0, \Sigma)$ given $Y = 0$; see, for example, Cox and Snell (1989). For this scenario, the conditional distribution of X_2 given X_1 and Y would follow a normal distribution; i.e.,

$$X_2|X_1, Y \sim N\Big(\xi_0 + \xi_1^T X_1 + \xi_2 Y, \xi_{\sigma^2}\Big). \tag{10.40}$$

Therefore, one strategy is to posit the model (10.40) for the conditional distribution of X_2 given X_1 and Y and estimate the parameters $\xi = (\xi_0, \xi_1^T, \xi_2, \xi_{\sigma^2})^T$ by maximizing (10.16). This is especially attractive because the model (10.40) is a traditional normally distributed linear model. Therefore, using the complete cases (i.e., $\{i : R_i = 1\}$), the MLE for $(\xi_0, \xi_1^T, \xi_2)^T$ can be obtained using standard least squares and the MLE for the variance parameter ξ_{σ^2} can be obtained as the average of the squared residuals. Denote this estimator by $\hat{\xi}_n^*$.

Finally, me must compute

$$\begin{aligned} &L_2^*(Y, X_1, \beta, \hat{\xi}_n^*) = E\{m(Y, X_1, X_2, \beta)|Y, X_1, \xi\}|_{\xi=\hat{\xi}_n^*} \\ &= \int m(Y, X_1, u, \beta)(2\pi\hat{\xi}_{\sigma_n^2}^*)^{-1/2} \exp - \left[\frac{\{u - (\hat{\xi}_{0n}^* + \hat{\xi}_{1n}^{*T} X_1 + \hat{\xi}_{2n}^* Y)\}^2}{2\hat{\xi}_{\sigma_n^2}^*}\right] du. \end{aligned}$$

This can be accomplished using numerical integration or Monte Carlo techniques.

The estimator for β is then obtained by solving the equation

$$\begin{aligned} &\sum_{i=1}^n \left[\frac{R_i}{\pi(Y_i, X_{1i}, \hat{\psi}_n)} X_i^* \left\{Y_i - \frac{\exp(\beta^T X_i^*)}{1 + \exp(\beta^T X_i^*)}\right\}\right. \\ &\left. - \left\{\frac{R_i - \pi(Y_i, X_{1i}, \hat{\psi}_n)}{\pi(Y_i, X_{1i}, \hat{\psi}_n)}\right\} L_2^*(Y_i, X_{1i}, \beta, \hat{\xi}_n^*)\right] = 0. \end{aligned} \tag{10.41}$$

This estimator is doubly robust; it is a consistent asymptotically normal RAL estimator for β if either the missingness model (10.38) or the model for the conditional distribution of X_2 given X_1 and Y (10.40) is correct.

The asymptotic variance for $\hat{\beta}_n$ can be obtained by using the sandwich variance estimator (10.28).

10.3 Improving Efficiency with Monotone Coarsening

Finding the Projection onto the Augmentation Space

Monotone missingness, or more generally monotone coarsening, occurs often in practice and is worth special consideration. We have shown that a natural way to obtain semiparametric coarsened-data estimators for β is to consider augmented inverse probability weighted complete-case estimators, estimators based on

$$\sum_{i=1}^{n}\left[\frac{I(\mathcal{C}_i=\infty)m(Z_i,\beta)}{\varpi(\infty,Z_i,\hat{\psi}_n)}+L_2\{\mathcal{C}_i,G_{\mathcal{C}_i}(Z_i)\}\right]=0, \tag{10.42}$$

where $m(Z_i,\beta)$ is an estimating function such that $m(Z,\beta_0)=\varphi^{*F}(Z)\in\Lambda^{F\perp}$ and $L_2\{\mathcal{C},G_{\mathcal{C}}(Z)\}\in\Lambda_2$. We also proved in Section 10.1 that we should choose $L_2\{\mathcal{C},G_{\mathcal{C}}(Z)\}=-\Pi\left[\frac{I(\mathcal{C}=\infty)m(Z,\beta)}{\varpi(\infty,Z,\psi)}\middle|\Lambda_2\right]$ in order to gain the greatest efficiency among estimators that solve (10.42); see (10.7). Therefore, to obtain estimators for β with improved efficiency, we now consider how to find the projection of $\frac{I(\mathcal{C}=\infty)m(Z,\beta)}{\varpi(\infty,Z,\psi)}$ onto the augmentation space Λ_2 when coarsening is monotone.

We remind the reader that a typical element of Λ_2, written in terms of the coarsening probabilities, is given by (7.37). However, under monotone coarsening, we showed (in Theorem 9.2) that a typical element of Λ_2 can be reparametrized, in terms of discrete hazard functions, as

$$\sum_{r\neq\infty}\left[\frac{I(\mathcal{C}=r)-\lambda_r\{G_r(Z)\}I(\mathcal{C}\geq r)}{K_r\{G_r(Z)\}}\right]L_r\{G_r(Z)\},$$

where the discrete hazard function $\lambda_r\{G_r(Z)\}$ and $K_r\{G_r(Z)\}$ are defined by (8.2) and (8.4), respectively, and $L_r\{G_r(Z)\}$ denotes an arbitrary function of $G_r(Z)$ for $r\neq\infty$.

By the projection theorem, if we want to find

$$\Pi\left[\frac{I(\mathcal{C}=\infty)m(Z,\beta)}{\varpi(\infty,Z)}\middle|\Lambda_2\right],$$

then we must derive the functions $L_{0r}\{G_r(Z)\},\ r\neq\infty$, such that

$$E\Bigg(\frac{I(\mathcal{C}=\infty)m^T(Z,\beta)}{\varpi(\infty,Z)} - \sum_{r\neq\infty}\left[\frac{I(\mathcal{C}=r)-\lambda_r\{G_r(Z)\}I(\mathcal{C}\geq r)}{K_r\{G_r(Z)\}}\right]L_{0r}^T\{G_r(Z)\}\Bigg) \times\left(\sum_{r\neq\infty}\left[\frac{I(\mathcal{C}=r)-\lambda_r\{G_r(Z)\}I(\mathcal{C}\geq r)}{K_r\{G_r(Z)\}}\right]L_r\{G_r(Z)\}\right) = 0 \quad \text{for all } L_r\{G_r(Z)\}, r\neq\infty. \tag{10.43}$$

Theorem 10.4. The projection of $\frac{I(\mathcal{C}=\infty)m(Z,\beta)}{\varpi(\infty,Z)}$ onto Λ_2 (i.e., the solution to (10.43)) is obtained by taking $L_{0r}\{G_r(Z)\} = -E\{m(Z,\beta)|G_r(Z)\}$.

Those readers who are familiar with the counting process notation and stochastic integral martingale processes that are used in an advanced course in survival analysis (see, for example, Fleming and Harrington, 1991) will find that the proof of Theorem 10.4 uses similar methods. We first derive some relationships in the following three lemmas that will simplify the calculations in (10.43).

Lemma 10.1. For $r \neq r'$

$$E\left(\left[\frac{I(\mathcal{C}=r)-\lambda_r\{G_r(Z)\}I(\mathcal{C}\geq r)}{K_r\{G_r(Z)\}}\right]L_{0r}^T\{G_r(Z)\}\right) \times\left(\left[\frac{I(\mathcal{C}=r')-\lambda_{r'}\{G_{r'}(Z)\}I(\mathcal{C}\geq r')}{K_{r'}\{G_{r'}(Z)\}}\right]L_{r'}\{G_{r'}(Z)\}\right)=0. \tag{10.44}$$

Proof. For a single observation, define $\mathcal{F}_r$ to be the random vector $\{I(\mathcal{C}=1), I(\mathcal{C}=2),\ldots, I(\mathcal{C}=r-1), Z\}$. Without loss of generality, take $r'>r$. The expectation in (10.44) can be evaluated as $E\{E(\cdot|\mathcal{F}_{r'})\}$. Conditional on $\mathcal{F}_{r'}$, however, the only term in (10.44) that is not known is $I(\mathcal{C}=r')$. Consequently, (10.44) can be written as

$$E\left(\left[\frac{I(\mathcal{C}=r)-\lambda_r\{G_r(Z)\}I(\mathcal{C}\geq r)}{K_r\{G_r(Z)\}}\right]L_{0r}^T\{G_r(Z)\}\right) \times\left(\left[\frac{E\{I(\mathcal{C}=r')|\mathcal{F}_{r'}\}-\lambda_{r'}\{G_{r'}(Z)\}I(\mathcal{C}\geq r')}{K_{r'}\{G_{r'}(Z)\}}\right]L_{r'}\{G_{r'}(Z)\}\right). \tag{10.45}$$

But

$$E\{I(\mathcal{C}=r')|\mathcal{F}_{r'}\} = P(\mathcal{C}=r'|\mathcal{C}\geq r', Z)I(\mathcal{C}\geq r'), \tag{10.46}$$

which by the coarsening at random assumption and the definition of a discrete hazard, given by (8.2), is equal to $\lambda_{r'}\{G_{r'}(Z)\}I(\mathcal{C}\geq r')$. Substituting (10.46) into (10.45) proves (10.44). □

Lemma 10.2. When $r = r'$, the left-hand side of (10.44) equals

$$E\left(\left[\frac{I(\mathcal{C}=r)-\lambda_r\{G_r(Z)\}I(\mathcal{C}\geq r)}{K_r\{G_r(Z)\}}\right]L_{0r}^T\{G_r(Z)\}\right)$$
$$\times\left(\left[\frac{I(\mathcal{C}=r)-\lambda_r\{G_r(Z)\}I(\mathcal{C}\geq r)}{K_r\{G_r(Z)\}}\right]L_r\{G_r(Z)\}\right)$$
$$=E\left[\frac{\lambda_r\{G_r(Z)\}}{K_r\{G_r(Z)\}}L_{0r}^T\{G_r(Z)\}L_r\{G_r(Z)\}\right]. \tag{10.47}$$

Proof. Computing the expectation of the left-hand side of equation (10.47) by first conditioning on $\mathcal{F}_r$ yields

$$E\left(\frac{E[\{I(\mathcal{C}=r)-\lambda_r\{G_r(Z)\}I(\mathcal{C}\geq r)\}^2|\mathcal{F}_r]}{K_r^2\{G_r(Z)\}}L_{0r}^T\{G_r(Z)\}L_r\{G_r(Z)\}\right). \tag{10.48}$$

The conditional expectation

$$E[\{I(\mathcal{C}=r)-\lambda_r\{G_r(Z)\}I(\mathcal{C}\geq r)\}^2|\mathcal{F}_r]$$

is the conditional variance of the Bernoulli indicator $I(\mathcal{C}=r)$ given $\mathcal{F}_r$, which equals

$$\lambda_r\{G_r(Z)\}[1-\lambda_r\{G_r(Z)\}]I(\mathcal{C}\geq r).$$

Hence, (10.48) equals

$$E\left(\frac{\lambda_r\{G_r(Z)\}[1-\lambda_r\{G_r(Z)\}]I(\mathcal{C}\geq r)}{K_r^2\{G_r(Z)\}}L_{0r}^T\{G_r(Z)\}L_r\{G_r(Z)\}\right). \tag{10.49}$$

Computing the expectation of (10.49) by first conditioning on Z gives

$$E\left\{\frac{\lambda_r(1-\lambda_r)P(\mathcal{C}\geq r|Z)}{K_r^2}L_{0r}^TL_r\right\}. \tag{10.50}$$

Since $P(\mathcal{C}\geq r|Z)=\prod_{j=1}^{r-1}(1-\lambda_j)$, we obtain that $(1-\lambda_r)P(\mathcal{C}\geq r|Z)=\prod_{j=1}^{r}(1-\lambda_j)=K_r$. Therefore, (10.50) equals $E\left[\frac{\lambda_r}{K_r}L_{0r}^TL_r\right]$, thus proving the lemma. □

Lemma 10.3.

$$E\left(\frac{I(\mathcal{C}=\infty)m^T(Z,\beta)}{\varpi(\infty,Z)}\left[\frac{I(\mathcal{C}=r)-\lambda_r\{G_r(Z)\}I(\mathcal{C}\geq r)}{K_r\{G_r(Z)\}}\right]L_r\{G_r(Z)\}\right)$$
$$=-E\left[\frac{\lambda_r\{G_r(Z)\}}{K_r\{G_r(Z)\}}m^T(Z,\beta)L_r\{G_r(Z)\}\right]. \tag{10.51}$$

Proof. Since $I(\mathcal{C}=\infty)I(\mathcal{C}=r)=0$ for $r\neq\infty$ and $I(R=\infty)I(\mathcal{C}\geq r)=I(\mathcal{C}=\infty)$, the left-hand side of (10.51) equals

$$-E\left[\frac{I(\mathcal{C}=\infty)}{\varpi(\infty,Z)}\frac{\lambda_r\{G_r(Z)\}}{K_r\{G_r(Z)\}}m^T(Z,\beta)L_r\{G_r(Z)\}\right].$$

The result (10.51) holds by first conditioning on Z and realizing that

$$E\left\{\frac{I(\mathcal{C}=\infty)}{\varpi(\infty,Z)}\middle|Z\right\}=1.\ \square$$

Proof. Theorem 10.4
Using the results of Lemmas 10.1–10.3, equation (10.43) can be written as

$$-\sum_{r\neq\infty}E\left(\frac{\lambda_r\{G_r(Z)\}}{K_r\{G_r(Z)\}}\left[m(Z,\beta)+L_{0r}\{G_r(Z)\}\right]^T L_r\{G_r(Z)\}\right)=0, \quad (10.52)$$

for all functions $L_r\{G_r(Z)\}, r\neq\infty$.

By conditioning each expectation in the sum on the left-hand side of (10.52) by $G_r(Z)$, this can be written as

$$-\sum_{r\neq\infty}E\Bigg(\frac{\lambda_r\{G_r(Z)\}}{K_r\{G_r(Z)\}}\Big[E\{m(Z,\beta)|G_r(Z)\} + L_{0r}\{G_r(Z)\}\Big]^T L_r\{G_r(Z)\}\Bigg)=0. \quad (10.53)$$

We now show that equation (10.53) holds if and only if

$$L_{0r}\{G_r(Z)\}=-E\left\{m(Z,\beta)|G_r(Z)\right\} \text{ for all } r\neq\infty. \quad (10.54)$$

Clearly, (10.53) follows when (10.54) holds. Conversely, if (10.54) were not true, then we could choose $L_r\{G_r(Z)\}=E\{m(Z,\beta)|\ G_r(Z)\}+L_{0r}\{G_r(Z)\}$ for all $r\neq\infty$ to get a contradiction.

Therefore, we have demonstrated that, with monotone CAR,

$$\Pi\left[\frac{I(\mathcal{C}=\infty)m(Z,\beta)}{\varpi(\infty,Z)}\middle|\Lambda_2\right] = -\sum_{r\neq\infty}\left[\frac{I(\mathcal{C}=r)-\lambda_r\{G_r(Z)\}I(\mathcal{C}\geq r)}{K_r\{G_r(Z)\}}\right]E\left\{m(Z,\beta)|G_r(Z)\right\}. \quad (10.55)$$

$\square$

In order to take advantage of the results above, we need to compute $E\{m(Z,\beta)|G_r(Z)\}$. This requires us to estimate the distribution of Z, or at least enough of the distribution to be able to compute these conditional expectations.

Remark 4. This last statement almost seems like circular reasoning. That is, we argue that to gain greater efficiency, we would need to estimate the distribution of Z. However, if we had methods to estimate the distribution of Z, then we wouldn't need to develop this theory in the first place. The rationale for considering semiparametric theory for coarsened data is that it led us to augmented inverse probability weighted complete-case estimators, which, we argued, build naturally on full-data estimators and are easier to derive than, say, likelihood methods with coarsened data. However, as we saw in the case with two levels of missingness, we will still obtain consistent asymptotically normal estimators using this inverse weighted methodology even if we construct estimators for the distribution of Z that are incorrect. This gives us greater flexibility and robustness and suggests the use of an adaptive approach, as we now describe. □

Adaptive Estimation

To take advantage of the results for increased efficiency, we consider an adaptive approach. That is, we posit simpler models for the distribution of Z only for the purpose of approximating the conditional expectations $E\{m(Z,\beta)|G_r(Z)\}, r \neq \infty$. We do not necessarily expect that these posited models are correct, although we do hope that they may serve as a reasonable approximation. However, even if the posited model is incorrect, the resulting expectation, which we denote as $E_{\text{INC}}\{m(Z,\beta)|G_r(Z)\}$, because

$$E_{\text{INC}}\{m(Z,\beta)|G_r(Z)\} \neq E_0\{m(Z,\beta)|G_r(Z)\},$$

still results in a function of $G_r(Z)$.

Consequently, computing (10.55) under incorrectly posited models would lead us to use

$$L_2\{\mathcal{C}, G_{\mathcal{C}}(Z)\} = \sum_{r\neq\infty}\left[\frac{I(\mathcal{C}=r)-\lambda_r\{G_r(Z)\}I(\mathcal{C}\geq r)}{K_r\{G_r(Z)\}}\right]E_{\text{INC}}\{m(Z,\beta)|G_r(Z)\} \quad (10.56)$$

in the estimating equation (10.42). Even though (10.56) is not the correct projection of $\dfrac{I(R=\infty)m(Z,\beta)}{\varpi(\infty,Z)}$ onto Λ_2, if the posited model for Z is incorrect, it is still an element of the augmentation space Λ_2, which implies that the solution to (10.42), using the augmented term $L_2(\cdot)$ as given by (10.56), would still result in an AIPWCC consistent, asymptotically normal semiparametric estimator for β. This protection against misspecified models argues in favor of using an adaptive approach.

In an adaptive strategy, to improve efficiency, we start by positing a simpler and possibly incorrect model for the distribution of the full data Z. Say we

assume $Z \sim p_Z^*(z,\xi)$, where ξ is finite-dimensional. Under this presumed model, we could compute the conditional expectations

$$E\{m(Z,\beta)|G_r(Z),\xi\} = \frac{\int\limits_{\{G_r(z)=G_r(Z)\}} m(z,\beta)p_Z^*(z,\xi)d\nu_Z(z)}{\int\limits_{\{G_r(z)=G_r(Z)\}} p_Z^*(z,\xi)d\nu_Z(z)}. \tag{10.57}$$

Of course, we need to estimate the parameter ξ in our posited model. Because the parameter ξ is finite-dimensional, we can estimate ξ using likelihood techniques as described in Section 7.1.

Using (7.10), the likelihood for a realization of such data, $(r_i, g_{r_i}), i = 1,\ldots,n$, when the coarsening mechanism is CAR, is equal to

$$\prod_{i=1}^{n} p^*_{G_{r_i}(Z_i)}(g_{r_i},\xi), \tag{10.58}$$

where

$$p^*_{G_r(Z)}(g_r,\xi) = \int\limits_{\{z:G_r(z)=g_r\}} p_Z^*(z,\xi)d\nu_Z(z).$$

Consequently, we would estimate $E\{m(Z,\beta)|G_r(Z),\xi\}$ by substituting $\hat{\xi}_n^*$ for ξ in (10.57), where $\hat{\xi}_n^*$ is the MLE obtained by maximizing (10.58). The estimated conditional expectation is denoted by $E\{m(Z,\beta)|G_r(Z),\hat{\xi}_n^*\}$.

Remark 5. Because of the monotone nature of the coarsening, it may be convenient to build models for the density of Z by considering models for the conditional density of $G_{r'}(Z)$ given $G_{r'-1}(Z)$. Specifically, the density of Z can be written as

$$P_Z^*(z) = \prod_{r'=1}^{\infty} p^*_{G_{r'}(Z)|G_{r'-1}(Z)}(g_{r'}|g_{r'-1}),$$

where $g_r = G_r(z)$, $p^*_{G_1(Z)|G_0(Z)}(g_1|g_0) = p^*_{G_1(Z)}(g_1)$,

$$p^*_{G_{r'}(Z)|G_{r'-1}(Z)}(g_{r'}|g_{r'-1}) = p^*_{Z|G_\ell(Z)}(z|g_\ell) \text{ when } r' = \infty,$$

and ℓ denotes the number of levels of coarsening. With this representation, we can write the density

$$p^*_{G_r(Z)}(g_r) = \prod_{r'\le r} p^*_{G_{r'}(Z)|G_{r'-1}(Z)}(g_{r'}|g_{r'-1}).$$

If, in addition, we consider models for the conditional density

$$p^*_{G_r(Z)|G_{r-1}(Z)}(g_r|g_{r-1},\xi_r),$$

in terms of parameter ξ_r, where ξ_r, for different r, are variationally independent, then the likelihood (10.58) can be written as

$$\prod_{i:r_i\geq 1} p^*_{G_1(Z_i)}(g_{1_i},\xi_1) \prod_{i:r_i\geq 2} p^*_{G_2(Z_i)|G_1(Z_i)}(g_{2_i}|g_{1_i},\xi_2)$$
$$\times \ldots \times \prod_{i:r_i=\infty} p^*_{Z_i|G_\ell(Z_i)}(z_i|g_{\ell_i},\xi_\infty). \tag{10.59}$$

The maximum likelihood estimator for $\xi = (\xi_1,\ldots,\xi_\infty)^T$ can then be obtained by maximizing each of the terms in (10.59) separately. □

Thus, the adaptive approach to finding estimators when data are monotonically coarsened is as follows.

1. We first consider the full-data problem. That is, how would semiparametric estimators for β be obtained if we had full data? For example, we may use a full-data m-estimator for β, which is the solution to

$$\sum_{i=1}^n m(Z_i,\beta) = 0,$$

which has influence function

$$-\left[E\left\{\frac{\partial m(Z_i,\beta_0)}{\partial \beta^T}\right\}\right]^{-1} m(Z_i,\beta_0) = \varphi^F(Z_i).$$

2. Next, we consider the augmented inverse probability weighted complete-case estimator that is the solution to (10.42), with $L_2(\cdot)$ being an estimator of (10.56). Specifically, we consider the estimator for β that solves the estimating equation

$$\sum_{i=1}^n \left(\frac{I(\mathcal{C}_i=\infty)m(Z_i,\beta)}{\varpi(\infty,Z_i,\hat{\psi}_n)} + \right.$$
$$\left. \sum_{r\neq\infty} \left[\frac{I(\mathcal{C}_i=r)-\lambda_r\{G_r(Z_i),\hat{\psi}_n\}I(\mathcal{C}_i\geq r)}{K_r\{G_r(Z_i),\hat{\psi}_n\}}\right] E\{m(Z,\beta)|G_r(Z_i),\hat{\xi}^*_n\}\right)$$
$$= 0, \tag{10.60}$$

where $\varpi(\infty,Z_i,\hat{\psi}_n) = \prod_{r<\infty}\left[1-\lambda_r\{G_r(Z_i),\hat{\psi}_n\}\right]$, (see (8.6)) $\hat{\psi}_n$ is the maximum likelihood estimator for ψ obtained by maximizing (8.12), and $E\{m(Z,\beta)|G_r(Z_i),\hat{\xi}^*_n\}$ is obtained using (10.57), substituting $\hat{\xi}^*_n$, which maximizes (10.58) or (10.59), for ξ. We denote this estimator by $\hat{\beta}_n$.

Even though the posited model $p^*_Z(z,\xi)$ may not be correctly specified, under suitable regularity conditions, the estimator $\hat{\xi}^*_n$ will converge in probability to a constant ξ^* and that $n^{1/2}(\hat{\xi}^*_n - \xi^*)$ will be bounded in probability. Also, even if the posited model is incorrect, the function $L_2\{\mathcal{C}_i, G_{\mathcal{C}_i}(Z_i), \beta, \psi_0, \xi^*\} \in \Lambda_2$, where

$$L_2\{\mathcal{C}_i, G_{\mathcal{C}_i}(Z_i), \beta, \psi, \xi\} = \sum_{r \neq \infty} \left[\frac{I(\mathcal{C}_i = r) - \lambda_r\{G_r(Z_i), \psi\} I(\mathcal{C}_i \geq r)}{K_r\{G_r(Z_i), \psi\}} \right] E\{m(Z,\beta) | G_r(Z_i), \xi\}.$$

Consequently, the solution to the estimating equation

$$\sum_{i=1}^{n} \left[\frac{I(\mathcal{C}_i = \infty) m(Z_i, \beta)}{\varpi(\infty, Z_i, \hat{\psi}_n)} + L_2\{\mathcal{C}_i, G_{\mathcal{C}_i}(Z_i), \beta_0, \hat{\psi}_n, \xi^*\} \right] = 0, \tag{10.61}$$

with β set to β_0 and ξ^* fixed in $L_2(\cdot)$, is an AIPWCC estimator for β, which we denote by $\hat{\beta}^*_n$.

Using an argument similar to that in Theorem 10.3, when we considered two levels of missingness, we can show that, under suitable regularity conditions, the estimator $\hat{\beta}_n$, which solves equation (10.60), is asymptotically equivalent to the AIPWCC estimator $\hat{\beta}^*_n$, which solves (10.61); that is,

$$n^{1/2}(\hat{\beta}_n - \hat{\beta}^*_n) \xrightarrow{P} 0.$$

(We leave this as an exercise for the reader.)

The resulting influence function for $\hat{\beta}_n$, which is the same as for $\hat{\beta}^*_n$, can now be derived by Theorem 9.1 and is equal to

$$\left(\frac{I(\mathcal{C}_i = \infty)\varphi^F(Z_i)}{\varpi(\infty, Z_i, \psi_0)} + L_2^*\{\mathcal{C}_i, G_{\mathcal{C}_i}(Z_i), \beta_0, \psi_0, \xi^*\} - \Pi\left[\frac{I(\mathcal{C}_i = \infty)\varphi^F(Z_i)}{\varpi(\infty, Z_i, \psi_0)} + L_2^*\{\mathcal{C}_i, G_{\mathcal{C}_i}(Z_i), \beta_0, \psi_0, \xi^*\} \middle| \Lambda_\psi \right] \right),$$

where

$$\varphi^F(Z_i) = \left[E\left\{ \frac{\partial m(Z_i, \beta_0)}{\partial \beta^T} \right\} \right]^{-1} m(Z_i, \beta_0)$$

and

$$L_2^*\{\mathcal{C}_i, G_{\mathcal{C}_i}, \beta_0, \psi_0, \xi^*\} = - \left[E\left\{ \frac{\partial m(Z_i, \beta_0)}{\partial \beta^T} \right\} \right]^{-1} L_2\{\mathcal{C}_i, G_{\mathcal{C}_i}(Z_i), \beta_0, \psi_0, \xi^*\}.$$

The asymptotic variance of $\hat{\beta}_n$ can also be obtained by using the sandwich variance estimator for AIPWCC estimators given by (9.19).

Remark 6. If the posited model $p^*(z, \xi)$ is correctly specified, then

$$E\{m(Z_i, \beta_0)|G_r(Z_i), \hat{\xi}_n^*\}$$

will be a consistent estimator of

$$E_0\{m(Z_i, \beta_0)|G_r(Z_i)\},$$

the true conditional expectation. In this case,

$$L_2\{\mathcal{C}_i, G_{\mathcal{C}_i}(Z_i), \beta_0, \hat{\psi}_n, \hat{\xi}_n^*\}$$

will converge to

$$\Pi\left[\frac{I(\mathcal{C}_i = \infty)m(Z_i, \beta_0)}{\varpi(\infty, Z_i, \psi_0)}\middle|\Lambda_2\right].$$

For the case of a correctly specified model, the influence function is

$$\left\{\frac{I(\mathcal{C}_i = \infty)\varphi^F(Z_i)}{\varpi(\infty, Z_i, \psi_0)} - \Pi\left[\frac{I(\mathcal{C}_i = \infty)\varphi^F(Z_i)}{\varpi(\infty, Z_i, \psi_0)}\middle|\Lambda_2\right]\right\}$$
$$- \Pi\left[\left\{\ \cdot\ \right\}\middle|\Lambda_\psi\right].$$
$$\Uparrow$$
This is orthogonal to Λ_2 and since $\Lambda_\psi \subset \Lambda_2$ must equal 0

In this case, the influence function equals

$$\frac{I(\mathcal{C} = \infty)\varphi^F(Z)}{\varpi(\infty, Z, \psi_0)} - \Pi\left[\frac{I(\mathcal{C} = \infty)\varphi^F(Z)}{\varpi(\infty, Z, \psi_0)}\middle|\Lambda_2\right] = \mathcal{J}\{\varphi^F(Z)\},$$

which is the most efficient among observed-data influence functions associated with the full-data influence function $\varphi^F(Z_i)$ and an element of $(IF)_{\text{DR}}$. □

Generally, the attempt to estimate

$$\Pi\left[\frac{I(\mathcal{C}_i = \infty)\varphi^F(Z_i)}{\varpi(\infty, Z_i)}\middle|\Lambda_2\right]$$

by positing a model $p_Z^*(z, \xi)$ often leads to more efficient estimators even if the model was incorrect. In fact, this attempt to gain efficiency also gives us the extra protection of double robustness similar to that seen in the previous section when we considered two levels of missingness. We now explore this double-robustness relationship for the case of monotone coarsening.

Double Robustness with Monotone Coarsening

Throughout this section, we have taken the point of view that the model for the coarsening probabilities was correctly specified. That is, for some ψ_0,

$$\lambda_r\{G_r(Z), \psi_0\} = P_0(\mathcal{C} = r|\mathcal{C} \geq r, Z), \ r \neq \infty,$$

where $P_0(\mathcal{C} = r|\mathcal{C} \geq r, Z)$ denotes the true discrete hazard rate. In actuality, this model may also be misspecified. Nonetheless, under suitable regularity conditions, the maximum likelihood estimator $\hat{\psi}_n$, which maximizes (8.12), will converge to a constant ψ^* even if the model for the coarsening hazards is not correctly specified.

When we developed the adaptive estimators for the purpose of improving efficiency, we considered the posited model $p^*_Z(z, \xi)$ and argued that the estimator $\hat{\xi}^*_n$ converged to a constant ξ^*, where $p^*_Z(z, \xi^*)$ may not be the correct distribution for the full data Z (i.e., $p^*_Z(z, \xi^*) \neq p_{0Z}(z)$), where $p_{0Z}(z) = p_Z(z, \beta_0, \eta_0)$ denotes the true density of Z. We now consider the double-robustness property of the proposed adaptive estimator $\hat{\beta}_n$ for β, the solution to (10.60). That is, we will prove that $\hat{\beta}_n$ is a consistent estimator if either the model for $\lambda_r\{G_r(Z), \psi\}, \ r \neq \infty$ or the posited model $p^*_Z(z, \xi)$ is correctly specified.

Using standard asymptotic arguments, the estimator $\hat{\beta}_n$ will be consistent and asymptotically normal if we can show that

$$E\left(\frac{I(\mathcal{C} = \infty)m(Z, \beta_0)}{\varpi(\infty, Z, \psi^*)} + \right.$$
$$\left.\sum_{r \neq \infty}\left[\frac{I(\mathcal{C} = r) - \lambda_r\{G_r(Z), \psi^*\}I(\mathcal{C} \geq r)}{K_r\{G_r(Z), \psi^*\}}\right] E\{m(Z, \beta_0)|G_r(Z), \xi^*\}\right) = 0. \tag{10.62}$$

It will be convenient to show first that the expression inside the expectation on the left-hand side of (10.62) can be written as

$$m(Z, \beta_0) - \sum_{r \neq \infty}\left(\left[\frac{I(\mathcal{C} = r) - \lambda_r\{G_r(Z), \psi^*\}I(\mathcal{C} \geq r)}{K_r\{G_r(Z), \psi^*\}}\right]\right.$$
$$\left.\times\left[m(Z, \beta_0) - E\{m(Z, \beta_0)|G_r(Z), \xi^*\}\right]\right). \tag{10.63}$$

This follows because

$$\frac{I(\mathcal{C} = \infty)m(Z, \beta_0)}{\varpi(\infty, Z, \psi^*)} = m(Z, \beta_0) + \left\{\frac{I(\mathcal{C} = \infty)}{\varpi(\infty, Z, \psi^*)} - 1\right\} m(Z, \beta_0) \tag{10.64}$$

and by the following lemma.

Lemma 10.4.

$$\sum_{r\neq\infty}\left[\frac{I(\mathcal{C}=r)-\lambda_r\{G_r(Z),\psi^*\}I(\mathcal{C}\geq r)}{K_r\{G_r(Z),\psi^*\}}\right]=\left\{1-\frac{I(\mathcal{C}=\infty)}{\varpi(\infty,Z,\psi^*)}\right\}. \quad (10.65)$$

Proof. Lemma 10.4
Because of the discreteness of $\mathcal{C}$, we can write

$$\sum_{r\neq\infty}\frac{I(\mathcal{C}=r)}{K_r\{G_r(Z),\psi^*\}}=\frac{I(\mathcal{C}\neq\infty)}{K_{\mathcal{C}}\{G_{\mathcal{C}}(Z),\psi^*\}}. \quad (10.66)$$

By the definitions of $\lambda_r(\cdot)$ and $K_r(\cdot)$ given by (8.2) and (8.4), respectively, we obtain that

$$\frac{\lambda_r(\cdot)}{K_r(\cdot)}=\frac{1}{K_r(\cdot)}-\frac{1}{K_{r-1}(\cdot)},$$

where $K_0(\cdot)=1$. Consequently,

$$\begin{aligned}-\sum_{r\neq\infty}\frac{\lambda_r(\cdot)I(\mathcal{C}\geq r)}{K_r(\cdot)}=&I(\mathcal{C}\neq\infty)\sum_{r\leq\mathcal{C}}\left\{\frac{1}{K_{r-1}(\cdot)}-\frac{1}{K_r(\cdot)}\right\}\\&+I(\mathcal{C}=\infty)\sum_{r\neq\infty}\left\{\frac{1}{K_{r-1}(\cdot)}-\frac{1}{K_r(\cdot)}\right\}\\=&I(\mathcal{C}\neq\infty)\left[1-\frac{1}{K_{\mathcal{C}}\{G_{\mathcal{C}}(Z),\psi^*\}}\right]\\&+I(\mathcal{C}=\infty)\left[1-\frac{1}{K_\ell\{G_\ell(Z),\psi^*\}}\right], \quad (10.67)\end{aligned}$$

where ℓ denotes the number of different coarsening levels (i.e., the largest integer $r<\infty$) and

$$K_\ell\{G_\ell(Z),\psi^*\}=\prod_{r\neq\infty}[1-\lambda_r\{G_r(Z),\psi^*\}]=\varpi(\infty,Z,\psi^*). \quad (10.68)$$

Taking the sum of (10.66) and (10.67) and substituting $\varpi(\infty,Z,\psi^*)$ for $K_\ell\{G_\ell(Z),\psi^*\}$ (see (10.68)), we obtain

$$I(\mathcal{C}\neq\infty)+I(\mathcal{C}=\infty)\left\{1-\frac{1}{\varpi(\infty,Z,\psi^*)}\right\}=\left\{1-\frac{I(\mathcal{C}=\infty)}{\varpi(\infty,Z,\psi^*)}\right\},$$

thus proving the lemma. □

Therefore, to prove the double-robustness property of $\hat{\beta}_n$, it suffices to show that the expected value of (10.63) is equal to zero if either the model for $\lambda_r\{G_r(Z),\psi\},\ r\neq\infty$ or the posited model $p^*_Z(z,\xi)$ is correctly specified, which we give by the following theorem.

Theorem 10.5.

$$E\Bigg\{m(Z,\beta_0) - \sum_{r\neq\infty}\Bigg(\left[\frac{I(\mathcal{C}=r) - \lambda_r\{G_r(Z),\psi^*\}I(\mathcal{C}\geq r)}{K_r\{G_r(Z),\psi^*\}}\right] \times \Big[m(Z,\beta_0) - E\{m(Z,\beta_0)|G_r(Z),\xi^*\}\Big]\Bigg)\Bigg\} = 0$$

if either the model for $\lambda_r\{G_r(Z),\psi\},\ r\neq\infty$ or the posited model $p^*_Z(z,\xi)$ is correctly specified.

Proof. By construction, $E\{m(Z,\beta_0)\} = 0$. Therefore, to prove Theorem 10.5, we must show that

$$E\Bigg(\left[\frac{I(\mathcal{C}=r) - \lambda_r\{G_r(Z),\psi^*\}I(\mathcal{C}\geq r)}{K_r\{G_r(Z),\psi^*\}}\right] \Big[m(Z,\beta_0) - E\{m(Z,\beta_0)|G_r(Z),\xi^*\}\Big]\Bigg) = 0, \tag{10.69}$$

for all $r\neq\infty$, if either the model for $\lambda_r\{G_r(Z),\psi\},\ r\neq\infty$ or the posited model $p^*_Z(z,\xi)$ is correctly specified.

We first consider the case when the model for the coarsening probabilities is correctly specified (i.e., $\lambda_r\{G_r(Z),\psi^*\} = \lambda_r\{G_r(Z)\} = P_0(\mathcal{C}=r|\mathcal{C}\geq r,Z)$), whether the posited model $p^*_Z(z,\xi)$ is correct or not. Defining the random vector $\mathcal{F}_r = \{I(\mathcal{C}=1),\ldots,I(\mathcal{C}=r-1),Z\}$, as we did in the proof of Lemma 10.1, and deriving the expectation of (10.69) by first conditioning on $\mathcal{F}_r$, we obtain

$$E\Bigg(\left[\frac{E\{I(\mathcal{C}=r)|\mathcal{F}_r\} - \lambda_r\{G_r(Z)\}I(\mathcal{C}\geq r)}{K_r\{G_r(Z)\}}\right] \times \Big[m(Z,\beta_0) - E\{m(Z,\beta_0)|G_r(Z),\xi^*\}\Big]\Bigg).$$

We showed in (10.46) that $E\{I(\mathcal{C}=r)|\mathcal{F}_r\} = \lambda_r\{G_r(Z)\}I(\mathcal{C}\geq r)$, which proves that (10.69) is equal to zero for all $r\neq\infty$ when the coarsening probabilities are modeled correctly, which, in turn, proves (10.62).

Now, let's consider the case when the posited model for the distribution of Z is correctly specified (i.e., $p^*_Z(z,\xi^*) = p_{0Z}(z)$), whether or not the model for the coarsening probabilities is correct. If this model is correctly specified, then the conditional expectation

$$E\{m(Z,\beta_0)|G_r(Z),\xi^*\} = E_0\{m(Z,\beta_0)|G_r(Z)\}.$$

Write the expectation (10.69) as the difference of two expectations, namely

$$E\left[\frac{I(\mathcal{C}=r)}{K_r\{G_r(Z),\psi^*\}}\right]\Big[m(Z,\beta_0) - E_0\{m(Z,\beta_0)|G_r(Z)\}\Big] \tag{10.70}$$

$$-E\left[\frac{\lambda_r\{G_r(Z),\psi^*\}I(\mathcal{C}\geq r)}{K_r\{G_r(Z),\psi^*\}}\right]\Big[m(Z,\beta_0) - E_0\{m(Z,\beta_0)|G_r(Z)\}\Big]. \tag{10.71}$$

We compute the expectation in (10.71) by first conditioning on $\{I(\mathcal{C} \geq r), G_r(Z)\}$ to obtain

$$E\left[\frac{\lambda_r\{G_r(Z), \psi^*\}}{K_r\{G_r(Z), \psi^*\}}\right]\Big[E\{I(\mathcal{C} \geq r)m(Z, \beta_0)|I(\mathcal{C} \geq r), G_r(Z)\} - I(\mathcal{C} \geq r)E_0\{m(Z, \beta_0)|G_r(Z)\}\Big]. \tag{10.72}$$

But

$$E\{I(\mathcal{C} \geq r)m(Z, \beta_0)|I(\mathcal{C} \geq r), G_r(Z)\} = I(\mathcal{C} \geq r)E\{m(Z, \beta_0)|\mathcal{C} \geq r, G_r(Z)\}. \tag{10.73}$$

Because of the coarsening at random (CAR) assumption, we obtain that

$$p_{Z|\mathcal{C}\geq r, G_r(Z)}(z|g_r) = p_{Z|G_r(Z)}(z|g_r). \tag{10.74}$$

This follows because

$$\begin{aligned} p_{Z|\mathcal{C}\geq r, G_r(Z)}(z|g_r) &= \frac{P(\mathcal{C} \geq r|Z = z)p_{0Z}(z)}{\int_{z:G_r(z)=g_r} P(\mathcal{C} \geq r|Z = z)p_{0Z}(z)d\nu_Z(z)} \\ &= \frac{K_r\{G_r(z)\}p_{0Z}(z)}{\int_{z:G_r(z)=g_r} K_r\{G_r(z)\}p_{0Z}(z)d\nu_Z(z)} \\ &= \frac{p_{0Z}(z)}{\int_{z:G_r(z)=g_r} p_{0Z}(z)d\nu_Z(z)} = p_{Z|G_r(Z)}(z|g_r). \end{aligned}$$

Equation (10.74), together with (10.73), implies that

$$E\{I(\mathcal{C} \geq r)m(Z, \beta_0)|I(\mathcal{C} \geq r), G_r(Z)\} = I(\mathcal{C} \geq r)E_0\{m(Z, \beta_0)|G_r(Z)\}$$

and hence (10.72) and (10.71) are equal to zero. A similar argument, where we condition on $\{I(\mathcal{C} = r), G_r(Z)\}$, can be used to show that (10.70) is equal to zero. This then implies that (10.69) is equal to zero, which, in turn, implies that (10.62) is true, thus demonstrating that $\hat{\beta}_n$ is a consistent estimator for β when the posited model $p^*_Z(z, \xi)$ is correctly specified. □

Example with Longitudinal Data

We return to Example 1, given in Section 9.2, where the interest was in estimating parameters that described the mean CD4 count over time, as a function of treatment, in a randomized study where CD4 counts were measured longitudinally at fixed time points. Specifically, we considered two treatments: $(X = 1)$ was the new treatment and $(X = 0)$ was the control treatment. The response $Y = (Y_1, \ldots, Y_l)^T$ was a vector of CD4 counts that were measured on each individual at times $0 = t_1 < \ldots < t_l$. The full data are denoted by $Z = (Y, X)$. It was assumed that CD4 counts follow a linear trajectory whose

slope may be treatment-dependent. Thus the model was given by (9.24) and assumes that

$$E(Y_{ji}|X_i) = \beta_1 + \beta_2 t_j + \beta_3 X_i t_j.$$

Therefore, the problem was to estimate the parameter $\beta = (\beta_1, \beta_2, \beta_3)^T$ from a sample of data $Z_i = (Y_i, X_i),\ i = 1, \ldots, n$, where $Y_i = (Y_{1i}, \ldots, Y_{li})^T$ are the longitudinally measured CD4 counts for subject i.

In this study, some patients dropped out, and for those patients we observed the CD4 count data prior to dropout, whereas all subsequent CD4 counts are missing. This is an example of monotone coarsening with $\ell = l - 1$ levels of coarsening. We introduce the notation Y^r to denote the vector of data $(Y_1, \ldots, Y_r)^T,\ r = 1, \ldots, l$ and $Y^{\bar{r}}$ to denote the vector of data $(Y_{r+1}, \ldots, Y_l)^T,\ r = 1, \ldots, l-1$. Therefore, when the coarsening variable is $\mathcal{C}_i = r$, we observe $G_r(Z_i) = (X_i, Y_i^r),\ r = 1, \ldots, l-1$, and, when $\mathcal{C}_i = \infty$, we observe the complete data $G_\infty(Z_i) = Z_i = (X_i, Y_i)$.

With such coarsened data $\{\mathcal{C}_i, G_{\mathcal{C}_i}(Z_i)\},\ i = 1, \ldots, n$, we considered estimating the parameter β using an AIPWCC estimator. To accommodate this, we introduced a logistic regression model for the discrete hazard of coarsening probabilities given by (9.25) in terms of a parameter ψ, which was estimated by maximizing the likelihood (8.13). The resulting estimator was denoted by $\hat{\psi}_n$. An AIPWCC estimator for β was proposed by solving the estimating equation (9.26), where, for simplicity, we chose

$$m(Z, \beta) = H^T(X)\{Y - H(X)\beta\}.$$

The definition of the design matrix $H(X)$ is given subsequent to (9.24), and the rationale for this estimating equation is given in Section 9.2. Notice, however, that we defined the augmentation term in equation (9.26) generically using arbitrary functions $L_r\{G_r(Z)\},\ r \neq \infty$. We now know that to improve efficiency as much as possible, we should choose

$$\{L_r\{G_r(Z)\} = Em(Z, \beta)|G_r(Z)\},$$

which requires adaptive estimation using a posited model for $p_Z^*(z, \xi)$. We propose the following.

Assume that the distribution of Y given X follows a multivariate normal distribution whose mean and variance matrix may depend on treatment X; that is,

$$Y|X = 1 \sim MVN(\xi_1, \Sigma_1)$$

and

$$Y|X = 0 \sim MVN(\xi_0, \Sigma_0),$$

where ξ_k is an l-dimensional vector and Σ_k is an $l \times l$ matrix for $k = 0, 1$. The parameter ξ will denote $(\xi_0, \xi_1, \Sigma_0, \Sigma_1)$.

Remark 7. Even though our model puts restrictions on the mean vectors ξ_k, in terms of the parameter β, we will let these be unrestricted and, as it turns out, these parameters will not come into play in our estimating equation. □

We denote $\xi_k^r = (\xi_{1k}, \ldots, \xi_{rk})^T$ and $\xi_k^{\bar{r}} = (\xi_{(r+1)k}, \ldots, \xi_{lk})^T$ for $r = 1, \ldots, l-1$ and $k = 0, 1$. We also denote the corresponding elements of the partitioned matrix for Σ_k by Σ_k^{rr}, $\Sigma_k^{r\bar{r}}$, and $\Sigma_k^{\bar{r}\bar{r}}$ to represent the variance matrix of Y^r, the covariance matrix of Y^r and $Y^{\bar{r}}$, and the variance matrix of $Y^{\bar{r}}$, respectively, given $X = k$. An estimator for ξ can be obtained by maximizing the likelihood

$$\prod_{i=1}^{n}\prod_{k=0}^{1}\Bigg(\prod_{r=1}^{l-1}\Bigg[\{(2\pi)^r|\Sigma_k^{rr}|\}^{-1/2} \times \exp\left\{-\frac{1}{2}(Y_i^r - \xi_k^r)^T(\Sigma_k^{rr})^{-1}(Y_i^r - \xi_k^r)\right\}\Bigg]^{I(\mathcal{C}_i=r, X_i=k)} \times \left[\{(2\pi)^l|\Sigma_k|\}^{-1/2}\exp\left\{-\frac{1}{2}(Y_i - \xi_k)^T\Sigma_k^{-1}(Y_i - \xi_k)\right\}\right]^{I(\mathcal{C}_i=\infty, X_i=k)}\Bigg). \tag{10.75}$$

This likelihood can be maximized by using standard statistical software such as SAS Proc Mixed; see Littell et al. (1996). Denote the estimators for the variance matrix by $\hat{\Sigma}_{kn}$ for $k = 0, 1$. Using standard results for the conditional distribution of a multivariate normal distribution, we obtain that

$$\begin{aligned} E\{m(Z,\beta)|G_r(Z),\xi\} &= E[H^T(X)\{Y - H(X)\beta\}|X, Y^r, \xi] \\ &= H^T(X)q(r, X, Y^r, \beta, \xi), \end{aligned}$$

where $q(r, X, Y^r, \beta, \xi)$ is an l-dimensional vector whose first r elements are $\{Y^r - H^r(X)\beta\}$, whose last $l - r$ elements are

$$(\Sigma_k^{r\bar{r}})^T(\Sigma_k^{rr})^{-1}\{Y^r - H^r(X)\beta\}, \text{ when } X = k,\ k = 0, 1,$$

and $H^r(X)$ is an $r \times 3$ matrix consisting of the first r rows of $H(X)$.

Therefore, the improved double-robust estimator for β is obtained by solving the equation

$$\sum_{i=1}^{n}\Bigg[\frac{I(\mathcal{C}_i = \infty)H^T(X_i)\{Y_i - H(X_i)\beta\}}{\varpi(\infty, Z_i, \hat{\psi}_n)} + \sum_{r=1}^{l-1}\left\{\frac{I(\mathcal{C}_i = r) - \lambda_r\{G_r(Z_i), \hat{\psi}_n\}I(\mathcal{C}_i \geq r)}{K_r\{G_r(Z_i), \hat{\psi}_n\}}\right\} \times H^T(X_i)q(r, X_i, Y_i^r, \beta, \hat{\xi}_n^*)\Bigg] = 0,$$

where, for this example, $\hat{\xi}_n^*$ corresponds to the estimators, $\hat{\Sigma}_{kn}$, $k = 0, 1$, derived by maximizing the likelihood (10.75).

10.4 Remarks Regarding Right Censoring

In Section 9.3, we showed that right censoring, which occurs frequently in survival analysis, can be viewed as a special case of monotone coarsening with continuous-time hazard rates representing the distribution of censoring. We remind the reader that censored-data estimators for the parameter β can also be written as augmented inverse probability weighted complete-case estimators. Specifically, we argued in (9.34) that estimators for β can be derived by solving the estimating equation

$$\sum_{i=1}^{n}\left[\frac{\Delta_i}{K_{U_i}\{\bar{X}_i(U_i)\}}m(Z_i,\beta)+\int_0^{\infty}\frac{dM_{\tilde{C}_i}\{r,\bar{X}_i(r)\}}{K_r\{\bar{X}_i(r)\}}L_r\{\bar{X}_i(r)\}\right]=0, \quad (10.76)$$

where Z denotes the full-data (i.e., $Z=\{T,\bar{X}(T)\}$), and $m(Z,\beta)$ denotes a full-data estimating function that would have been used to obtain estimators for β had there been no censoring. The definition of $dN_{\tilde{C}}(r)$, $\lambda_{\tilde{C}}\{r,\bar{X}(r)\}$, $Y(r)$, $dM_{\tilde{C}}\{r,\bar{X}(r)\}=dN_{\tilde{C}}(r)-\lambda_{\tilde{C}}\{r,\bar{X}(r)\}Y(r)dr$, and $K_r\{\bar{X}(r)\}$ were all defined in Section 9.3.

By analogy between the censored-data estimating equation (10.76) and the monotonically coarsened data estimating equation (10.42), we can show that the most efficient augmented inverse probability weighted complete-case estimator for β that uses the full-data estimating function $m(Z,\beta)$ is obtained by choosing

$$L_r\{\bar{X}(r)\}=E\{m(Z,\beta)|T\geq r,\bar{X}(r)\}.$$

To actually implement these methods with censored data, we need to

1. develop models for the censoring distribution $\lambda_{\tilde{C}}\{r,\bar{X}(r),\psi\}$ and find estimators for ψ, and
2. estimate the conditional expectation $E\{m(Z,\beta)|T\geq r,\bar{X}(r)\}$.

A popular model for the censoring hazard function $\lambda_{\tilde{C}}\{r,\bar{X}(r),\psi\}$ is the semiparametric proportional hazards regression model (Cox, 1972) using maximum partial likelihood estimators to estimate the regression parameters and Breslow's (1974) estimator to estimate the underlying cumulative hazard function.

In order to estimate the conditional expectation $E\{m(Z,\beta)|T\geq r,\bar{X}(r)\}$, we can posit a simpler full-data model, say $p^*_Z(z,\xi)=p^*_{T,\bar{X}(T)}\{t,\bar{x}(t),\xi\}$, and then estimate ξ using the observed data $\{U_i,\Delta_i,\bar{X}_i(U_i)\}, i=1,\ldots,n$ by maximizing the observed-data likelihood

$$\prod_{i=1}^{n}\left[p^*_{T,\bar{X}(T)}\{U_i,\bar{X}_i(U_i),\xi\}\right]^{\Delta_i}\times$$
$$\left[\int_{\{t,\bar{x}(t)\}:t\geq U_i,\{x(s)=X_i(s),s\leq U_i\}}p^*_{T,\bar{X}(T)}\{t,\bar{x}(t),\xi\}d\nu_{T,\bar{X}(T)}\{t,\bar{x}(t)\}\right]^{1-\Delta_i}. \quad (10.77)$$

Building models for $p^*_{T,\bar{X}(T)}\{t, \bar{x}(t), \xi\}$ with time-dependent covariate and maximizing (10.77) can be a daunting task. Nonetheless, the theory that has been developed can often be useful in developing more efficient estimators even if we don't necessarily derive the most efficient one.

In the example of censored medical cost data that was described in Example 2 of Section 9.3, Bang and Tsiatis (2000) used augmented inverse probability weighted complete-case estimators to estimate the mean medical cost and showed various methods for gaining efficiency by judiciously choosing the augmented term.

Other examples where this methodology was used include Robins, Rotnitzky, and Bonetti (2001), who used AIPWCC estimators of the survival distribution under double sampling with follow-up of dropouts. Hu and Tsiatis (1996) and van der Laan and Hubbard (1998) constructed estimators of the survival distribution from survival data that are subject to reporting delays. Zhao and Tsiatis (1997, 1999, 2000) and van der Laan and Hubbard (1999) derived estimators of the quality-adjusted-lifetime distribution from right-censored data. Bang and Tsiatis (2002) derived estimators for the parameters in a median regression model of right-censored medical costs. Strawderman (2000) used these methods to derive an estimator of the mean of an increasing stopped stochastic process. Van der Laan, Hubbard, and Robins (2002) and Quale, van der Laan and Robins (2003) constructed locally efficient estimators of a multivariate survival distribution when failure times are subject to a common censoring time and to failure-time-specific censoring.

10.5 Improving Efficiency when Coarsening Is Nonmonotone

We have discussed how to derive AIPWCC estimators with improved efficiency when there are two levels of coarsening or when the coarsening is monotone and have given several examples to illustrate these methods. This theory can also be extended to the case when the coarsening is nonmonotone. However, we must caution the reader that the use of AIPWCC estimators in this setting is very difficult to implement. At the end of Section 8.1, we already remarked that developing coherent models for the missingness probabilities when the missingness is nonmonotone, is not trivial. There has been very little work in this area, with the exception of the paper by Robins and Gill (1997). Even if one were able to develop models for the missingness probabilities, finding projections onto the augmentation space, as is necessary to obtain more efficient AIPWCC estimators, is not straightforward and requires an iterative process that is numerically difficult to implement. Consequently, the semiparametric theory that leads to AIPWCC estimators has not been well developed with nonmonotone coarsened data and there is still a great deal of research that needs to be done in this area. Nonetheless, many of the theoretical results

have been worked out for nonmonotone coarsened data using the general theory developed by Robins, Rotnitzky, and Zhao (1994). For completeness, we present these results in this section, but again we caution the reader that there are many challenges yet to be tackled before these methods can be feasibly implemented.

Finding the Projection onto the Augmentation Space

We have already argued that among all coarsened-data influence functions given by (10.1) with $\varphi^F(Z)$ fixed, the optimal choice is given by (10.2). We have also shown how to find the projection of $\frac{I(\mathcal{C}=\infty)\varphi^{*F}(Z)}{\varpi(\infty,Z)}$ onto the augmentation space, Λ_2, when there are two levels of coarsening or when the coarsening is monotone, where $\varphi^{*F}(Z) \in \Lambda^{F\perp}$. We now consider how to derive the projection of $\frac{I(\mathcal{C}=\infty)\varphi^{*F}(Z)}{\varpi(\infty,Z)}$ onto Λ_2 in general; that is, even if the coarsening is nonmonotone.

We begin by defining two linear operators.

Definition 4. The linear operator $\mathcal{L}$ is a mapping from the full-data Hilbert space $\mathcal{H}^F$ to the observed-data Hilbert space $\mathcal{H}$, where

$$\mathcal{L} : \mathcal{H}^F \to \mathcal{H}$$

is defined as

$$\mathcal{L}\{h^F(\cdot)\} = E\{h^F(Z)|\mathcal{C}, G_{\mathcal{C}}(Z)\} \tag{10.78}$$

for $h^F(Z) \in \mathcal{H}^F$. Specifically,

$$\mathcal{L}\{h^F(\cdot)\} = \sum_{r=1}^{\infty} I(\mathcal{C}=r)E\{h^F(Z)|\mathcal{C}=r, G_r(Z)\},$$

and because of the coarsening-at-random assumption, we obtain

$$\mathcal{L}\{h^F(Z)\} = \sum_{r=1}^{\infty} I(\mathcal{C}=r)E\{h^F(Z)|G_r(Z)\}. \ \square \tag{10.79}$$

Definition 5. The linear operator $\mathcal{M}$ is a mapping from the full-data Hilbert space to the full-data Hilbert space. Specifically,

$$\mathcal{M} : \mathcal{H}^F \to \mathcal{H}^F$$

is defined as

$$\mathcal{M}\{h^F(\cdot)\} = E[\mathcal{L}\{h^F(\cdot)\}|Z]. \tag{10.80}$$

Using (10.79), we obtain

$$\begin{aligned}\mathcal{M}\{h^F(\cdot)\} &= E\left[\sum_{r=1}^{\infty} I(\mathcal{C}=r)E\{h^F(Z)|G_r(Z)\}|Z\right] \\ &= \sum_{r=1}^{\infty} \varpi\{r, G_r(Z)\}E\{h^F(Z)|G_r(Z)\}. \ \square \end{aligned} \tag{10.81}$$

The projection of $\frac{I(\mathcal{C}=\infty)h^F(Z)}{\varpi(\infty,Z)}$ onto Λ_2 is now given by the following theorem.

Theorem 10.6.

(i) The inverse mapping $\mathcal{M}^{-1}$ exists and is uniquely defined.
(ii) The projection is

$$\Pi\left[\frac{I(\mathcal{C}=\infty)h^F(Z)}{\varpi(\infty,Z)}\middle|\Lambda_2\right] = \frac{I(\mathcal{C}=\infty)h^F(Z)}{\varpi(\infty,Z)} - \mathcal{L}[\mathcal{M}^{-1}\{h^F(\cdot)\}]. \quad (10.82)$$

Proof. Theorem 10.6 part (i)
We will defer the proof of (i) and assume for the time being that it is true.

Proof of Theorem 10.6 part (ii)
If we can show that

a. $\frac{I(\mathcal{C}=\infty)h^F(Z)}{\varpi(\infty,Z)} - \mathcal{L}[\mathcal{M}^{-1}\{h^F(\cdot)\}] \in \Lambda_2$
b. and that

$$\frac{I(\mathcal{C}=\infty)h^F(Z)}{\varpi(\infty,Z)} - \left(\frac{I(\mathcal{C}=\infty)h^F(Z)}{\varpi(\infty,Z)} - \mathcal{L}[\mathcal{M}^{-1}\{h^F(\cdot)\}]\right) = \mathcal{L}[\mathcal{M}^{-1}\{h^F(\cdot)\}]$$

is orthogonal to every element in Λ_2,

then, by the projection theorem, $\frac{I(\mathcal{C}=\infty)h^F(Z)}{\varpi(\infty,Z)} - \mathcal{L}[\mathcal{M}^{-1}\{h^F(\cdot)\}]$ is the unique projection of $\frac{I(\mathcal{C}=\infty)h^F(Z)}{\varpi(\infty,Z)}$ onto Λ_2.

We first note that

$$E\left(\frac{I(\mathcal{C}=\infty)h^F(Z)}{\varpi(\infty,Z)} - \mathcal{L}[\mathcal{M}^{-1}\{h^F(\cdot)\}]\middle|Z\right) = h^F(Z) - \mathcal{M}[\mathcal{M}^{-1}\{h^F(\cdot)\}] = h^F(Z) - h^F(Z) = 0,$$

thus proving (a).

If we let $L_2\{\mathcal{C}, G_{\mathcal{C}}(Z)\}$ be an arbitrary element of Λ_2, then the inner product

$$
\begin{aligned}
&E\Big\{\big(\mathcal{L}[\mathcal{M}^{-1}\{h^F(\cdot)\}]\big)^T L_2\{\mathcal{C}, G_{\mathcal{C}}(Z)\}\Big\} \\
&= E\bigg[E\Big(\big[\mathcal{M}^{-1}\{h^F(\cdot)\}\big]^T L_2\{\mathcal{C}, G_{\mathcal{C}}(Z)\}\Big|\mathcal{C}, G_{\mathcal{C}}(Z)\Big)\bigg] \\
&= E\Big(\big[\mathcal{M}^{-1}\{h^F(\cdot)\}\big]^T L_2\{\mathcal{C}, G_{\mathcal{C}}(Z)\}\Big) \\
&= E\bigg\{E\Big(\big[\mathcal{M}^{-1}\{h^F(\cdot)\}\big]^T L_2\{\mathcal{C}, G_{\mathcal{C}}(Z)\}\Big|Z\Big)\bigg\} \\
&= E\bigg\{\big[\mathcal{M}^{-1}\{h^F(\cdot)\}\big]^T E\Big[L_2\{\mathcal{C}, G_{\mathcal{C}}(Z)\}\Big|Z\Big]\bigg\} = 0,
\end{aligned}
$$

where the last equality follows because $\mathcal{M}^{-1}(h^F) \in \mathcal{H}^F$ and hence as a function of Z, allowing it to come outside the inner conditional expectation, and $L_2(\cdot) \in \Lambda_2$, which implies that $E\{L_2(\cdot)|Z\} = 0$, thus proving (b). □

Uniqueness of $\mathcal{M}^{-1}(\cdot)$

In order to complete the proof of Theorem 10.6, we must show that the linear operator $\mathcal{M}$ has a unique inverse. We will prove the existence and uniqueness of the inverse of the linear mapping $\mathcal{M}$ by showing that the linear operator $(I - \mathcal{M})$ is a contraction mapping, where I denotes the identity mapping $\mathcal{H}^F \to \mathcal{H}^F$; i.e., $I\{h^F(Z)\} = h^F(Z)$ for $h^F \in \mathcal{H}^F$. For more details on these methods, we refer the reader to Kress (1989).

We begin by first defining what we mean by a contraction mapping and proving why $(I - \mathcal{M})$ being a contraction mapping implies that $\mathcal{M}$ has a unique inverse.

Definition 6. A linear operator, say $(I - \mathcal{M})$, is a contraction mapping if the norm of the operator satisfies $\|I - \mathcal{M}\| \leq (1 - \varepsilon)$ for some $\varepsilon > 0$, where the norm of a linear operator, say $\|I - \mathcal{M}\|$, is defined as

$$\sup_{h^F \in \mathcal{H}^F} \frac{\|(I - \mathcal{M})(h^F)\|}{\|h^F\|},$$

or equivalently, $\|I - \mathcal{M}\| \leq (1 - \varepsilon)$, if

$$\|(I - \mathcal{M})(h^F)\| \leq (1 - \varepsilon)\|h^F\| \quad \text{for all } h^F \in \mathcal{H}^F. \square$$

Lemma 10.5. If $(I - \mathcal{M})$ is a contraction mapping, then $\mathcal{M}^{-1}$ exists, is unique, and is equal to the operator

$$S = \sum_{k=0}^{\infty} (I - \mathcal{M})^k. \tag{10.83}$$

Also, $\mathcal{M}^{-1}\{h^F(Z)\}$ can be obtained by successive approximation, where

$$\varphi_{n+1}(Z) = (I - \mathcal{M})\varphi_n(Z) + h^F(Z); \tag{10.84}$$

i.e.,

$$\begin{aligned}
&\varphi_0(Z) = h^F(Z)\\
&\varphi_1(Z) = (I - \mathcal{M})h^F(Z) + h^F(Z)\\
&\varphi_2(Z) = (I - \mathcal{M})^2 h^F(Z) + (I - \mathcal{M})h^F(Z) + h^F(Z)\\
&\quad\vdots
\end{aligned}$$

and $\varphi_n(Z) \to \mathcal{M}^{-1}\{h^F(\cdot)\}$.

Proof. To demonstrate existence, we must show

$$\mathcal{M}[S\{h^F(Z)\}] = h^F(Z).$$

However,

$$\begin{aligned}
\mathcal{M}\big[S\{h^F(Z)\}\big] &= \mathcal{M}\left\{\sum_{k=0}^{\infty}(I - \mathcal{M})^k h^F(Z)\right\}\\
&= \{I - (I - \mathcal{M})\}\left\{\sum_{k=0}^{\infty}(I - \mathcal{M})^k h^F(Z)\right\}.
\end{aligned}$$

By a telescoping argument, this equals

$$\lim_{k\to\infty}\left\{I - (I - \mathcal{M})^k\right\} h^F(Z),$$

but $(I - \mathcal{M})^k h^F(Z)$ will have a norm that converges to zero as $k \to \infty$. This follows because $(I - \mathcal{M})$ is a contraction mapping and $\|(I - \mathcal{M})^k h^F\| \leq (1 - \epsilon)^k \|h^F\|$. Therefore $\lim_{k\to\infty}(I - \mathcal{M})^k h^F(Z) = 0$ a.s. Consequently,

$$\mathcal{M}[S\{h^F(Z)\}] = h^F(Z).$$

We will demonstrate uniqueness by contradiction. Suppose $\mathcal{M}^{-1}(\cdot)$ were not unique. Then there exists $S^*\{h^F(Z)\}$ such that

$$\begin{aligned}
&\mathcal{M}[S^*\{h^F(Z)\}] = h^F(Z) \quad \text{but}\\
&S^*\{h^F(Z)\} \neq S\{h^F(Z)\}.
\end{aligned}$$

In that case,

$$\begin{aligned}
(I - \mathcal{M})(S - S^*)h^F(Z) &= (I - \mathcal{M})[S\{h^F(Z)\} - S^*\{h^F(Z)\}]\\
&= S\{h^F(Z)\} - S^*\{h^F(Z)\}
\end{aligned}$$

and

$$\|(I-\mathcal{M})[S\{h^F(Z)\}-S^*\{h^F(Z)\}]\| = \|S\{h^F(Z)\}-S^*\{h^F(Z)\}\|.$$

But since $(I-\mathcal{M})$ is a contraction mapping, this implies that

$$\|(I-\mathcal{M})[S\{h^F(Z)\}-S^*\{h^F(Z)\}]\| \leq (1-\epsilon)\|S\{h^F(Z)\}-S^*\{h^F(Z)\}\|.$$

This can only happen when

$$S(h^F)-S^*(h^F)=0. \ \square$$

We now complete the proof of Theorem 10.6 by showing that $(I-\mathcal{M})$ is a contraction mapping when $\mathcal{M}$ is defined by (10.80).

Proof. Theorem 10.6 part (i)
Consider the Hilbert space $\mathcal{H}^{\mathcal{C}Z}$ of all q-dimensional mean-zero measurable functions of $(\mathcal{C}, Z)$ equipped with the covariance inner product. The observed-data Hilbert space $\mathcal{H}$ and the full-data Hilbert space $\mathcal{H}^F$ are both contained in $\mathcal{H}^{\mathcal{C}Z}$. That is, $\mathcal{H} \subset \mathcal{H}^{\mathcal{C}Z}, \mathcal{H}^F \subset \mathcal{H}^{\mathcal{C}Z}$ are linear subspaces within this space.

If we consider any arbitrary element $h^{\mathcal{C}Z}(\mathcal{C}, Z) \in \mathcal{H}^{\mathcal{C}Z}$, then

$$\Pi[h^{\mathcal{C}Z}|\mathcal{H}] = E\left\{h^{\mathcal{C},Z}(\mathcal{C}, Z)|\mathcal{C}, G_{\mathcal{C}}(Z)\right\} \tag{10.85}$$

and

$$\Pi[h^{\mathcal{C}Z}|\mathcal{H}^F] = E\left\{h^{\mathcal{C}Z}(\mathcal{C}, Z)|Z\right\}, \tag{10.86}$$

where equations (10.85) and (10.86) can be easily shown to hold by checking that the definitions of a projection are satisfied.

Therefore, deriving $\mathcal{M}\{h(\cdot)\} = \Pi[\Pi[h|\mathcal{H}]|\mathcal{H}^F]$ corresponds to finding two subsequent projections onto these two linear subspaces. What we want to prove is that $(I-\mathcal{M})$ is a contraction mapping from $\mathcal{H}^F$ to $\mathcal{H}^F$. First note that

$$(I-\mathcal{M})h^F(Z) = \Pi[\left\{h^F(Z)-\Pi[h^F(Z)|\mathcal{H}]\right\}|\mathcal{H}^F].$$

Hence, by the Pythagorean theorem,

$$\|(I-\mathcal{M})h^F(Z)\| \leq \|h^F(Z)-\Pi[h^F(Z)|\mathcal{H}]\|. \tag{10.87}$$

Also by the Pythagorean theorem, the right-hand side of (10.87) is equal to

$$\left\{\|h^F(Z)\|^2-\|\Pi\left[h^F(Z)|\mathcal{H}\right]\|^2\right\}^{1/2}. \tag{10.88}$$

The projection $\Pi[h^F(Z)|\mathcal{H}] = E\left\{h^F(Z)|\mathcal{C}, G_{\mathcal{C}}(Z)\right\}$, which, by (10.79), equals

$$I(\mathcal{C}=\infty)h^F(Z)+\sum_{r\neq\infty} I(\mathcal{C}=r)E\{h^F(Z)|G_r(Z)\}.$$

Hence,

$$\|\Pi[h^F(Z)|\mathcal{H}]\|^2 = E\left(\left[I(\mathcal{C}=\infty)h^F(Z) + \sum_{r\neq\infty} I(\mathcal{C}=r)E\{h^F(Z)|G_r(Z)\}\right]^T \right.$$
$$\left.\left[I(\mathcal{C}=\infty)h^F(Z) + \sum_{r\neq\infty} I(\mathcal{C}=r)E\{h^F(Z)|G_r(Z)\}\right]\right)$$
$$= E\Bigg(I(\mathcal{C}=\infty)\{h^F(Z)\}^T h^F(Z) + \sum_{r\neq\infty} I(\mathcal{C}=r)$$
$$\left[E\{h^F(Z)|G_r(Z)\}\right]^T \times \left[E\{h^F(Z)|G_r(Z)\}\right]\Bigg)$$
$$\geq E\left[I(\mathcal{C}=\infty)\{h^F(Z)\}^T h^F(Z)\right]. \tag{10.89}$$

Conditioning on Z, (10.89) equals

$$E\left[\varpi(\infty, Z)\{h^F(Z)\}^T h^F(Z)\right]. \tag{10.90}$$

By assumption, $\varpi(\infty, Z) \geq \epsilon^* > 0$ for all Z, and hence (10.90) is

$$\geq \epsilon^* E[\{h^F(Z)\}^T h^F(Z)] = \epsilon^* \|h^F(Z)\|^2.$$

Consequently, (10.88) is less than or equal to

$$\left\{\|h^F(Z)\|^2 - \epsilon^*\|h^F(Z)\|^2\right\}^{1/2} = (1-\epsilon^*)^{1/2}\|h^F(Z)\|,\ \epsilon^* > 0. \tag{10.91}$$

Therefore, by (10.87), (10.88), and (10.91), we have shown that

$$\|(I-\mathcal{M})h^F(Z)\| \leq (1-\epsilon^*)^{1/2}\|h^F(Z)\|$$

for all $h^F \in \mathcal{H}^F$. Hence $(I-\mathcal{M})$ is a contraction mapping. □

Obtaining Improved Estimators with Nonmonotone Coarsening

In (10.2), we showed that the optimal observed-data influence function of RAL estimators for β associated with the full-data influence function $\varphi^F(Z)$ is obtained by considering the residual after projecting the inverse probability weighted complete-case influence function $\frac{I(\mathcal{C}=\infty)\varphi^F(Z)}{\varpi(\infty,Z)}$ onto Λ_2. When the coarsening is nonmonotone, we demonstrated in the previous section (see (10.82)) that this residual is equal to $\mathcal{L}[\mathcal{M}^{-1}\{\varphi^F(Z)\}]$; that is, $\mathcal{L}[\mathcal{M}^{-1}\{\varphi^F(Z)\}] = \mathcal{J}\{\varphi^F(Z)\}$, where $\mathcal{J}(\varphi^F)$ was defined by (10.5) and is an element of the space of influence functions $(IF)_{\text{DR}}$ (see Definition 2). We also argued (see Remark 2) that if we are interested in deriving more efficient estimators (i.e., estimators whose influence function is an element of $(IF)_{\text{DR}}$), then we should consider estimating functions, which, at the truth, are elements

of the DR linear space $\mathcal{J}(\Lambda^{F\perp})$ (see Definition 3) or, equivalently, the space $\mathcal{L}\{\mathcal{M}^{-1}(\Lambda^{F\perp})\}$.

Consequently, if we defined a full-data estimating function $m(Z,\beta)$ such that $m(Z,\beta_0) \in \Lambda^{F\perp}$, then we should use $\mathcal{L}[\mathcal{M}^{-1}\{m(Z,\beta)\}]$ as our observed-data estimating function. That is, the estimator for β would be the solution to the estimating equation

$$\sum_{i=1}^{n} \mathcal{L}_i[\mathcal{M}_i^{-1}\{m(Z_i,\beta)\}] = 0. \tag{10.92}$$

Of course, deriving the estimating equation in (10.92) is not trivial. The operator $\mathcal{L}(\cdot)$, defined by (10.79), involves finding conditional expectations of functions of Z given $G_r(Z)$ for $r \neq \infty$, and the operator $\mathcal{M}(\cdot)$, defined by (10.81), involves finding such conditional expectations as well as deriving the coarsening probabilities $\varpi\{r, G_r(Z)\}$. Consequently, in order to proceed, these coarsening probabilities and conditional expectations need to be estimated.

The coarsening probabilities are modeled using a parametric model with parameter ψ; that is, $P(\mathcal{C} = r|Z) = \varpi\{r, G_r(Z), \psi\}$, where ψ is estimated by maximizing (8.8) and the resulting estimator is denoted by $\hat{\psi}_n$.

In order to estimate the conditional expectation of functions of Z given $G_r(Z)$ for $r \neq \infty$, we need to estimate the conditional density of Z given $G_r(Z)$. An adaptive strategy is to posit a simplifying parametric model for the density of Z, namely $p_Z^*(z,\xi)$, in terms of a parameter ξ, and then estimate ξ by maximizing (10.58), where the estimator is denoted by $\hat{\xi}_n^*$. We remind the reader that the "$*$" notation is used to emphasize that such a model is not necessarily believed to contain the true distribution of Z. We assume sufficient regularity conditions so that $\hat{\xi}_n^* \xrightarrow{P} \xi^*$, where $p_Z^*(z,\xi^*)$ may or may not be the true density of Z. If it is the true density of Z, we denote this by taking $\xi^* = \xi_0$. With such an estimator for ξ, we can now estimate the conditional expectation $E\{m(Z,\beta)|G_r(Z)\}$ by using $E\{m(Z,\beta)|G_r(Z), \hat{\xi}_n^*\}$, where $E\{m(Z,\beta)|G_r(Z),\xi\}$ is defined by (10.57).

The linear operator $\mathcal{M}$ and its inverse $\mathcal{M}^{-1}$ are functions of the parameters ψ and ξ, and the linear operator $\mathcal{L}$ is a function of ξ. To make this explicit, we define these operators as $\mathcal{M}(\cdot,\psi,\xi)$, $\mathcal{M}^{-1}(\cdot,\psi,\xi)$, and $\mathcal{L}(\cdot,\xi)$. Consequently, the improved estimator for β would be the solution to

$$\sum_{i=1}^{n} \mathcal{L}_i[\mathcal{M}_i^{-1}\{m(Z_i,\beta),\hat{\psi}_n,\hat{\xi}_n^*\},\hat{\xi}_n^*] = 0. \tag{10.93}$$

We will now demonstrate that the estimator for β that solves (10.93) is an example of an AIPWCC estimator. This follows because of the following theorem.

Theorem 10.7. Let $d^F(Z,\beta,\psi,\xi) = \mathcal{M}^{-1}\{m(Z,\beta),\psi,\xi\}$. Then,

$$\mathcal{L}[\mathcal{M}^{-1}\{m(Z,\beta),\psi,\xi\},\xi] = \mathcal{L}\{d^F(Z,\beta,\psi,\xi),\xi\}$$

can be written as

$$\mathcal{L}\{d^F(Z,\beta,\psi,\xi),\xi\} = \frac{I(\mathcal{C}=\infty)m(Z,\beta)}{\varpi(\infty,Z,\psi)} + L_2^*\{\mathcal{C},G_{\mathcal{C}}(Z),\beta,\psi,\xi\}, \quad (10.94)$$

where

$$\begin{aligned} &L_2^*\{\mathcal{C},G_{\mathcal{C}}(Z),\beta,\psi,\xi\} = \\ &-\frac{I(\mathcal{C}=\infty)}{\varpi(\infty,Z,\psi)}\left[\sum_{r\neq\infty}\varpi\{r,G_r(Z),\psi\}E\{d^F(Z,\beta,\psi,\xi)|G_r(Z),\xi\}\right] \\ &+\sum_{r\neq\infty} I(\mathcal{C}=r)E\{d^F(Z,\beta,\psi,\xi)|G_r(Z),\xi\}. \end{aligned} \quad (10.95)$$

Proof. By definition,

$$\begin{aligned} &\mathcal{L}\{d^F(Z,\beta,\psi,\xi),\xi\} = \\ &I(\mathcal{C}=\infty)d^F(Z,\beta,\psi,\xi) + \sum_{r\neq\infty} I(\mathcal{C}=r)E\{d^F(Z,\beta,\psi,\xi)|G_r(Z),\xi\}. \end{aligned} \quad (10.96)$$

Because

$$\begin{aligned} &\mathcal{M}\{d^F(Z,\beta,\psi,\xi),\psi,\xi\} = \mathcal{M}[\mathcal{M}^{-1}\{m(Z,\beta),\psi,\xi\},\psi,\xi] = m(Z,\beta) \\ &= \varpi(\infty,Z,\psi)d^F(Z,\beta,\psi,\xi) + \sum_{r\neq\infty}\varpi\{r,G_r(Z),\psi\}E\{d^F(Z,\beta,\psi,\xi)|G_r(Z),\xi\}, \end{aligned}$$

this implies that

$$\begin{aligned} d^F(Z,\beta,\psi,\xi) =&\{\varpi(\infty,Z,\psi)\}^{-1}\Bigg[m(Z,\beta) \\ &-\sum_{r\neq\infty}\varpi\{r,G_r(Z),\psi\}E\{d^F(Z,\beta,\psi,\xi)|G_r(Z),\xi\}\Bigg]. \end{aligned} \quad (10.97)$$

Substituting the right-hand side of (10.97) for $d^F(Z,\beta,\psi,\xi)$ in (10.96) gives us (10.94) and hence proves the theorem. □

Let us denote $L_{2r}\{G_r(Z)\}$ to be $-E\{d^F(Z,\beta,\psi_0,\xi^*)|G_r(Z),\xi^*\}$. This is a function of $G_r(Z)$ whether the model $p_Z^*(z,\xi^*)$ is correctly specified or not. Therefore, because of (7.37), where elements of the augmentation space are defined, we note that $L_2^*\{\mathcal{C},G_{\mathcal{C}}(Z),\beta,\psi_0,\xi^*\}$, defined by (10.95), is an element of the augmentation space Λ_2 as long as the model for the coarsening probabilities $\varpi\{r,G_r(Z),\psi_0\}$ is correctly specified, regardless of whether the posited model for the density of Z, $p_Z^*(z,\xi^*)$, is or not. Finally, because of Theorem 10.7, the estimator (10.93) is the same as the solution to

$$\sum_{i=1}^n \frac{I(\mathcal{C}_i=\infty)m(Z_i,\beta)}{\varpi(\infty,Z,\hat{\psi}_n)} + L_2^*\{\mathcal{C}_i,G_{\mathcal{C}_i}(Z_i),\beta,\hat{\psi}_n,\hat{\xi}_n^*\} = 0 \quad (10.98)$$

and therefore is an AIPWCC estimator.

The result above assumes that we can derive $\mathcal{M}^{-1}\{m(Z,\beta),\psi,\xi\}$. In Lemma 10.5, we showed that the inverse operator $\mathcal{M}^{-1}$ exists. Nonetheless, this inverse operator is not necessarily easy to compute with nonmonotone coarsened data, and an iterative procedure using successive approximations was also given in Lemma 10.5. Therefore, let us denote, by $d^F_{(j)}(Z,\beta,\psi,\xi)$ the approximation of $\mathcal{M}^{-1}\{m(Z,\beta),\psi,\xi\}$ after, say, (j) iterations of successive approximations given by (10.84). Because $d^F_{(j)}(Z,\beta,\psi,\xi)$ is not exactly $\mathcal{M}^{-1}\{m(Z,\beta),\psi,\xi\}$, then $\mathcal{L}\{d^F_{(j)}(Z,\beta_0,\psi_0,\xi^*),\xi^*\}$ may not be an element of $\Lambda^\perp$ and therefore not appropriate as the basis of an estimating function. We therefore suggest the following strategy.

Define

$$\begin{aligned}
&L^*_{2(j)}\{\mathcal{C},G_{\mathcal{C}}(Z),\beta,\psi,\xi\}\\
&=-\frac{I(\mathcal{C}=\infty)}{\varpi(\infty,Z,\psi)}\left[\sum_{r\neq\infty}\varpi\{r,G_r(Z),\psi\}E\{d^F_{(j)}(Z,\beta,\psi,\xi)|G_r(Z),\xi\}\right]\\
&\quad+\sum_{r\neq\infty}I(\mathcal{C}=r)E\{d^F_{(j)}(Z,\beta,\psi,\xi)|G_r(Z),\xi\}. \qquad (10.99)
\end{aligned}$$

By construction, $L^*_{2(j)}\{\mathcal{C},G_{\mathcal{C}}(Z),\beta,\psi_0,\xi^*\}$ is an element of Λ_2, whether $d^F_{(j)}(Z,\beta,\psi,\xi)$ equals $\mathcal{M}^{-1}\{m(Z,\beta),\psi,\xi\}$ or not. This implies that

$$\frac{I(\mathcal{C}=\infty)m(Z,\beta_0)}{\varpi(\infty,Z,\psi_0)}+L^*_{2(j)}\{\mathcal{C},G_{\mathcal{C}}(Z),\beta_0,\psi_0,\xi^*\}$$

is guaranteed to be an element of $\Lambda^\perp$ when ψ_0 is correctly specified.

By defining $L^*_{2(j)}(\cdot)$ in this manner, we are guaranteed that the solution to the equation

$$\sum_{i=1}^n\frac{I(\mathcal{C}_i=\infty)m(Z_i,\beta)}{\varpi(\infty,Z,\hat\psi_n)}+L^*_{2(j)}\{\mathcal{C}_i,G_{\mathcal{C}_i}(Z_i),\beta,\hat\psi_n,\hat\xi^*_n\}=0 \qquad (10.100)$$

is an AIPWCC estimator. Moreover, because of Theorem 10.7, if we take the number of iterations (j) to be sufficiently large so that $d^F_{(j)}(Z,\beta,\psi,\xi)$ is equal (or as close as we want) to $\mathcal{M}^{-1}\{m(Z,\beta),\psi,\xi\}$, then solving equation (10.100) will lead to the estimator (10.93).

As long as the model for the coarsening probabilities is correctly specified, the estimator, (10.100), for β, under suitable regularity conditions, will be an RAL estimator for β with influence function

$$\left[\frac{I(\mathcal{C}=\infty)\varphi^F(Z)}{\varpi(\infty,Z,\psi_0)}+L^{**}_{2(j)}\{\mathcal{C},G_{\mathcal{C}}(Z),\beta_0,\psi_0,\xi^*\}\right]-\Pi\left[[\cdot]\middle|\Lambda_\psi\right], \qquad (10.101)$$

where

$$\varphi^F(Z) = \left[-E\left\{\frac{\partial m(Z,\beta_0)}{\partial \beta^T}\right\}\right]^{-1} m(Z,\beta_0),$$

$$L^{**}_{2(j)}\{\mathcal{C}, G_{\mathcal{C}}(Z), \beta_0, \psi_0, \xi^*\}$$
$$= \left[-E\left\{\frac{\partial m(Z,\beta_0)}{\partial \beta^T}\right\}\right]^{-1} L^{*}_{2(j)}\{\mathcal{C}, G_{\mathcal{C}}(Z), \beta_0, \psi_0, \xi^*\},$$

and ξ^* denotes the limit (in probability) of $\hat{\xi}^*_n$.

If, in addition, the posited model $p^*_Z(z,\xi)$ contains the truth, then the influence function (10.101) is equal to $\mathcal{L}[\mathcal{M}^{-1}\{\varphi^F(Z)\}] = \mathcal{J}\{\varphi^F(Z)\}$ and hence is an element of $(IF)_{\rm DR}$.

Double Robustness

In constructing the estimator in (10.100), we took the point of view that the coarsening probabilities are correctly specified. We also defined a posited model for Z, namely $p^*_Z(z,\xi)$, for the purpose of constructing more efficient estimators. Such a model, for instance, enabled us to construct functions $d^F(Z)$ and $E\{d^F(Z)|G_r(Z)\}$, which were used to derive projections onto the space Λ_2. As we showed above, the model for $p^*_Z(z,\xi)$ does not need to be correctly specified in order for our estimator to be consistent and asymptotically normal as long as the model for the coarsening probabilities is correct.

However, as we will now demonstrate, the attempt to gain efficiency also gives us the added protection of double robustness. That is, if the posited model for the density of Z is correct (i.e., the true density of Z, $p_0(z)$, is contained in the model $p^*_Z(z,\xi)$ for some ξ, which we denote by ξ_0), and if we choose $d^F(Z,\beta,\psi,\xi)$ to be exactly $\mathcal{M}^{-1}\{m(Z,\beta),\psi,\xi\}$, then the estimator β that is the solution to (10.98) will be consistent and asymptotically normal even if the model for the coarsening probabilities is not correct.

Such a double-robustness property was shown previously for two levels of missingness or for monotone coarsening. This result now generalizes the double-robustness property for all coarsened-data models where the coarsening mechanism is CAR.

In order for the double-robustness property to hold, we emphasize that $d^F(Z,\beta,\psi,\xi) = \mathcal{M}^{-1}\{m(Z,\beta),\psi,\xi\}$, which, as we showed in the previous section, would result in the estimating equation (10.98) being identical to the estimating equation (10.93). With sufficient regularity conditions, the solution to (10.93) will lead to a consistent, asymptotically normal estimator for β if the estimating function of (10.93), $\mathcal{L}[\mathcal{M}^{-1}\{m(Z,\beta),\psi,\xi\},\xi]$, evaluated at $\beta = \beta_0$, $\psi = \psi^*$, $\xi = \xi^*$, where ψ^* and ξ^* are the probabilistic limits of $\hat{\psi}_n$ and $\hat{\xi}^*_n$ respectively, is an unbiased estimator of zero. Namely, we must show that

$$E\Big(\mathcal{L}[\mathcal{M}^{-1}\{m(Z,\beta_0),\psi^*,\xi^*\},\xi^*]\Big) = 0. \tag{10.102}$$

Remark 8. The expectation in (10.102) is with respect to the truth. We remind the reader that the true coarsened-data density involves both the marginal density of Z, $p_0(z)$, and the model for the coarsening probabilities, $P_0(\mathcal{C} = r|Z) = \varpi_0\{r, G_r(Z)\}$. We are considering the case where we posit a model for the marginal density, namely $p^*_Z(z, \xi)$, which might be incorrect. We denote this situation by letting the estimator $\hat{\xi}^*_n$ converge in probability to ξ^*, where $p^*_Z(z, \xi^*) \neq p_0(z)$. However, in the special case where the posited model is correctly specified, we will denote this by taking $\xi^* = \xi_0$, where $p^*_Z(z, \xi_0) = p_0(z)$. We also posit a model for the coarsening probabilities; namely, $P(\mathcal{C} = r|Z) = \varpi\{r, G_r(Z), \psi\}$. Using the same convention, if this model is misspecified, we denote this by letting the estimator $\hat{\psi}_n$ converge in probability to ψ^*, where $\varpi\{r, G_r(Z), \psi^*\} \neq \varpi_0\{r, G_r(Z)\}$. If this model is correctly specified, then we take $\psi^* = \psi_0$, where $\varpi\{r, G_r(Z), \psi_0\} = \varpi_0\{r, G_r(Z)\}$. To emphasize this notation, we write the expectation in (10.102) as

$$E_{\xi_0,\psi_0}\Big(\mathcal{L}[\mathcal{M}^{-1}\{m(Z,\beta_0),\psi^*,\xi^*\},\xi^*]\Big). \ \square$$

To demonstrate double robustness, we need to prove the following theorem.

Theorem 10.8.

(i) $E_{\xi_0,\psi_0}\Big(\mathcal{L}[\mathcal{M}^{-1}\{m(Z,\beta_0),\psi_0,\xi^*\},\xi^*]\Big) = 0,$

and

(ii) $E_{\xi_0,\psi_0}\Big(\mathcal{L}[\mathcal{M}^{-1}\{m(Z,\beta_0),\psi^*,\xi_0\},\xi_0]\Big) = 0.$

Proof of (i)
Because the conditional distribution of $\mathcal{C}|Z$ involves the parameter ψ only and the marginal distribution of Z involves ξ only, we can write (i) as

$$E_{\xi_0}\left\{E_{\psi_0}\left(\mathcal{L}[\mathcal{M}^{-1}\{m(Z,\beta_0),\psi^*,\xi_0\},\xi_0]\Big|Z\right)\right\}. \tag{10.103}$$

But, for any function $q(Z)$, $E_{\psi_0}[\mathcal{L}\{q(Z),\xi^*\}|Z]$ is, by definition, equal to $\mathcal{M}\{q(Z),\psi_0,\xi^*\}$. Therefore, (10.103) equals

$$\begin{aligned} &E_{\xi_0}\left(\mathcal{M}\left[\mathcal{M}^{-1}\{m(Z,\beta_0),\psi_0,\xi^*\},\psi_0,\xi^*\right]\right) \\ &= E_{\xi_0}\{m(Z,\beta_0)\} = 0. \quad \square \end{aligned}$$

Proof of (ii)
Because $\mathcal{L}\{q(Z),\xi_0\} = E_{\xi_0}\{q(Z)|\mathcal{C}, G_{\mathcal{C}}(Z)\}$, then

$$E_{\xi_0,\psi_0}\left[\mathcal{L}\{q(Z),\xi_0\}\right] = E_{\xi_0}\{q(Z)\}.$$

Notice that the argument above didn't involve the parameter ψ; hence

$$E_{\xi_0,\psi_0}\left[\mathcal{L}\{q(Z),\xi_0\}\right] = E_{\xi_0,\psi^*}\left[\mathcal{L}\{q(Z),\xi_0\}\right] = E_{\xi_0}\{q(Z)\}$$

for any parameter ψ^*. Applying this to the left-hand side of equation (ii), we obtain

$$\begin{aligned}
&E_{\xi_0,\psi_0}\Big(\mathcal{L}[\mathcal{M}^{-1}\{m(Z,\beta_0),\psi^*,\xi_0\},\xi_0]\Big)\\
&= E_{\xi_0,\psi^*}\Big(\mathcal{L}[\mathcal{M}^{-1}\{m(Z,\beta_0),\psi^*,\xi_0\},\xi_0]\Big)\\
&= E_{\xi_0}\left\{E_{\psi^*}\left(\mathcal{L}[\mathcal{M}^{-1}\{m(Z,\beta_0),\psi^*,\xi_0\},\xi_0]\middle| Z\right)\right\}\\
&= E_{\xi_0}\left(\mathcal{M}\left[\mathcal{M}^{-1}\{m(Z,\beta_0),\psi^*,\xi_0\},\psi^*,\xi_0\right]\right)\\
&= E_{\xi_0}\{m(Z,\beta_0)\} = 0. \quad \square
\end{aligned}$$

Remark 9. As we indicated earlier, in order for our estimator to be double robust, we must make sure that if the posited model $p_Z^*(z,\xi)$ contains the truth, then the estimator for ξ, namely $\hat{\xi}_n^*$, is a consistent estimator for ξ_0, even if the model for the coarsening probabilities is misspecified. Therefore, likelihood estimators such as those that maximize (10.58) would be appropriate for this purpose, whereas AIPWCC estimators for ξ would not. $\square$

10.6 Recap and Review of Notation

General results

- Among observed-data influence functions $\varphi\{\mathcal{C}, G_{\mathcal{C}}(Z)\} \in (IF)$,

  $$\varphi\{\mathcal{C}, G_{\mathcal{C}}(Z)\} = \left\{\left[\frac{I(\mathcal{C}=\infty)\varphi^F(Z)}{\varpi(\infty,Z,\psi_0)} + L_2\{\mathcal{C}, G_{\mathcal{C}}(Z)\}\right] - \Pi\{[\ ]|\Lambda_\psi\},\right.$$

  $$\left.\text{where } \varphi^F(Z) \text{ is a full-data influence function and } L_2\{\mathcal{C}, G_{\mathcal{C}}(Z)\} \in \Lambda_2\right\},$$

  optimal influence functions can be obtained by taking

  $$L_2\{\mathcal{C}, G_{\mathcal{C}}(Z)\} = -\Pi\left[\frac{I(\mathcal{C}=\infty)\varphi^F(Z)}{\varpi(\infty,Z,\psi_0)}\middle|\Lambda_2\right].$$

- We denote such influence functions using the linear operator $\mathcal{J}(\cdot)$, where $\mathcal{J}: \mathcal{H}^F \to \mathcal{H}$ is defined as

  $$\mathcal{J}(\varphi^F) = \frac{I(\mathcal{C}=\infty)\varphi^F(Z)}{\varpi(\infty,Z,\psi_0)} - \Pi\left[\frac{I(\mathcal{C}=\infty)\varphi^F(Z)}{\varpi(\infty,Z,\psi_0)}\middle|\Lambda_2\right].$$

- We denote the class of influence functions

$$\mathcal{J}\{(IF)^F\} = \left\{ \mathcal{J}(\varphi^F) : \varphi^F(Z) \in (IF)^F \right\}$$

  by $(IF)_{\text{DR}}$ (i.e., the space of double-robust influence functions).
- The corresponding linear space used to derive estimating functions that lead to observed-data estimators for β with influence functions in $(IF)_{\text{DR}}$ is denoted by the DR linear space and is defined as $\mathcal{J}(\Lambda^{F\perp})$.

Two levels of missingness

- Let the full data be given by $Z = (Z_1^T, Z_2^T)^T$, where Z_1 is always observed but Z_2 may be missing. Let R denote the complete-case indicator, and $P(R = 1|Z) = \pi(Z_1, \psi)$ is a model that describes the complete-case probabilities. Then, for $\varphi^{*F}(Z) \in \Lambda^{F\perp}$,

$$\mathcal{J}(\varphi^{*F}) = \left[\frac{R\varphi^{*F}(Z)}{\pi(Z_1, \psi_0)} - \left\{ \frac{R - \pi(Z_1, \psi_0)}{\pi(Z_1, \psi_0)} \right\} E\{\varphi^{*F}(Z)|Z_1\} \right].$$

- Adaptive double-robust AIPWCC estimators for β are obtained by solving the equation

$$\sum_{i=1}^{n} \left[\frac{R_i m(Z_i, \beta)}{\pi(Z_{1i}, \hat{\psi}_n)} - \left\{ \frac{R_i - \pi(Z_{1i}, \hat{\psi}_n)}{\pi(Z_{1i}, \hat{\psi}_n)} \right\} h_2^*(Z_{1i}, \beta, \hat{\xi}_n^*) \right] = 0,$$

  where $m(Z, \beta)$ is a full-data estimating function, $\hat{\psi}_n$ is the MLE for ψ obtained by maximizing

$$\prod_{i=1}^{n} \{\pi(Z_{1i}, \psi)\}^{R_i} \{1 - \pi(Z_{1i}, \psi)\}^{1-R_i},$$

  $\hat{\xi}_n^*$ is an estimator for the parameter ξ in a posited model $p_Z^*(z, \xi)$, and

$$h_2^*(Z_{1i}, \beta, \xi) = E\left\{ m(Z_i, \beta)|Z_{1i}, \xi \right\}.$$

Monotone coarsening

- When data are monotonically coarsened, the coarsening probabilities can be modeled, as a function of the parameter ψ, using the discrete hazard probability of coarsening,

$$\lambda_r\{G_r(Z), \psi\} = P(\mathcal{C} = r|\mathcal{C} \geq r, Z, \psi), r \neq \infty.$$

- For $\varphi^{*F}(Z) \in \Lambda^{F\perp}$,

$$\mathcal{J}(\varphi^{*F}) = \frac{I(\mathcal{C}=\infty)\varphi^{*F}(Z)}{\varpi(\infty, Z, \psi_0)} + \sum_{r\neq\infty}\left[\frac{I(\mathcal{C}=r) - \lambda_r\{G_r(Z),\psi_0\}I(\mathcal{C}\geq r)}{K_r\{G_r(Z),\psi_0\}}\right] E\left\{\varphi^{*F}(Z)|G_r(Z)\right\},$$

where

$$K_r\{G_r(Z),\psi_0\} = P(\mathcal{C}>r|Z,\psi_0) = \prod_{r'=1}^{r}[1-\lambda_{r'}\{G_{r'}(Z),\psi_0\}], r\neq\infty,$$

and

$$\varpi(\infty, Z, \psi_0) = \prod_{r\neq\infty}[1-\lambda_r\{G_r(Z),\psi_0\}].$$

- Adaptive double-robust AIPWCC estimators for β are obtained by solving the equation

$$\sum_{i=1}^{n}\left(\frac{I(\mathcal{C}_i=\infty)m(Z_i,\beta)}{\varpi(\infty,Z_i,\hat{\psi}_n)} + \sum_{r\neq\infty}\left[\frac{I(\mathcal{C}_i=r)-\lambda_r\{G_r(Z_i),\hat{\psi}_n\}I(\mathcal{C}_i\geq r)}{K_r\{G_r(Z_i),\hat{\psi}_n\}}\right] E\{m(Z,\beta)|G_r(Z_i),\hat{\xi}_n^*\}\right)=0,$$

where $m(Z,\beta)$ is a full-data estimating function, $\hat{\psi}_n$ is the MLE for ψ, and $\hat{\xi}_n^*$ is an estimator for the parameter ξ in a posited model $p_Z^*(z,\xi)$.

Nonmonotone coarsening

- In general, for $\varphi^{*F}(Z) \in \Lambda^{F\perp}$,

$$\mathcal{J}(\varphi^{*F}) = \mathcal{L}\{\mathcal{M}^{-1}(\varphi^{*F})\},$$

where

$\mathcal{L}(h^F) = E\{h^F(Z)|\mathcal{C}, G_{\mathcal{C}}(Z)\}$ is equal to

$$\sum_r I(\mathcal{C}=r)E\{h^F(Z)|G_r(Z)\},$$

$\mathcal{M}(h^F) = E\{\mathcal{L}(h^F)|Z\}$ is equal to

$$\sum_r \varpi\{r, G_r(Z)\}E\{h^F(Z)|G_r(Z)\},$$

and where the inverse operator $\mathcal{M}^{-1}(h^F)$ exists and can be obtained by successive approximation, where

$$\varphi_{n+1}(Z) = (I - \mathcal{M})\varphi_n(Z) + h^F(Z),$$

and $\varphi_n(Z) \xrightarrow{n\to\infty} \mathcal{M}^{-1}\{h^F(Z)\}$.

- More efficient adaptive AIPWCC estimators for β can be obtained by solving the equation

$$\sum_{i=1}^{n} \frac{I(\mathcal{C}_i = \infty)m(Z_i, \beta)}{\varpi(\infty, Z, \hat{\psi}_n)} + L^*_{2(j)}\{\mathcal{C}_i, G_{\mathcal{C}_i}(Z_i), \beta, \hat{\psi}_n, \hat{\xi}^*_n\} = 0,$$

where $m(Z, \beta)$ is a full-data estimating function, $\hat{\psi}_n$ is the MLE for ψ, $\hat{\xi}^*_n$ is an estimator for the parameter ξ in a posited model $p^*_Z(z, \xi)$,

$$\begin{aligned} &L^*_{2(j)}\{\mathcal{C}, G_{\mathcal{C}}(Z), \beta, \psi, \xi\} \\ &= -\frac{I(\mathcal{C} = \infty)}{\varpi(\infty, Z, \psi)} \left[\sum_{r \neq \infty} \varpi\{r, G_r(Z), \psi\} E\{d^F_{(j)}(Z, \beta, \psi, \xi) | G_r(Z), \xi\} \right] \\ &\quad + \sum_{r \neq \infty} I(\mathcal{C} = r) E\{d^F_{(j)}(Z, \beta, \psi, \xi) | G_r(Z), \xi\}, \end{aligned}$$

and $d^F_{(j)}(Z, \beta, \psi, \xi)$ is the approximation to $\mathcal{M}^{-1}\{m(Z, \beta, \psi, \xi)\}$ after (j) iterations of successive approximations.

- As $j \to \infty$, the AIPWCC estimator becomes a double-robust estimator.

10.7 Exercises for Chapter 10

1. In Definition 6 of Chapter 3, we defined a q-replicating linear space. In Theorem 10.1, we considered the linear space $M = \Pi[\Lambda_2 | \Lambda^{\perp}_{\psi}] \subset \mathcal{H}$.
   a) Prove that Λ_2 is a q-replicating linear space.
   b) Prove that Λ_ψ is a q-replicating linear space. (Recall that Λ_ψ is the finite-dimensional linear space, contained in $\mathcal{H}$, that is spanned by the score vector $S_\psi\{\mathcal{C}, G_{\mathcal{C}}(Z)\}$.)
   c) Prove that M is a q-replicating linear space.
2. When we considered monotone coarsening, we stated that the adaptive estimator $\hat{\beta}_n$, which solves equation (10.60), is asymptotically equivalent to the AIPWCC estimator $\hat{\beta}^*_n$, which solves equation (10.61), when the model for the coarsening probabilities is correctly specified. Give a heuristic proof that
$$n^{1/2}(\hat{\beta}_n - \hat{\beta}^*_n) \xrightarrow{P} 0.$$
(You can use arguments similar to the proof of Theorem 10.3.)

3. Consider the simple linear regression restricted moment model where with full data $(Y_i, X_{1i}, X_{2i}), i = 1, \ldots, n$, we assume

$$E(Y_i|X_{1i}, X_{2i}) = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i}.$$

In such a model, we can estimate the parameters $(\beta_0, \beta_1, \beta_2)^T$ using ordinary least squares; that is, the solution to the estimating equation

$$\sum_{i=1}^{n} (1, X_{1i}, X_{2i})^T (Y_i - \beta_0 - \beta_1 X_{1i} - \beta_2 X_{2i}) = 0. \tag{10.104}$$

In fact, this estimator is locally efficient when $\text{var}(Y_i|X_{1i}, X_{2i})$ is constant. The data, however, are missing at random with a monotone missing pattern. That is, Y_i is observed on all individuals in the sample; however, for some individuals, only X_{2i} is missing, and for others both X_{1i} and X_{2i} are missing. Therefore, we define the missingness indicator

$$\begin{aligned} &(\mathcal{C}_i = 1) \text{ if we only observe } Y_i, \\ &(\mathcal{C}_i = 2) \text{ if we only observe } (Y_i, X_{1i}), \end{aligned}$$

and

$$(\mathcal{C}_i = \infty) \text{ if we observe } (Y_i, X_{1i}, X_{2i}).$$

We will define the missingness probability model using discrete-time hazards, namely

$$\begin{aligned} \lambda_1(Y) &= P(\mathcal{C} = 1|Y), \\ \lambda_2(Y, X_1) &= P(\mathcal{C} = 2|\mathcal{C} \geq 2, Y, X_1). \end{aligned}$$

a) In terms of λ_1 and λ_2, what is

$$P(\mathcal{C} = \infty|Y, X_1, X_2)?$$

In order to model the missingness process, we assume logistic regression models; namely,

$$\text{logit}\{\lambda_1(Y)\} = \psi_{10} + \psi_{11} Y, \text{where logit}(p) = \log\left(\frac{p}{1-p}\right),$$

and

$$\text{logit}\{\lambda_2(Y, X_1)\} = \psi_{20} + \psi_{21} X_1 + \psi_{22} Y.$$

b) Using some consistent notation to describe the observed data, write out the estimating equations that need to be solved to derive the maximum likelihood estimator for

$$\psi = (\psi_{10}, \psi_{11}, \psi_{20}, \psi_{21}, \psi_{22})^T.$$

c) Describe the linear subspace Λ_ψ.
d) Describe the linear subspace Λ_2. Verify that $\Lambda_\psi \subset \Lambda_2$.
e) Describe the subspace $\Lambda^\perp$, the linear space orthogonal to the observed-data nuisance tangent space. An initial estimator for β can be obtained by using an inverse probability weighted complete-case estimator that solves the equation

$$\sum_{i=1}^{n} \frac{I(\mathcal{C}_i = \infty)}{\varpi(\infty, Y_i, X_{1i}, X_{2i}, \hat{\psi}_n)}(1, X_{1i}, X_{2i})^T(Y_i - \beta_0 - \beta_1 X_{1i} - \beta_2 X_{2i}) = 0,$$

where $\hat{\psi}_n$ is the maximum likelihood estimator derived in (b). Denote this estimator by $\hat{\beta}_n^I$.
f) What is the i-th influence function for $\hat{\beta}_n^I$?
g) Derive a consistent estimator for the asymptotic variance of $\hat{\beta}_n^I$.
In an attempt to gain efficiency, we consider

$$\frac{I(\mathcal{C}_i = \infty)}{\varpi(\infty, Y_i, X_{1i}, X_{2i}, \psi_o)}\varphi^{*F}(Y_i, X_{1i}, X_{2i}) - \Pi\left[\frac{I(\mathcal{C}_i = \infty)\varphi^{*F}(Y_{1i}, X_{1i}, X_{2i})}{\varpi(\infty, Y_i, X_{1i}, X_{2i}, \psi_o)}\middle|\Lambda_2\right],$$

where $\varphi^{*F}(\cdot) = (1, X_{1i}, X_{2i})^T(Y_i - \beta_0 - \beta_1 X_{1i} - \beta_2 X_{2i})$.
h) Compute

$$\Pi\left[\frac{I(\mathcal{C}_i = \infty)\varphi^{*F}(Y_{1i}, X_{1i}, X_{2i})}{\varpi(\infty, Y_i, X_{1i}, X_{2i}, \psi_o)}\middle|\Lambda_2\right].$$

In practice, we need to estimate (h) using a simplifying model. For simplicity, let us use as a working model that $(Y_i, X_{1i}, X_{2i})^T$ is multivariate normal with mean $(\mu_Y, \mu_{X_1}, \mu_{X_2})^T$ and covariance matrix

$$\begin{bmatrix} \sigma_{YY} & \sigma_{YX_1} & \sigma_{YX_2} \\ \sigma_{YX_1} & \sigma_{X_1X_2} & \sigma_{X_1X_2} \\ \sigma_{YX_2} & \sigma_{X_1X_2} & \sigma_{X_2X_2} \end{bmatrix}.$$

i) With the observed data, how would you estimate the parameters in the multivariate normal?
j) Assuming the simplifying multivariate normal model and the estimates derived in (i), estimate the projection in (h).
k) Write out the estimating equation that needs to be solved to get an improved estimator.
l) Find a consistent estimator for the asymptotic variance of the estimator in (k). (Keep in mind that the simplifying model of multivariate normality may not be correct.)

11

Locally Efficient Estimators for Coarsened-Data Semiparametric Models

Using semiparametric theory, we have demonstrated that RAL estimators for the parameter β in a semiparametric model with coarsened data can be obtained using AIPWCC estimators. That is, estimators for β can be obtained from a sample of coarsened data $\{\mathcal{C}_i, G_{\mathcal{C}_i}(Z_i)\}, i = 1, \ldots, n$, by solving the estimating equation

$$\sum_{i=1}^{n} \left[\frac{I(\mathcal{C}_i = \infty) m(Z_i, \beta)}{\varpi(\infty, Z_i, \hat{\psi}_n)} + L_2\{\mathcal{C}_i, G_{\mathcal{C}_i}(Z_i), \beta, \hat{\psi}_n\} \right] = 0, \tag{11.1}$$

where $m(Z, \beta)$ is a full-data estimating function, $L_2(\cdot)$ is an element of the augmentation space, Λ_2, and $\hat{\psi}_n$ is an estimator for the parameters in the coarsening model. In Chapter 10, we demonstrated that, among the AIPWCC estimators, improved double-robust estimators for β can be obtained by considering observed-data estimating functions within the class $\mathcal{J}(\Lambda^{F\perp})$ (i.e., the so-called DR linear space), where, for $\varphi^{*F} \in \Lambda^{F\perp}$,

$$\mathcal{J}(\varphi^{*F}) = \frac{I(\mathcal{C} = \infty)\varphi^{*F}}{\varpi(\infty, Z)} - \Pi\left[\frac{I(\mathcal{C} = \infty)\varphi^{*F}}{\varpi(\infty, Z)} | \Lambda_2 \right].$$

This led us to develop adaptive estimators that were the solution to

$$\sum_{i=1}^{n} \left[\frac{I(\mathcal{C}_i = \infty) m(Z_i, \beta)}{\varpi(\infty, Z_i, \hat{\psi}_n)} + L_2^*\{\mathcal{C}_i, G_{\mathcal{C}_i}(Z_i), \beta, \hat{\psi}_n, \hat{\xi}_n^*\} \right] = 0, \tag{11.2}$$

where $L_2^*\{\mathcal{C}, G_{\mathcal{C}}(Z), \beta, \psi, \xi\}$ is equal to minus the projection onto the augmentation space; i.e.,

$$-\Pi\left[\frac{I(\mathcal{C} = \infty) m(Z, \beta)}{\varpi(\infty, Z)} | \Lambda_2 \right].$$

To compute this projection, we need estimates for the parameter ψ that describes the coarsening probabilities and an estimate for the marginal distribution of Z. The latter is accomplished by positing a simpler, and possibly

incorrect, model for the density of Z as $p^*_Z(z,\xi)$ and deriving an estimator $\hat{\xi}^*_n$ for ξ.

Among the class of double-robust estimators is the efficient estimator, the estimator that achieves the semiparametric efficiency bound. Finding this efficient estimator within this class of double-robust estimators entails deriving the proper choice of the full-data estimating function $m(Z,\beta)$, where $m(Z,\beta_0) = \varphi^{*F} \in \Lambda^{F\perp}$. In this chapter, we will study how to find the efficient estimator and the appropriate choice for $m(Z,\beta)$.

As we will see, the efficient estimator will depend on the true marginal distribution of Z, which, of course, is unknown to us. Consequently, we will develop adaptive methods where the efficient estimator will be computed based on a posited model $p^*_Z(z,\xi)$ for the density of Z. Hence, the proposed methods will lead to a locally efficient estimator, an estimator for β that will achieve the semiparametric efficiency bound if the posited model is correct but will still be a consistent, asymptotically normal RAL semiparametric estimator for β even if the posited model does not contain the truth.

As we indicated in Chapter 10, finding improved double-robust estimators often involves computationally intensive methods. In fact, when the coarsening of the data is nonmonotone, these computational challenges could be overwhelming. Similarly here, deriving locally efficient estimators involves numerical difficulties. Nonetheless, the theory developed by Robins, Rotnitzky, and Zhao (1994) gives us a prescription for how to derive locally efficient estimators. We present this theory in this chapter and discuss strategies for finding locally efficient estimators. The methods build on the full-data semiparametric theory. Therefore, it will be assumed that we have a good understanding of the full-data semiparametric model. That is, we can identify the space orthogonal to the full-data nuisance tangent space $\Lambda^{F\perp}$, the class of full-data influence functions $(IF)^F$, the full-data efficient score $S^F_{\text{eff}}(Z,\beta_0)$, and the full-data influence function $\varphi^F_{\text{eff}}(Z)$.

However, we caution the reader that these methods may be very difficult to implement in practice, and we believe a great deal of research still needs to be done in developing feasible computational algorithms. In Chapter 12, we will discuss approximations that may be used to derive AIPWCC estimators for β that although not locally efficient are easier to implement and can result in substantial gains in efficiency.

There is, however, one class of problems where locally efficient estimators are obtained readily, and this is the case when only one full-data influence function exists. This occurs, for example, when the full-data tangent space is the entire Hilbert space $\mathcal{H}^F$, as is the case when no restrictions are put on the class of densities for Z; i.e., the nonparametric problem (see Theorem 4.4). In Section 5.3, we showed that only one full-data influence function exists when we are interested in estimating the mean of a random variable in a nonparametric problem and that this estimator can be obtained using the sample average. When only one full-data influence function, $\varphi^F(Z)$, exists, then the class of observed-data influence functions is given by

$$\left\{ \left[\frac{I(\mathcal{C}=\infty)\varphi^F(Z)}{\varpi(\infty, Z, \psi_0)} + L_2\{\mathcal{C}, G_{\mathcal{C}}(Z)\} \right] - \Pi\{[\ ]|\Lambda_\psi\},\ L_2(\cdot) \in \Lambda_2 \right\},$$

and because of Theorem 10.1, the optimal observed-data influence function is

$$\frac{I(\mathcal{C}=\infty)\varphi^F(Z)}{\varpi(\infty, Z, \psi_0)} - \Pi\left[\frac{I(\mathcal{C}=\infty)\varphi^F(Z)}{\varpi(\infty, Z, \psi_0)} \middle| \Lambda_2 \right],$$

which is also the efficient influence function. Consequently, when there is only one full-data influence function, then the adaptive double-robust estimator outlined in Chapter 10 will lead to a locally efficient estimator. We illustrate with a simple example.

Example: Estimating the Mean with Missing Data

In Section 7.4, we considered a problem where interest focused on estimating the relationship of a response variable Y as a function of covariates X. Because one of the covariates X_2 was expensive to collect, only a subsample of this covariate was collected with probability $\pi(Y, X_1)$, which depends on the response Y and the other covariates X_1. Thus, for this problem, $X = (X_1^T, X_2)^T$ and the full data $Z = (Y, X)$. The probability $\pi(Y, X_1)$ was chosen by the investigator and therefore this is an example of two levels of missingness by design. The complete-case indicator is denoted by R, where $P(R = 1|Z) = P(R = 1|Y, X_1) = \pi(Y, X_1)$, and the observed data are denoted by $O = (R, Y, X_1, RX_2)$. The statistical question, as originally stated in Section 7.4, was to estimate the parameter in a restricted moment model with a sample of observed data $(R_i, Y_i, X_{1i}, R_i X_{2i}),\ i = 1, \ldots, n$.

However, we now want to consider the simpler problem of estimating the mean of X_2 using the observed data. Let us denote this parameter as $\beta = E(X_2)$. Also, to be as robust to model misspecification as possible, we no longer assume any specific relationship between the response Y and the covariates X. Consequently, with no assumptions on the joint distribution of the full data (Y, X_1, X_2) (i.e., the nonparametric problem), we know that there is only one full-data influence function of RAL estimators for $\beta = E(X_2)$, namely $\varphi^F(Z) = (X_2 - \beta_0)$, and that the solution to the full-data estimating equation is

$$\sum_{i=1}^{n} (X_{2i} - \beta) = 0;$$

i.e., the sample mean $\hat{\beta}_n^F = n^{-1} \sum_{i=1}^n X_{2i}$ is an RAL full-data estimator for β with this influence function.

We also know that all observed-data influence functions for this problem are given by

$$\left\{ \frac{R\varphi^F(Z)}{\pi(Y, X_1)} + \frac{R - \pi(Y, X_1)}{\pi(Y, X_1)} L(Y, X_1) \right\}, \tag{11.3}$$

where $L(Y, X_1)$ is an arbitrary function of Y and X_1, and, because of Theorem 10.2, the optimal choice for $L(Y, X_1)$ is given by $-E(X_2|Y, X_1)$. Therefore, among the class of influence functions (11.3), the one with the smallest variance is

$$\left\{ \frac{R\varphi^F(Z)}{\pi(Y, X_1)} - \frac{R - \pi(Y, X_1)}{\pi(Y, X_1)} E(X_2|Y, X_1) \right\}. \tag{11.4}$$

This also is the semiparametric efficient observed-data influence function.

Since the efficient influence function depends on $E(X_2|Y, X_1)$, we consider an adaptive strategy. Using methods described in Section 10.2 for adaptive estimation with two levels of missingness, we posit a model for the conditional distribution of X_2 given (Y, X_1). One simple model we may consider is that the distribution of X_2 given (Y, X_1) is normally distributed with mean $\xi_1 + \xi_2 Y + \xi_3^T X_1$ and variance σ_ξ^2. This is attractive because the MLE estimator for the parameter ξ (i.e., the estimator that maximizes (10.16)) can be obtained using ordinary least squares among the complete cases $\{i : R_i = 1\}$. That is, the estimator for ξ is obtained as the solution to the estimating equation

$$\sum_{i=1}^{n} R_i (1, Y_i, X_{1i}^T)^T (X_{2i} - \xi_1 - \xi_2 Y_i - \xi_3^T X_{1i}) = 0.$$

Denote the least-squares estimator for ξ by $\hat{\xi}_n^*$. Then, the adaptive observed-data estimator for β is given as the solution to

$$\sum_{i=1}^{n} \left\{ \frac{R_i}{\pi(Y_1, X_{1i})} (X_{2i} - \beta) - \frac{R_i - \pi(Y_1, X_{1i})}{\pi(Y_1, X_{1i})} (\hat{\xi}_{1n}^* + \hat{\xi}_{2n}^* Y_i + \hat{\xi}_{2n}^{*T} X_{1i} - \beta) \right\} = 0,$$

which, after solving, yields

$$\hat{\beta}_n = n^{-1} \sum_{i=1}^{n} \left\{ \frac{R_i X_{2i}}{\pi(Y_1, X_{1i})} - \frac{R_i - \pi(Y_1, X_{1i})}{\pi(Y_1, X_{1i})} (\hat{\xi}_{1n}^* + \hat{\xi}_{2n}^* Y_i + \hat{\xi}_{2n}^{*T} X_{1i}) \right\}. \tag{11.5}$$

This estimator is a consistent, asymptotically normal observed-data RAL estimator for β regardless of whether the posited model is correct or not. Moreover, if the posited model is correctly specified, then this estimator is semiparametric efficient. Therefore, (11.5) is a locally efficient semiparametric estimator for $\beta = E(X_2)$. We also note that because the least-squares estimator leads to a consistent estimator for ξ, if the conditional expectation of X_2 given Y and X_1 is linear, namely

$$E(X_2|Y, X_1) = \xi_1 + \xi_2 Y + \xi_3^T X_1, \tag{11.6}$$

whether the distribution is normal or not, the locally efficient estimator $\hat{\beta}_n$ is also fully efficient whenever (11.6) is satisfied.

11.1 The Observed-Data Efficient Score

As we know, the efficient RAL estimator for β will have an influence function that is proportional to the efficient score. Therefore, it is useful to study the properties of the efficient score with coarsened data. Toward that end, we give two different representations for the efficient score. The first is likelihood-based and the second is based on AIPWCC estimators. The relationship of these two representations to each other will be key in the development of the proposed adaptive locally efficient estimators.

Representation 1 (Likelihood-Based)

We remind the reader that the efficient observed-data estimator for β has an influence function that is proportional to the observed-data efficient score and that the efficient score is unique and equal to

$$S_{\text{eff}}\{\mathcal{C}, G_{\mathcal{C}}(Z)\} = S_{\beta}\{\mathcal{C}, G_{\mathcal{C}}(Z)\} - \Pi[S_{\beta}\{\mathcal{C}, G_{\mathcal{C}}(Z)\}|\Lambda],$$

where $S_{\beta}\{\mathcal{C}, G_{\mathcal{C}}(Z)\} = E\{S_{\beta}^{F}(Z)|\mathcal{C}, G_{\mathcal{C}}(Z)\}$, and $\Lambda = \Lambda_{\psi} \oplus \Lambda_{\eta}$, $\Lambda_{\psi} \perp \Lambda_{\eta}$. Because $\Lambda_{\psi} \perp \Lambda_{\eta}$, this implies that

$$\Pi[S_{\beta}\{\mathcal{C}, G_{\mathcal{C}}(Z)\}|\Lambda] = \Pi[S_{\beta}\{\mathcal{C}, G_{\mathcal{C}}(Z)\}|\Lambda_{\psi}] + \Pi[S_{\beta}\{\mathcal{C}, G_{\mathcal{C}}(Z)\}|\Lambda_{\eta}].$$

The same argument that was used to show that $\Lambda_{\eta} \perp \Lambda_2$ in Theorem 8.2 can also be used to show that $S_{\beta}\{\mathcal{C}, G_{\mathcal{C}}(Z)\} \perp \Lambda_2$. Since $\Lambda_{\psi} \subset \Lambda_2$, this implies that $S_{\beta}\{\mathcal{C}, G_{\mathcal{C}}(Z)\}$ is orthogonal to Λ_{ψ}. Therefore,

$$\Pi[S_{\beta}\{\mathcal{C}, G_{\mathcal{C}}(Z)\}|\Lambda_{\psi}] = 0, \tag{11.7}$$

which implies that

$$\Pi[S_{\beta}\{\mathcal{C}, G_{\mathcal{C}}(Z)\}|\Lambda] = \Pi[S_{\beta}\{\mathcal{C}, G_{\mathcal{C}}(Z)\}|\Lambda_{\eta}],$$

and the efficient score is

$$S_{\text{eff}}\{\mathcal{C}, G_{\mathcal{C}}(Z)\} = S_{\beta}\{\mathcal{C}, G_{\mathcal{C}}(Z)\} - \Pi[S_{\beta}\{\mathcal{C}, G_{\mathcal{C}}(Z)\}|\Lambda_{\eta}]. \tag{11.8}$$

Recall that

$$\Lambda_{\eta} = \left\{ E\left\{\alpha^{F}(Z)|\mathcal{C}, G_{\mathcal{C}}(Z)\right\} \text{ for all } \alpha^{F}(Z) \in \Lambda^{F} \right\}.$$

This means that the unique projection $\Pi[S_{\beta}\{\mathcal{C}, G_{\mathcal{C}}(Z)\}|\Lambda_{\eta}]$ corresponds to some element in Λ_{η}, which we will denote by

$$E\left\{\alpha_{\text{eff}}^{F}(Z)|\mathcal{C}, G_{\mathcal{C}}(Z)\right\}, \ \alpha_{\text{eff}}^{F}(Z) \in \Lambda^{F}.$$

With this representation,

$$\begin{aligned} S_{\text{eff}}\{\mathcal{C}, G_{\mathcal{C}}(Z)\} &= E\left\{S_{\beta}^{F}(Z)|\mathcal{C}, G_{\mathcal{C}}(Z)\right\} - E\left\{\alpha_{\text{eff}}^{F}(Z)|\mathcal{C}, G_{\mathcal{C}}(Z)\right\} \\ &= E[\{S_{\beta}^{F}(Z) - \alpha_{\text{eff}}^{F}(Z)\}|\mathcal{C}, G_{\mathcal{C}}(Z)]. \end{aligned} \tag{11.9}$$

Remark 1. The full-data efficient score is given by

$$S^F_{\text{eff}}(Z) = S^F_\beta(Z) - \Pi[S^F_\beta(Z)|\Lambda^F].$$

However, the element $\alpha^F_{\text{eff}}(Z)$ is not necessarily the same as $\Pi[S^F_\beta(Z)|\Lambda^F]$. This means that, in general,

$$S_{\text{eff}}\{\mathcal{C}, G_{\mathcal{C}}(Z)\} \neq E\left\{S^F_{\text{eff}}(Z)|\mathcal{C}, G_{\mathcal{C}}(Z)\right\}. \square \tag{11.10}$$

Representation 2 (AIPWCC-Based)

In (10.2), we derived the optimal (smallest variance matrix) influence function among the class of AIPWCC influence functions associated with a fixed full-data influence function $\varphi^F(Z)$. Consequently, we can restrict our search for the efficient observed-data influence function to the class of influence functions

$$\left\{\frac{I(\mathcal{C}=\infty)\varphi^F(Z)}{\varpi(\infty, Z, \psi_0)} - \Pi\left[\frac{I(\mathcal{C}=\infty)\varphi^F(Z)}{\varpi(\infty, Z, \psi_0)}\middle|\Lambda_2\right] : \varphi^F(Z) \in (IF)^F\right\},$$

which we denote as the space $(IF)_{\text{DR}} = \mathcal{J}\{(IF)^F\}$. Since the observed-data efficient score is defined up to a proportionality constant matrix times the efficient influence function, this implies that the observed-data efficient score must be an element in the DR linear space $\mathcal{J}(\Lambda^{F\perp})$,

$$\left\{\frac{I(\mathcal{C}=\infty)B^F(Z)}{\varpi(\infty, Z, \psi_0)} - \Pi\left[\frac{I(\mathcal{C}=\infty)B^F(Z)}{\varpi(\infty, Z, \psi_0)}\middle|\Lambda_2\right] : \text{ for } B^F(Z) \in \Lambda^{F\perp}\right\}, \tag{11.11}$$

with the corresponding element denoted by $B^F_{\text{eff}}(Z)$.

Relationship between the Two Representations

Thus, we have shown that there are two equivalent representations for the observed-data efficient score, which are given by (11.9) and (11.11):

(i) $S_{\text{eff}}\{\mathcal{C}, G_{\mathcal{C}}(Z)\} = E\left[\left\{S^F_\beta(Z) - \alpha^F_{\text{eff}}(Z)\right\}|\mathcal{C}, G_{\mathcal{C}}(Z)\right]$, where $\alpha^F_{\text{eff}}(Z) \in \Lambda^F$,

and

(ii) $S_{\text{eff}}\{\mathcal{C}, G_{\mathcal{C}}(Z)\} = \dfrac{I(\mathcal{C}=\infty)B^F_{\text{eff}}(Z)}{\varpi(\infty, Z)} - \Pi\left[\dfrac{I(\mathcal{C}=\infty)B^F_{\text{eff}}(Z)}{\varpi(\infty, Z)}\middle|\Lambda_2\right]$, where $B^F_{\text{eff}}(Z) \in \Lambda^{F\perp}$.

Remark 2. If the full-data model were parametric (i.e., if η were finite-dimensional), then representation (i) would correspond to the estimating function that would be used to derive the coarsened-data MLE for β. This may

be preferable, as it does not involve the parameter ψ specifying the coarsening process. However, if the model is complicated, this approach may become formidable due to the curse of dimensionality. The second representation leads us to augmented inverse-probability weighted complete-case (AIPWCC) estimating equations. These estimators, which build on full-data estimators, may be easier to derive, even in some complicated situations. However, this approach requires that the data analyst model the coarsening process and estimate the parameter ψ. Also, to obtain gains in efficiency, an adaptive approach is required, where the data analyst posits simpler models for the full data $p_Z^*(z, \xi)$, in terms of the parameter ξ that needs to be estimated. Which method is preferable often depends on the specific application. However, because of the robustness of the AIPWCC estimators to misspecification, we will focus attention on these estimators.

Nonetheless, the two representations will aid us in getting a better understanding of the geometry of observed-data efficient influence functions and guide us in finding as good an AIPWCC estimator as is feasible. □

If we knew, or could reasonably approximate, the element $B_{\text{eff}}^F(Z) \in \Lambda^{F\perp}$ of representation (ii) above, then we could construct an AIPWCC estimator for β by using the full-data estimating function $m(Z, \beta)$, where $m(Z, \beta_0) = B_{\text{eff}}^F(Z)$, and applying the methods outlined in Chapter 10. Subject to the accuracy of different posited models, this methodology will give us as efficient an AIPWCC estimator as possible while still affording us maximum robustness to misspecification. Toward that end, we now show how to derive $B_{\text{eff}}^F(Z)$ in the following theorem.

Theorem 11.1. The element $B_{\text{eff}}^F(Z)$ is the unique $B^F(Z) \in \Lambda^{F\perp}$ that solves the equation

$$\Pi[\mathcal{M}^{-1}\{B^F(Z)\}|\Lambda^{F\perp}] = S_{\text{eff}}^F(Z), \tag{11.12}$$

where $\mathcal{M}(\cdot)$ denotes the linear operator, given by Definition 5 of Chapter 10, equation (10.80), which maps $\mathcal{H}^F$ (full-data Hilbert space) to $\mathcal{H}^F$ as

$$\mathcal{M}\{h^{F(q\times 1)}(Z)\} = E\big[E\big\{h^F(Z)\big|\mathcal{C}, G_{\mathcal{C}}(Z)\big\}\big|Z\big].$$

Proof. Because of the equivalence of the two representations (i) and (ii) above, the efficient score can be written as

$$\begin{aligned} &E\left[\left\{S_\beta^F(F) - \alpha_{\text{eff}}^F(Z)\right\}\big|\mathcal{C}, G_{\mathcal{C}}(Z)\right] \\ &= \frac{I(\mathcal{C}=\infty)B_{\text{eff}}^F(Z)}{\varpi(\infty, Z)} - \Pi\left[\frac{I(\mathcal{C}=\infty)B_{\text{eff}}^F(Z)}{\varpi(\infty, Z)}\bigg|\Lambda_2\right] \end{aligned} \tag{11.13}$$

for some $\alpha_{\text{eff}}^F(Z) \in \Lambda^F$ and $B_{\text{eff}}^F(Z) \in \Lambda^{F\perp}$. Taking the conditional expectation of both sides of equation (11.13) with respect to Z, we obtain the equation

$$\mathcal{M}\big\{S_\beta^F(Z) - \alpha_{\text{eff}}^F(Z)\big\} = B_{\text{eff}}^F(Z). \tag{11.14}$$

In Theorem 10.6 and Lemma 10.5, we showed that the linear operator $\mathcal{M}(\cdot)$ has a unique inverse, $\mathcal{M}^{-1}$. Therefore, we can write (11.14) as

$$\mathcal{M}^{-1}\{B^F_{\text{eff}}(Z)\} = \{S^F_\beta(Z) - \alpha^F_{\text{eff}}(Z)\}. \tag{11.15}$$

Projecting both sides of equation (11.15) onto $\Lambda^{F\perp}$, we obtain

$$\Pi[\mathcal{M}^{-1}\{B^F_{\text{eff}}(Z)\}|\Lambda^{F\perp}] = \underset{||}{\Pi[S^F_\beta(Z)|\Lambda^{F\perp}]} - \underbrace{\underset{||}{\Pi[\alpha^F_{\text{eff}}(Z)|\Lambda^{F\perp}]}}.$$

$$S^F_{\text{eff}}(Z) \qquad \underset{\text{since } \alpha^F_{\text{eff}}(Z) \in \Lambda^F}{0}$$

This leads us to the important relationship that the efficient element $B^F_{\text{eff}}(Z) \in \Lambda^{F\perp}$, which is used to construct the efficient score

$$S_{\text{eff}}\{\mathcal{C}, G_{\mathcal{C}}(Z)\} = \frac{I(\mathcal{C}=\infty)B^F_{\text{eff}}(Z)}{\varpi(\infty, Z)} - \Pi\left[\frac{I(\mathcal{C}=\infty)B^F_{\text{eff}}(Z)}{\varpi(\infty, Z)}|\Lambda_2\right], \tag{11.16}$$

must satisfy the relationship (11.12).

We still need to show that a unique $B^F_{\text{eff}}(Z) \in \Lambda^{F\perp}$ exists that satisfies (11.12) and find a method for computing $B^F_{\text{eff}}(Z)$. □

Uniqueness of $\boldsymbol{B^F_{\text{eff}}(Z)}$

Lemma 11.1. There exists a unique $B^F_{\text{eff}}(Z) \in \Lambda^{F\perp}$ that solves the equation

$$\Pi[\mathcal{M}^{-1}\{B^F_{\text{eff}}(Z)\}|\Lambda^{F\perp}] = S^F_{\text{eff}}(Z), \tag{11.17}$$

and this solution can be obtained through successive approximations.

Proof. Notice that (11.17) involves a mapping from the linear subspace $\Lambda^{F\perp} \subset \mathcal{H}^F$ to $\Lambda^{F\perp} \subset \mathcal{H}^F$. The way we will prove this lemma is by defining another linear mapping $(I-\mathcal{Q})\{\mathcal{M}^{-1}\}(\cdot) : \mathcal{H}^F \to \mathcal{H}^F$, which maps the entire full-data Hilbert space to the entire full-data Hilbert space in such a way that

(i) $(I-\mathcal{Q})\{\mathcal{M}^{-1}\}(\cdot)$ coincides with the mapping $\Pi[\mathcal{M}^{-1}(h^F)|\Lambda^{F\perp}]$ whenever $h^F \in \Lambda^{F\perp}$, and
(ii) $(I-\mathcal{Q})$ is a contraction mapping and hence has a unique inverse.

Define

$$\mathcal{D}^F_{\text{eff}}(Z) = \mathcal{M}^{-1}\{B^F_{\text{eff}}(Z)\}.$$

Because of the existence and uniqueness of $\mathcal{M}^{-1}$, if $B^F_{\text{eff}}(Z)$ exists, then so does $\mathcal{D}^F_{\text{eff}}(Z)$ such that

$$\mathcal{M}\{\mathcal{D}^F_{\text{eff}}(Z)\} \in \Lambda^{F\perp}. \tag{11.18}$$

Motivated by the fact that $\mathcal{D}^F_{\text{eff}}(Z)$ must satisfy

(a) equation (11.17), namely

$$\Pi[\mathcal{D}^F_{\text{eff}}(Z)|\Lambda^{F\perp}] = S^F_{\text{eff}}(Z), \tag{11.19}$$

or, equivalently,

$$\{I(\cdot) - \Pi[I(\cdot)|\Lambda^F]\}\{\mathcal{D}^F_{\text{eff}}(Z)\} = S^F_{\text{eff}}(Z), \tag{11.20}$$

where we view $\{I(\cdot) - \Pi[I(\cdot)|\Lambda^F]\}(\cdot)$ as a linear operator from $\mathcal{H}^F$ to $\mathcal{H}^F$ with $I(\cdot)$ denoting the identity operator,

and

(b) equation (11.18), or, equivalently,

$$\Pi[\mathcal{M}(\cdot)|\Lambda^F]\{\mathcal{D}^F_{\text{eff}}(Z)\} = 0, \tag{11.21}$$

where $\Pi[\mathcal{M}(\cdot)|\Lambda^F](\cdot)$ is also viewed as a linear operator from $\mathcal{H}^F$ to $\mathcal{H}^F$,

we combine (11.19)–(11.21) to consider the equations

$$S_{\text{eff}}(Z) = \Pi[h^F(Z)|\Lambda^{F\perp}] + \Pi[\mathcal{M}\{h^F(Z)\}|\Lambda^F] \tag{11.22}$$

$$= (I - \mathcal{Q})\{h^F(Z)\}, \tag{11.23}$$

where $(I - \mathcal{Q})(\cdot)$ is a linear operator, with $\mathcal{Q}(\cdot)$ defined as

$$\mathcal{Q}\{\mathcal{D}^F_{\text{eff}}(Z)\} = \Pi[(I - \mathcal{M})\{\mathcal{D}^F_{\text{eff}}(Z)\}\big|\Lambda^F].$$

We first argue that the solution, $h^F(Z) \in \mathcal{H}^F$, to equation (11.23) exists and is unique and then argue that this solution $h^F(Z)$ must equal $\mathcal{D}^F_{\text{eff}}(Z)$.

According to Lemma 10.5, the linear operator $(I-\mathcal{Q})(\cdot)$ will have a unique inverse if we can show that the linear operator "$\mathcal{Q}$" is a contraction mapping. Also, if $\mathcal{Q}$ is a contraction mapping, then by Lemma 10.5, the unique inverse is equal to

$$(I - \mathcal{Q})^{-1} = \sum_{i=0}^{\infty} \mathcal{Q}^i.$$

In the proof of Theorem 10.6, part (i), we already showed that $(I - \mathcal{M})$ is a contraction mapping; i.e.,

$$||(I - \mathcal{M})\{h^F\}|| \leq (1 - \varepsilon)||h^F||. \tag{11.24}$$

By the Pythagorean theorem,

$$\begin{aligned} ||\mathcal{Q}(h^F)|| &= ||\Pi[(I - \mathcal{M})h^F\big|\Lambda^F]|| \\ &\leq ||(I - \mathcal{M})h^F|| \\ &\leq (1 - \varepsilon)||h^F|| \quad \text{by (11.24)}. \end{aligned}$$

Hence, $\mathcal{Q}$ is a contraction mapping and $(I - \mathcal{Q})^{-1}$ exists and is unique.

To complete the proof, we must show that the unique solution $h^F(Z)$ to equation (11.23) or, equivalently, (11.22), is identical to $\mathcal{D}^F_{\text{eff}}(Z)$ satisfying (11.19) and (11.21).

Clearly, any element $\mathcal{D}^F_{\text{eff}}(Z)$ satisfying (11.19) and (11.21) must satisfy (11.22). Conversely, since $S_{\text{eff}}(Z) \in \Lambda^{F\perp}$, then the solution $h^F(Z)$ of (11.22) must be such that $\Pi[\mathcal{M}\{h^F(Z)\}|\Lambda^F] = 0$ and $\Pi[h^F(Z)|\Lambda^{F\perp}] = S_{\text{eff}}(Z)$; that is, $h^F(Z)$ satisfies (11.21) and (11.19), respectively. This completes the proof that $\mathcal{D}^F_{\text{eff}}(Z)$ exists and is the unique element satisfying equations (11.19) and (11.21) or, equivalently, that $B_{\text{eff}}(Z) = \mathcal{M}\{\mathcal{D}^F_{\text{eff}}(Z)\}$ exists and is the unique solution to (11.17).

In Lemma 10.5, we showed that the solution $\mathcal{D}^F_{\text{eff}}(Z)$ can be obtained by successive approximation; that is,

$$\mathcal{D}^{(i+1)}(Z) = \Pi\big[(I - \mathcal{M})\mathcal{D}^{(i)}(Z)\big|\Lambda^F\big] + S^F_{\text{eff}}(Z), \tag{11.25}$$

and

$$\mathcal{D}^{(i)}(Z) \xrightarrow{i\to\infty} \mathcal{D}^F_{\text{eff}}(Z).$$

If we define

$$B^{(i)}(Z) = \Pi\big[\mathcal{M}\{\mathcal{D}^{(i)}(Z)\}\big|\Lambda^{F\perp}\big],$$

where, by construction, $B^{(i)}(Z) \in \Lambda^{F\perp}$, then

$$B^{(i)}(Z) = \underbrace{\mathcal{M}\{\mathcal{D}^{(i)}(Z)\}}_{\Downarrow} - \underbrace{\Pi\big[\mathcal{M}\{\mathcal{D}^{(i)}(Z)\}\big|\Lambda^F\big]}_{\Downarrow} \xrightarrow{i\to\infty} B^F_{\text{eff}}(Z). \tag{11.26}$$

$$\mathcal{M}\{\mathcal{D}^F_{\text{eff}}(Z)\} = B^F_{\text{eff}}(Z) \qquad \Pi\left[B^F_{\text{eff}}(Z)|\Lambda^F\right] \qquad \square$$

$$\Big|\Big|$$

$$0$$

The inverse operator $\mathcal{M}^{-1}$ plays an important role in the definition of $B^F_{\text{eff}}(Z)$ given by Theorem 11.1. As we showed in Chapter 10, $\mathcal{M}^{-1}$ exists, is unique, and can be computed using successive approximations. However, a closed-form solution exists when the coarsening is monotone. For completeness, we give this result.

$\mathcal{M}^{-1}$ for Monotone Coarsening

Recall that, for monotone coarsening, we defined coarsening probabilities using discrete hazard probabilities

$$\lambda_r\{G_r(Z)\} = P(\mathcal{C} = r|\mathcal{C} \geq r, Z)$$

and

$$K_r\{G_r(Z)\} = P(\mathcal{C} \geq r+1|Z) = \prod_{j=1}^{r}[1 - \lambda_j\{G_j(Z)\}].$$

Theorem 11.2. When coarsening is monotone, the inverse operator $\mathcal{M}^{-1}$ is given by

$$a^F(Z) = \mathcal{M}^{-1}\{h^F(Z)\} = \frac{h^F(Z)}{\varpi(\infty, Z)} - \sum_{r \neq \infty} \frac{\lambda_r}{K_r} E\{h^F(Z)|G_r(Z)\}, \quad (11.27)$$

where we use the shorthand notation $\lambda_r = \lambda_r\{G_r(Z)\}$ and $K_r = K_r\{G_r(Z)\}$. An equivalent representation is also given by

$$\mathcal{M}^{-1}\{h^F(Z)\} = h^F(Z) + \sum_{r \neq \infty} \frac{\lambda_r}{K_r} \left[h^F(Z) - E\{h^F(Z)|G_r(Z)\}\right]. \quad (11.28)$$

Proof. In Theorem 10.6, we showed that $\mathcal{M}^{-1}$ exists and is uniquely defined. Therefore, we only need to show that $\mathcal{M}\{a^F(Z)\} = h^F(Z)$, where $a^F(Z)$ is defined by (11.27) and

$$\mathcal{M}(a^F) = \varpi_\infty a^F + \sum_{r \neq \infty} \varpi_r E(a^F|G_r).$$

We note that

$$\frac{\lambda_r}{K_r} = \left(\frac{1}{K_r} - \frac{1}{K_{r-1}}\right) \quad (11.29)$$

and

$$\varpi_\infty = K_\ell,$$

where ℓ denotes the number of coarsening levels; i.e., ℓ denotes the largest value of $\mathcal{C}$ less than ∞. After substituting $\left(\frac{1}{K_r} - \frac{1}{K_{r-1}}\right)$ for $\frac{\lambda_r}{K_r}$ and $\frac{1}{K_\ell}$ for ϖ_∞ in (11.27) and rearranging terms, we obtain

$$a^F = \sum_r \frac{1}{K_{r-1}} \Big\{ E(h^F|G_r) - E(h^F|G_{r-1}) \Big\}, \quad (11.30)$$

where $K_0 = 1$ and $E(h^F|G_0) = 0$. Therefore,

$$\begin{aligned} \mathcal{M}(a^F) &= \sum_{r'} \varpi_{r'} E\left[\sum_r \frac{1}{K_{r-1}} \left\{ E(h^F|G_r) - E(h^F|G_{r-1}) \right\} \middle| G_{r'} \right] \\ &= \sum_{r'} \varpi_{r'} E\left[\sum_r \left\{ E\left(\frac{h^F}{K_{r-1}} \middle| G_r \right) - E\left(\frac{h^F}{K_{r-1}} \middle| G_{r-1} \right) \right\} \middle| G_{r'} \right]. \end{aligned} \quad (11.31)$$

As a consequence of monotone coarsening and the laws of conditional expectations, we obtain that

$$
\begin{aligned}
&E\left[\left\{E\left(\frac{h^F}{K_{r-1}}\middle|G_r\right)-E\left(\frac{h^F}{K_{r-1}}\middle|G_{r-1}\right)\right\}\middle|G_{r'}\right]\\
&\quad= \qquad 0 \qquad\qquad\qquad \text{if } r^{'}<r\\
&\quad=\frac{1}{K_{r-1}}\left\{E(h^F|G_r)-E(h^F|G_{r-1})\right\} \text{ if } r^{'}\geq r. \qquad (11.32)
\end{aligned}
$$

Substituting (11.32) into (11.31), we obtain

$$
\begin{aligned}
\mathcal{M}(a^F)&=\sum_{r'}\sum_{r}\varpi_{r'}\frac{1}{K_{r-1}}\left\{E(h^F|G_r)-E(h^F|G_{r-1})\right\}I(r^{'}\geq r)\\
&=\sum_{r}\frac{1}{K_{r-1}}\left\{E(h^F|G_r)-E(h^F|G_{r-1})\right\}\sum_{r'}\varpi_{r'}I(r^{'}\geq r)\\
&=\sum_{r}\frac{1}{K_{r-1}}\left\{E(h^F|G_r)-E(h^F|G_{r-1})\right\}K_{r-1}\\
&=\sum_{r}\left\{E(h^F|G_r)-E(h^F|G_{r-1})\right\}\\
&=E(h^F|G_\infty)-E(h^F|G_0)=h^F.
\end{aligned}
$$

The second representation (11.28) will follow if we can show that

$$
1+\sum_{r\neq\infty}\frac{\lambda_r}{K_r}=\frac{1}{\varpi_\infty}. \qquad (11.33)
$$

Substituting (11.29) into (11.33), we obtain

$$
1+\sum_{r\neq\infty}\left(\frac{1}{K_r}-\frac{1}{K_{r-1}}\right)=1+\frac{1}{K_\ell}-\frac{1}{K_0}=\frac{1}{K_\ell}=\frac{1}{\varpi_\infty}. \quad \square
$$

$\mathcal{M}^{-1}$ with Right Censored Data

Because of the correspondence that was developed between monotone coarsening and right-censored data in Section 9.3, we immediately obtain the following result, which was first given by Robins and Rotnitzky (1992).

Lemma 11.2. In a survival analysis problem, full data are represented as $\{T,\bar{X}(T)\}$. Using the notation of Section 9.3, we obtain that

$$
\begin{aligned}
&\mathcal{M}^{-1}[h^F\{T,\bar{X}(T)\}]\\
&=\frac{h^F\{T,\bar{X}(T)\}}{K_T\{\bar{X}(T)\}}-\int_0^T E\left[h^F\{T,\bar{X}(T)\}\middle|T\geq r,\bar{X}(r)\right]\frac{\lambda_{\tilde{C}}\{r,\bar{X}(r)\}}{K_r\{\bar{X}(r)\}}dr,
\end{aligned}
$$

or

$$\mathcal{M}^{-1}[h^F\{T,\bar{X}(T)\}] = h^F\{T,\bar{X}(T)\}+$$
$$\int_0^T \left(h^F\{T,\bar{X}(T)\} - E\left[h^F\{T,\bar{X}(T)\}\middle| T \geq r, \bar{X}(r)\right]\right)\frac{\lambda_{\tilde{C}}\{r,\bar{X}(r)\}}{K_r\{\bar{X}(r)\}}dr,$$

where $\lambda_{\tilde{C}}\{r,\bar{X}(r)\}$ was defined in (9.30) and $K_r\{\bar{X}(r)\}$ was defined in (9.33).

11.2 Strategy for Obtaining Improved Estimators

The goal of this section is to outline a method for obtaining an AIPWCC estimator for β that is as efficient as possible while still remaining as robust to model misspecification as possible. Many of the calculations necessary to derive conditional expectations, projections, linear operators, etc., involve the coarsening probabilities as well as the marginal distribution of the full-data Z. In Chapter 10, we discussed methods for positing models for the coarsening probabilities and the distribution of Z in terms of ψ and ξ and finding estimators for these parameters.

We first consider finding a full-data estimating function $m(Z,\beta)$ such that $m(Z,\beta_0) = \varphi^{*F}(Z)$, where $\varphi^{*F}(Z) \in \Lambda^{F\perp}$ is an approximation to $B^F_{\text{eff}}(Z)$. Because in general it is too difficult to find such an estimating function explicitly, we may instead use successive approximations as given by (11.48) and (11.26). That is, we start with $m^{(0)}(Z,\beta)$ such that $m^{(0)}(Z,\beta_0) = S^F_{\text{eff}}(Z)$ and iteratively compute

$$\mathcal{D}^{(i)}(Z,\beta,\hat{\psi}_n,\hat{\xi}^*_n) = \Pi[(I-\mathcal{M})\mathcal{D}^{(i-1)}(Z,\beta,\hat{\psi}_n,\hat{\xi}^*_n)|\Lambda^F] + m^{(0)}(Z,\beta), \quad (11.34)$$

where we index $\mathcal{D}(\cdot)$ by ψ and ξ to make clear that we need these parameters when computing the projection $\Pi[(\cdot)|\Lambda^F]$ and the linear operator $\mathcal{M}(\cdot)$. After, say, j iterations, we compute

$$m(Z,\beta,\hat{\psi}_n,\hat{\xi}^*_n) = \Pi[\mathcal{M}\{\mathcal{D}^{(j)}(Z,\beta,\hat{\psi}_n,\hat{\xi}^*_n)\}|\Lambda^{F\perp}] \quad (11.35)$$

to serve as an approximation to $B^F_{\text{eff}}(Z)$.

Now that we've computed $m(Z,\beta,\hat{\psi}_n,\hat{\xi}^*_n)$, where $m(Z,\beta_0,\psi^*,\xi^*) \in \Lambda^{F\perp}$, we need to compute in accordance with (11.16), $\Pi\left[\frac{I(\mathcal{C}=\infty)m(Z,\beta)}{\varpi(\infty,Z)}\middle|\Lambda_2\right]$. If we have two levels of coarsening or monotone coarsening, then such a projection can be defined explicitly, and in Chapter 10 we discussed how such projections can be obtained using adaptive methods. With nonmonotone coarsened data, we may proceed as follows.

In Chapter 10, equation (10.95), we showed that $\Pi\left[\frac{I(\mathcal{C}=\infty)m(Z,\beta)}{\varpi(\infty,Z)}\middle|\Lambda_2\right]$ equals

$$- \frac{I(\mathcal{C}=\infty)}{\varpi(\infty, Z)} \left[\sum_{r \neq \infty} \varpi\{r, G_r(Z)\} E\{d^F(Z)|G_r(Z)\} \right]$$
$$+ \sum_{r \neq \infty} I(\mathcal{C}=r) E\{d^F(Z)|G_r(Z)\}, \tag{11.36}$$

where $d^F(Z) = \mathcal{M}^{-1}\{m(Z,\beta)\}$. Since $\mathcal{D}^{(j)}(Z, \beta, \hat{\psi}_n, \hat{\xi}_n^*)$ is an approximation to $\mathcal{M}^{-1}\{m(Z, \beta, \hat{\psi}_n, \hat{\xi}_n^*)\}$, we propose the following adaptive estimating equation:

$$\sum_{i=1}^{n} \Bigg(\frac{I(\mathcal{C}_i=\infty) m(Z_i, \beta, \hat{\psi}_n, \hat{\xi}_n^*)}{\varpi(\infty, Z_i, \hat{\psi}_n)}$$
$$- \frac{I(\mathcal{C}_i=\infty)}{\varpi(\infty, Z_i, \hat{\psi}_n)} \left[\sum_{r \neq \infty} \varpi\{r, G_r(Z_i), \hat{\psi}_n\} E\{\mathcal{D}^{(j)}(Z, \beta, \hat{\psi}_n, \hat{\xi}_n^*)|G_r(Z_i), \hat{\xi}_n^*\} \right]$$
$$+ \sum_{r \neq \infty} I(\mathcal{C}_i=r) E\{\mathcal{D}^{(j)}(Z, \beta, \hat{\psi}_n, \hat{\xi}_n^*)|G_r(Z_i), \hat{\xi}_n^*\} \Bigg) = 0. \tag{11.37}$$

The important thing to notice is that, by construction,

$$m(Z, \beta_0, \psi^*, \xi^*) \in \Lambda^{F\perp}$$

and

$$- \frac{I(\mathcal{C}=\infty)}{\varpi(\infty, Z, \psi_0)} \left[\sum_{r \neq \infty} \varpi\{r, G_r(Z), \psi_0\} E\{\mathcal{D}^{(j)}(Z, \beta, \psi_0, \xi^*)|G_r(Z), \xi^*\} \right]$$
$$+ \sum_{r \neq \infty} I(\mathcal{C}=r) E\{\mathcal{D}^{(j)}(Z, \beta, \psi_0, \xi^*)|G_r(Z), \xi^*\} \in \Lambda_2$$

as long as the model for the coarsening probabilities is correctly specified.

In some cases, it may be possible to derive $B_{\text{eff}}^F(Z)$ directly. We illustrate with an example.

Example: Restricted Moment Model with Monotone Coarsening

When the coarsening of the data is monotone, we showed in Theorem 11.2, equation (11.27), how to derive the inverse operator $\mathcal{M}^{-1}$ in closed form. Let us now examine how we would go about finding a locally efficient estimator for β with monotonically coarsened data. Specifically, we will consider the restricted moment model, as the semiparametric theory for such a model has been studied thoroughly throughout this book.

We remind the reader that, for the restricted moment model, the full data are given by

$$Z = (Y, X),$$

and the model assumes that $E(Y|X) = \mu(X, \beta)$, or, equivalently,

$$Y = \mu(X, \beta) + \varepsilon, \quad \text{where } E(\varepsilon|X) = 0.$$

For this semiparametric full-data model, we also derived a series of results regarding the geometry of the full-data influence functions and full-data estimating functions. Specifically, we showed in (4.48) that all elements of $\Lambda^{F\perp}$ are given by $A(X)\varepsilon$, where $A(X)$ is a conformable matrix of functions of X. We also showed in (4.44) that

$$\Pi[h^F(Z)|\Lambda^{F\perp}] = E\left\{h^F(Z)\varepsilon^T|X\right\} V^{-1}(X)\varepsilon, \tag{11.38}$$

where $V(X) = \text{var}(Y|X)$. The full-data efficient score was given by (4.53),

$$S^F_{\text{eff}}(Z) = D^T(X)V^{-1}(X)\varepsilon, \tag{11.39}$$

where

$$D(X) = \frac{\partial \mu(X, \beta)}{\partial \beta^T}.$$

Suppose the data are coarsened at random with a monotone coarsening pattern. An example with monotone missing longitudinal data was given in Example 1 in Section 9.2 and also studied further in Section 10.3, where double-robust estimators were proposed. The question is, how do we go about finding a locally efficient estimator for β with a sample of monotonically coarsened data $\{\mathcal{C}_i, G_{\mathcal{C}_i}(Z_i)\},\ i = 1, \ldots, n$?

We first develop a model for the coarsening probabilities in terms of a parameter ψ. Because in this example we are assuming that coarsening is monotone, it is convenient to develop models for the discrete hazards $\lambda_r\{G_r(Z), \psi\},\ r \neq \infty$ and obtain estimators for ψ by maximizing (8.12). We denote these estimators by $\hat{\psi}_n$.

We also posit a simpler, possibly incorrect model for the full data $Z = (Y, X)$, $Z \sim p^*_Z(z, \xi)$, and obtain an estimator for ξ, say $\hat{\xi}^*_n$, by maximizing the observed-data likelihood; see, for example, (10.58),

$$\prod_{i=1}^{n} p^*_{G_{r_i}(Z_i)}(g_{r_i}, \xi).$$

This model also gives us an estimate for $\text{var}(Y|X, \hat{\xi}^*_n) = V(X, \hat{\xi}^*_n)$.

If the data are coarsened at random, then by (11.16) and Theorem 11.1, the efficient observed-data score is given by

$$\frac{I(\mathcal{C} = \infty)B^F_{\text{eff}}(Z, \beta, \psi, \xi)}{\varpi(\infty, Z, \psi)} - \Pi\left[\frac{I(\mathcal{C} = \infty)B^F_{\text{eff}}(Z, \beta, \psi, \xi)}{\varpi(\infty, Z, \psi)}\middle|\Lambda_2\right], \tag{11.40}$$

where $B^F_{\text{eff}}(Z, \beta, \psi, \xi) \in \Lambda^{F\perp}$ must satisfy

$$\Pi\big[\mathcal{M}^{-1}\{B_{\text{eff}}^F(Z,\beta,\psi,\xi)\}|\Lambda^{F\perp}\big] = S_{\text{eff}}^F(Z,\beta,\xi).$$

For the restricted moment model, $B_{\text{eff}}^F(Z,\beta,\psi,\xi) = A(X,\beta,\psi,\xi)\varepsilon(\beta)$, where $\varepsilon(\beta) = Y - \mu(X,\beta)$, and the matrix $A(X,\beta,\psi,\xi)$ is obtained by solving the equation

$$\Pi\big[\mathcal{M}^{-1}\left\{A(X,\beta,\psi,\xi)\varepsilon(\beta)\right\}|\Lambda^{F\perp}\big] = S_{\text{eff}}^F(Z,\beta,\xi),$$

which by (11.38) and (11.39) is equal to

$$\begin{aligned} &E\big[\mathcal{M}^{-1}\left\{A(X,\beta,\psi,\xi)\varepsilon(\beta)\right\}\varepsilon^T(\beta)|X,\xi\big]V^{-1}(X,\xi)\varepsilon(\beta) \\ &= D^T(X,\beta)V^{-1}(X,\xi)\varepsilon(\beta), \end{aligned}$$

or, equivalently,

$$E\big[\mathcal{M}^{-1}\{A(X,\beta,\psi,\xi)\varepsilon(\beta)\}\varepsilon^T(\beta)|X,\xi\big] = D^T(X,\beta). \tag{11.41}$$

If, in addition, the coarsening is monotone, then using the results from Theorem 11.2, equation (11.27), we obtain

$$\begin{aligned} \mathcal{M}^{-1}\{A(X,\beta,\psi,\xi)\varepsilon(\beta)\} &= \frac{A(X,\beta,\psi,\xi)\varepsilon(\beta)}{\varpi(\infty,Y,X,\psi)} \\ &- \sum_{r\neq\infty}\frac{\lambda_r\{G_r(Y,X),\psi\}}{K_r\{G_r(Y,X),\psi\}}E\{A(X,\beta,\psi,\xi)\varepsilon(\beta)|G_r(Y,X),\xi\}. \end{aligned}$$

Combining this with (11.41), we obtain

$$\begin{aligned} &E\Bigg(\Bigg[\frac{A(X,\beta,\psi,\xi)\varepsilon(\beta)}{\varpi(\infty,Y,X,\psi)} \\ &-\sum_{r\neq\infty}\frac{\lambda_r\{G_r(Y,X),\psi\}}{K_r\{G_r(Y,X),\psi\}}E\left\{A(X,\beta,\psi,\xi)\varepsilon(\beta)|G_r(Y,X),\xi\right\}\Bigg]\times\varepsilon^T(\beta)\Big|X,\xi\Bigg) \\ &= D^T(X,\beta), \end{aligned}$$

or

$$\begin{aligned} &A(X,\beta,\psi,\xi)E\left\{\frac{\varepsilon(\beta)\varepsilon^T(\beta)}{\varpi(\infty,Y,X,\psi)}\Big|X,\xi\right\} \\ &-\sum_{r\neq\infty}E\left[\frac{\lambda_r\{G_r(Y,X),\psi\}}{K_r\{G_r(Y,X),\psi\}}E\left\{A(X,\beta,\psi,\xi)\varepsilon(\beta)\big|G_r(Y,X),\xi\right\}\varepsilon^T(\beta)\Big|X,\xi\right] \\ &= D^T(X,\beta). \end{aligned} \tag{11.42}$$

Remarks

(i) In general, equation (11.42) is difficult to solve. We do, however, get a simplification for problems where the covariates X are always observed.

For instance, this was the case in Example 1 of Section 9.2, which was further developed in Section 10.3, where the responses $Y = (Y_1, \ldots, Y_l)^T$ were longitudinal data intended to be measured at times $t_1 < \ldots < t_l$ but were missing for some subjects in the study in a monotone fashion due to patient dropout. For this example, the covariate X (treatment assignment) was always observed but some of the longitudinal measurements that made up Y were missing. The coarsening was described as $G_r(Z) = (X, Y^r)$, where $Y^r = (Y_1, \ldots, Y_r)^T,\ r = 1, \ldots, l-1$. Equation (11.42) can now be written as

$$\begin{aligned} &A(X,\beta,\psi,\xi)E\left\{\frac{\varepsilon(\beta)\varepsilon^T(\beta)}{\varpi(\infty,Y,X,\psi)}\middle|X,\xi\right\} \\ -&A(X,\beta,\psi,\xi)\sum_{r\neq\infty}E\left[\frac{\lambda_r(X,Y^r,\psi)}{K_r(X,Y^r,\psi)}E\left\{\varepsilon(\beta)\middle|X,Y^r,\xi\right\}\varepsilon^T(\beta)\middle|X,\xi\right] \\ &= D^T(X,\beta). \end{aligned}$$

Therefore, the solution is given by

$$A(X,\beta,\psi,\xi) = D^T(X,\beta)\tilde{V}^{-1}(X,\beta,\psi,\xi),$$

where

$$\begin{aligned} \tilde{V}(X,\beta,\psi,\xi) = &\left(E\left\{\frac{\varepsilon(\beta)\varepsilon^T(\beta)}{\varpi(\infty,Y,X,\psi)}\middle|X,\xi\right\}\right. \\ &\left.-\sum_{r\neq\infty}E\left[\frac{\lambda_r(X,Y^r,\psi)}{K_r(X,Y^r,\psi)}E\left\{\varepsilon(\beta)\middle|X,Y^r,\xi\right\}\varepsilon^T(\beta)\middle|X,\xi\right]\right). \end{aligned}$$

(ii) Except for special cases, such as the example above, the equation for solving $A(X)$ in (11.42) is generally a complicated integral equation. Approximate methods for solving such integral equations are given in Kress (1989). However, these computations may be so difficult as not to be feasible in practice.

(iii) In Chapter 12, we will give some approximate methods for obtaining improved estimators that although not locally efficient do have increased efficiency and are easier to implement. □

Suppose we were able to overcome these numerical difficulties and obtain an approximate solution for $A(X,\beta,\psi,\xi)$. Denoting this solution by $A_{\text{imp}}(X,\beta,\hat{\psi}_n,\hat{\xi}_n^*)$ and going back to (11.40), we approximate the efficient score by

$$\begin{aligned} &\frac{I(\mathcal{C}=\infty)A_{\text{imp}}(X,\beta,\hat{\psi}_n,\hat{\xi}_n^*)\{Y-\mu(X,\beta)\}}{\pi(\infty,Y,X,\hat{\psi}_n)} \\ &-\Pi\left[\frac{I(\mathcal{C}=\infty)A_{\text{imp}}(X,\beta,\hat{\psi}_n,\hat{\xi}_n^*)\{Y-\mu(X,\beta)\}}{\pi(\infty,Y,X,\hat{\psi}_n)}\middle|\Lambda_2\right]. \end{aligned} \tag{11.43}$$

Because the coarsening is monotone, we can use (10.55) to estimate the projection onto Λ_2 by $-L_2\{\mathcal{C}, G_{\mathcal{C}}(Y,X), \beta, \hat{\psi}_n, \hat{\xi}_n^*\}$, where

$$L_2\{\mathcal{C}, G_{\mathcal{C}}(Y,X), \beta, \hat{\psi}_n, \hat{\xi}_n^*\}$$
$$\sum_{r \neq \infty} \left(\left[\frac{I(\mathcal{C}=r) - \lambda_r\{G_r(Y,X), \hat{\psi}_n\} I(\mathcal{C} \geq r)}{K_r\{G_r(Y,X), \hat{\psi}_n\}} \right] \times \right.$$
$$\left. E\left[A_{\text{imp}}(X, \beta, \hat{\psi}_n, \hat{\xi}_n^*)\{Y - \mu(X,\beta)\} \middle| G_r(Y,X), \hat{\xi}_n \right] \right). \tag{11.44}$$

Finally, the estimator for β is given as the solution to the estimating equation

$$\sum_{i=1}^{n} \left[\frac{I(\mathcal{C}_i = \infty) A_{\text{imp}}(X_i, \beta, \hat{\psi}_n, \hat{\xi}_n^*)\{Y_i - \mu(X_i,\beta)\}}{\varpi(\infty, Y_i, X_i, \hat{\psi}_n)} \right.$$
$$\left. + L_2\{\mathcal{C}_i, G_{\mathcal{C}_i}(Y_i, X_i), \beta, \hat{\psi}_n, \hat{\xi}_n^*\} \right] = 0. \tag{11.45}$$

Some Brief Remarks Regarding Robustness

The estimator for β given by the solution to equation (11.45) used

$$A_{\text{imp}}(X_i, \beta, \hat{\psi}_n, \hat{\xi}_n^*)\{Y_i - \mu(X_i,\beta)\} \tag{11.46}$$

to represent the full-data estimating function $m(Z_i, \beta)$. Strictly speaking, (11.46) is not a full-data estimating function, as it involves the parameter estimators $\hat{\psi}_n$ and $\hat{\xi}_n^*$. However, as we discussed in Remark 1 of Chapter 9, the solution to the estimating equation (11.45) had we substituted

$$A_{\text{imp}}(X_i, \beta, \psi^*, \xi^*)\{Y_i - \mu(X_i,\beta)\}$$

as $m(Z_i, \beta)$, where ψ^* and ξ^* are the limits (in probability) of $\hat{\psi}_n$ and $\hat{\xi}_n^*$, would result in an asymptotically equivalent estimator for β. What is important to note here is that

$$A_{\text{imp}}(X, \beta_0, \psi^*, \xi^*)\{Y - \mu(X,\beta_0)\} \in \Lambda^{F\perp} \tag{11.47}$$

regardless of what the converging values ψ^* and ξ^* are, and therefore the solution to (11.45) is an example of an AIPWCC estimator for β.

Because of (11.47), the estimator given as the solution to (11.45) with $L_2(\cdot)$ computed by (11.44) is an example of an improved estimator as described in Section 10.4. As such, this estimator is a double robust estimator in the sense that it will be consistent and asymptotically normal if either the coarsening model or the model for the posited marginal distribution of Z is correctly specified. This double-robustness property holds regardless of whether

$$A_{\text{imp}}(X, \beta_0, \psi^*, \xi^*)\{Y - \mu(X, \beta_0)\} = B^F_{\text{eff}}(Z)$$

or not.

Finally, if both models are correctly specified, and if

$$A_{\text{imp}}(X, \beta_0, \psi_0, \xi_0)\{Y - \mu(X, \beta_0)\} = B^F_{\text{eff}}(Z),$$

then the resulting estimator will be semiparametric efficient. Thus, this methodology, assuming the numerical complexities could be overcome, would lead to locally efficient observed-data estimators for β.

11.3 Concluding Thoughts

In the last two chapters, we have outlined methods for obtaining increasingly efficient estimators while trying to keep them as robust as possible. The key was always to use AIPWCC estimators. By no means do we want to give the impression that these methods are easily implemented. Deriving parameter estimates for ξ in a simpler posited parametric model for the marginal distribution of Z, which is used to obtain adaptive estimators, may require maximizing a coarsened-data likelihood. Such maximization algorithms may be complicated and may need specialized software. Even if these estimators are obtained, finding projections, deriving linear operators (such as $\mathcal{M}(\cdot)$ or $\mathcal{M}^{-1}(\cdot)$) may require complicated integrals. Therefore, although the theory for improved adaptive estimation has been laid out, the actual implementation needs to be considered on a case-by-case basis.

The use of this inverse weighted methodology can be thought of as a balance between simplicity of implementation and relative efficiency. The simplest estimator is the inverse probability weighted complete-case estimator based on some prespecified full-data estimating function. Since this estimator only uses complete cases, it may be throwing away a great deal of information from data that are coarsened. Depending on how much of the data are coarsened, this estimator may be inadequate. Also, the consistency of the simpler IPWCC estimator depends on correctly modeling the coarsening probabilities.

Improving the performance of the estimator by augmentation while using the same full-data estimating function is the next step. To implement these methods, one needs to develop simpler and possibly incorrect models for the marginal distribution of Z to be used as part of an adaptive approach. This can improve the efficiency considerably but at the cost of increased computations and model building. The attempt to gain efficiency also gives you the extra protection of double robustness in that the resulting AIPWCC estimator will be consistent if either the model for the coarsening probabilities or the posited model for the marginal distribution of Z is correctly specified.

Finally, the attempt to adaptively obtain the locally efficient estimator is the most complex numerically. Here we actually attempt to find the optimal

full-data estimating function, $B^F_{\text{eff}}(Z) \in \Lambda^{F\perp}$, as well as the optimal augmentation. It is not clear whether the efficiency gains of such an estimator would make such a complicated procedure attractive for practical use.

Because of the complexity of these methods, we offer in the next chapter some simpler methods for gaining efficiency that are easier to implement. These methods will not generally result in locally efficient estimators, but they are, however, more feasible.

11.4 Recap and Review of Notation

- The observed-data efficient score, which can be used to derive adaptive AIPWCC estimators that are locally efficient, is given by

$$S_{\text{eff}}\{\mathcal{C}, G_{\mathcal{C}}(Z)\} = \mathcal{J}\{B^F_{\text{eff}}(Z)\},$$

where

$$\mathcal{J}\{h^F(Z)\} = \frac{I(\mathcal{C}=\infty)h^F(Z)}{\varpi(\infty, Z)} - \Pi\left[\frac{I(\mathcal{C}=\infty)h^F(Z)}{\varpi(\infty, Z)}\middle|\Lambda_2\right],$$

$B^F_{\text{eff}}(Z)$ is the unique element in $\Lambda^{F\perp}$ that satisfies

$$\Pi[\mathcal{M}^{-1}\{B^F_{\text{eff}}(Z)\}|\Lambda^{F\perp}] = S^F_{\text{eff}}(Z),$$

$\mathcal{M}^{-1}(\cdot)$ is the inverse of the linear operator

$$\begin{aligned}\mathcal{M}\{h^F(Z)\} &= E[E\{h^F(Z)|\mathcal{C}, G_{\mathcal{C}}(Z)\}|Z] \\ &= \sum_r \varpi\{r, G_r(Z)\}E\{h^F(Z)|G_r(Z)\},\end{aligned}$$

and $S^F_{\text{eff}}(Z)$ is the full-data efficient score vector.

- We can derive $B^F_{\text{eff}}(Z)$ by solving the equation

$$(I-\mathcal{Q})\mathcal{M}^{-1}\{B^F_{\text{eff}}(Z)\} = S^F_{\text{eff}}(Z),$$

where $(I-\mathcal{Q})(\cdot)$ is a linear operator, with $\mathcal{Q}(\cdot)$ defined as

$$\mathcal{Q}\{h^F(Z)\} = \Pi[(I-\mathcal{M})\{h^F(Z)\}|\Lambda^F].$$

$(I-\mathcal{Q})(\cdot)$ is a contraction mapping and hence has a unique inverse. Therefore, $B^F_{\text{eff}}(Z) = \mathcal{M}\{\mathcal{D}^F_{\text{eff}}(Z)\}$, where $\mathcal{D}^F_{\text{eff}}(Z) = (I-\mathcal{Q})^{-1}\{S^F_{\text{eff}}(Z)\}$. The solution $\mathcal{D}^F_{\text{eff}}(Z)$ can be obtained by successive approximation,

$$\mathcal{D}^{(i+1)}(Z) = \Pi\big[(I-\mathcal{M})\mathcal{D}^{(i)}(Z)\big|\Lambda^F\big] + S^F_{\text{eff}}(Z), \tag{11.48}$$

and

$$\mathcal{D}^{(i)}(Z) \xrightarrow{i\to\infty} \mathcal{D}^F_{\text{eff}}(Z).$$

If we define

$$B^{(i)}(Z) = \Pi\big[\mathcal{M}\{\mathcal{D}^{(i)}(Z)\}\big|\Lambda^{F\perp}\big],$$

where, by construction, $B^{(i)}(Z) \in \Lambda^{F\perp}$, then

$$B^{(i)}(Z) \xrightarrow{i\to\infty} B^F_{\text{eff}}(Z).$$

11.5 Exercises for Chapter 11

1. Recall that, with two levels of missingness, the data $Z = (Z_1^T, Z_2^T)^T$, where Z_1 is always observed and Z_2 may be missing. We denote by R the complete-case indicator and assume $P(R = 1|Z) = P(R = 1|Z_1) = \pi(Z_1)$ (i.e., MAR). In Theorem 11.2, we derived the inverse operator $\mathcal{M}^{-1}$ when coarsening is monotone. You should derive an explicit expression for $\mathcal{M}^{-1}$ when there are two levels of missingness.
   (Note: Two levels of missingness can be viewed as a special case of monotone coarsening.)
2. In Section 11.2, we outlined the steps necessary to obtain a locally efficient estimator for β in a restricted moment model,

   $$E(Y|X) = \mu(X, \beta),$$

   when the data are monotonically coarsened. Similarly, outline the steps necessary to obtain a locally efficient estimator for β if there are two levels of missingness.

12

Approximate Methods for Gaining Efficiency

12.1 Restricted Class of AIPWCC Estimators

In Chapters 10 and 11, we described various methods to increase the efficiency of AIPWCC estimators. Although feasible in some situations, these methods are often computationally very challenging. For example, to use these methods, one needs to posit simpler models for the marginal distribution of the full-data Z in terms of a parameter ξ (i.e., $p^*_Z(z,\xi)$) and then estimate the parameter. This in itself can be a challenging numerical problem. But, even if estimators for ξ can be derived, it is necessary to compute a series of conditional expectations that might involve complicated integrals that are difficult to compute numerically. Moreover, as discussed in Chapter 11, an attempt to derive locally efficient estimators could result in complicated integral equations that are difficult to solve. Therefore, in this chapter, we explore other methods that are numerically easier to implement but will result in gains in efficiency.

We remind the reader that all semiparametric observed-data RAL estimators for β, when the coarsening is CAR, are asymptotically equivalent to an AIPWCC estimator, which is the solution to

$$\sum_{i=1}^{n} \left\{ \frac{I(\mathcal{C}_i = \infty) m(Z_i, \beta)}{\varpi(\infty, Z_i, \hat{\psi}_n)} + L_2\{\mathcal{C}_i, G_{\mathcal{C}_i}(Z_i), \hat{\psi}_n\} \right\} = 0, \tag{12.1}$$

where the estimating function $m(Z,\beta)$ is chosen so that $m(Z,\beta_0) = \varphi^{*F}(Z) \in \Lambda^{F\perp} \subset \mathcal{H}^F$ and $L_2(\cdot) \in \Lambda_2 \subset \mathcal{H}$, and $\hat{\psi}_n$ denotes the MLE for ψ obtained by maximizing

$$\prod_{i=1}^{n} \varpi\{\mathcal{C}_i, G_{\mathcal{C}_i}(Z_i), \psi\}.$$

Rather than searching for the optimal AIPWCC estimator, which involves finding the optimal $L_{2\text{eff}}(\cdot) \in \Lambda_2$ and the optimal $B^F_{\text{eff}}(Z) \in \Lambda^{F\perp}$, we instead restrict the search to linear subspaces of $\Lambda_2 \subset \mathcal{H}$ and $\Lambda^{F\perp} \subset \mathcal{H}^F$. That is,

we will only consider AIPWCC estimators that are the solution to (12.1) for $m(Z,\beta_0) = \varphi^{*F}(Z) \in \mathcal{G}^F$ and $L_2(\cdot) \in \mathcal{G}_2$, where $\mathcal{G}^F$ is a q-replicating linear subspace of $\Lambda^{F\perp}$ and $\mathcal{G}_2$ is a q-replicating linear subspace of Λ_2, where a q-replicating linear space is defined by Definition 6 of Chapter 3.

Remark 1. We remind the reader that a full-data Hilbert space $\mathcal{H}^F$ consists of mean-zero finite-variance q-dimensional functions of Z and the observed-data Hilbert space $\mathcal{H}$ consists of mean-zero finite-variance q-dimensional functions of $\{\mathcal{C}, G_{\mathcal{C}}(Z)\}$. In Definition 6 of Chapter 3, we noted that a q-replicating linear subspace can be written as $\{\mathcal{U}^{(1)}\}^q$, where $\mathcal{U}^{(1)}$ is a linear subspace contained in the Hilbert space $\mathcal{H}^{(1)}$ of one-dimensional mean-zero finite-variance functions and where $h(\cdot) = \{h_1(\cdot), \ldots, h_q(\cdot)\}^T$ is defined to be an element of $\{\mathcal{U}^{(1)}\}^q$ if and only if each element $h_j(\cdot) \in \mathcal{U}^{(1)}$, $j = 1, \ldots, q$. Linear subspaces, such as $\Lambda^{F\perp} \subset \mathcal{H}^F$ and $\Lambda_2 \subset \mathcal{H}$, are examples of q-replicating spaces. The importance of defining q-replicating linear spaces is given by Theorem 3.3, which allows a generalization of the Pythagorean theorem to q dimensions. One of the consequences of Theorem 3.3 is that if an element h is orthogonal to a q-replicating linear space, say $\{\mathcal{U}^{(1)}\}^q$, then not only is $E(h^T u) = 0$ for all $u \in \{\mathcal{U}^{(1)}\}^q$ (the definition of orthogonality) but also

$$E(hu^T) = 0^{q\times q} \text{ for all } u \in \{\mathcal{U}^{(1)}\}^q. \square \tag{12.2}$$

We will consider two specific classes of restricted estimators.

1. For the first class of restricted estimators, we will take both $\mathcal{G}^F \subset \Lambda^{F\perp}$ and $\mathcal{G}_2 \subset \Lambda_2$ to be finite-dimensional linear subspaces.
2. For the second class of restricted estimators, we will take $\mathcal{G}^F \subset \Lambda^{F\perp}$ to be a finite-dimensional linear subspace contained in the orthogonal complement of the full-data nuisance tangent space but will let $\mathcal{G}_2 = \Lambda_2$ be the entire augmentation space.

We note that a finite-dimensional linear subspace $\mathcal{G}^F$ of $\Lambda^{F\perp}$ can be defined by choosing a t_1-dimensional function of Z, say $J^F(Z) = \{J_1^F(Z), \ldots, J_{t_1}^F(Z)\}^T$, where $J_j^F(Z) \in \Lambda^{F\perp^{(1)}}$, $j = 1, \ldots, t_1$, and letting the space spanned by $J^F(Z)$ be

$$\mathcal{G}^F = \left\{ A^{q\times t_1} J^F(Z) \text{ for all constant matrices } A^{q\times t_1} \right\}.$$

Similarly, for (class 1) restricted estimators, a finite-dimensional linear subspace $\mathcal{G}_2$ of Λ_2 can be defined by choosing a t_2-dimensional function of $\{\mathcal{C}, G_{\mathcal{C}}(Z)\}$, say $J_2\{\mathcal{C}, G_{\mathcal{C}}(Z)\} = [J_{21}\{\mathcal{C}, G_{\mathcal{C}}(Z)\}, \ldots, J_{2t_2}\{\mathcal{C}, G_{\mathcal{C}}(Z)\}]^T$, where $J_{2j}\{\mathcal{C}, G_{\mathcal{C}}(Z)\} \in \Lambda_2^{(1)}$, $j = 1, \ldots, t_2$, and letting the space spanned by $J_2\{\mathcal{C}, G_{\mathcal{C}}(Z)\}$ be

$$\mathcal{G}_2 = \left\{ A^{q\times t_2} J_2\{\mathcal{C}, G_{\mathcal{C}}(Z)\} \text{ for all constant matrices } A^{q\times t_2} \right\}.$$

We will always assume that the t_1 elements of $J^F(Z)$ and the t_2 elements of $J_2\{\mathcal{C}, G_{\mathcal{C}}(Z)\}$ are linearly independent; that is, $c^T J^F(Z) = 0$ implies that $c = 0$, where c is a t_1 vector of constants (similarly for $J_2\{\mathcal{C}, G_{\mathcal{C}}(Z)\}$). By construction, the finite-dimensional linear spaces defined above are q-replicating linear spaces.

Therefore, we will restrict attention to estimators for β that are the solution to the estimating equation

$$\sum_{i=1}^{n} \left[\frac{I(\mathcal{C}_i = \infty) A^{F^{q\times t_1}} m^*(Z_i, \beta)}{\varpi(\infty, Z_i, \hat{\psi}_n)} + L_2\{\mathcal{C}_i, G_{\mathcal{C}_i}(Z_i), \hat{\psi}_n\} \right] = 0, \tag{12.3}$$

where $m^*(Z, \beta)$ is a t_1-dimensional estimating function such that $m^*(Z, \beta_0) = J^F(Z)$ and $A^{F^{q\times t_1}}$ is an arbitrary $q \times t_1$ constant matrix, and $L_2(\cdot) \in \mathcal{G}_2$, either $L_2\{\mathcal{C}, G_{\mathcal{C}}(Z)\} = A_2^{q\times t_2} J_2\{\mathcal{C}, G_{\mathcal{C}}(Z)\}$, where $A_2^{q\times t_2}$ is an arbitrary $q \times t_2$ matrix (class 1), or $L_2(\cdot) \in \Lambda_2$ (class 2).

Remark 2. The observed-data estimating function (12.1) uses the full-data estimating function $m(Z, \beta)$ to build estimators. Because the parameter β is q-dimensional, the estimating function $m(Z, \beta)$ is also q-dimensional, and, at the minimum, the elements of $m(Z, \beta)$ must be linearly independent. When we consider restricted estimators that solve (12.3), the full-data estimating function $m(Z, \beta)$ is now equal to $A^{F^{q\times t_1}} m^*(Z, \beta)$, where $m^*(Z, \beta)$ is t_1-dimensional. To ensure that the elements of $m(Z, \beta)$, constructed in this way, are linearly independent, the dimension t_1 must be greater than or equal to q. Moreover, since estimating equations, such as (12.3), can be defined up to a proportional constant matrix (that is, multiplying the left-hand side of (12.3) by a nonsingular $q \times q$ matrix will not affect the resulting estimator), we must choose the dimension t_1 to be strictly greater than q so that the strategy of choosing from this class of restricted estimators has an effect on the resulting estimator. Therefore, from here on, we will always assume that $m^*(Z, \beta)$, chosen so that $m^*(Z, \beta_0) = J^F(Z)$, is made up of t_1 linearly independent elements with $t_1 > q$. □

Let us define the q-replicating linear space $\Xi \subset \mathcal{H}$ to be

$$\Xi = \left\{ \frac{I(\mathcal{C} = \infty)\mathcal{G}^F}{\varpi(\infty, Z)} \oplus \mathcal{G}_2 \right\}; \tag{12.4}$$

that is, Ξ consists of the elements

$$\frac{I(\mathcal{C} = \infty) A^F m^*(Z, \beta_0)}{\varpi(\infty, Z)} + L_2\{\mathcal{C}, G_{\mathcal{C}}(Z)\} \tag{12.5}$$

in $\mathcal{H}$ for any constant matrix A^F and $L_2 \in \mathcal{G}_2$. We remind the reader that the orthogonal complement of the nuisance tangent space associated with the full-data nuisance parameter η is denoted by $\Lambda_\eta^\perp$, which, by Theorem 7.2, is equal to

$$\Lambda_\eta^\perp = \frac{I(\mathcal{C}=\infty)\Lambda^{F\perp}}{\varpi(\infty, Z)} \oplus \Lambda_2,$$

and hence

$$\Xi \subset \Lambda_\eta^\perp. \tag{12.6}$$

The elements in the linear subspace Ξ are associated with estimating functions that will lead to the restricted class of estimating equations (12.3) that we are considering. However, in Theorem 9.1, we showed that substituting the MLE $\hat{\psi}_n$ for the parameter ψ in an AIPWCC estimating equation resulted in an influence function that subtracts off a projection onto the space Λ_ψ, where Λ_ψ denotes the linear subspace spanned by the score vector

$$S_\psi\{\mathcal{C}, G_\mathcal{C}(Z)\} = \frac{\partial \log \varpi\{\mathcal{C}, G_\mathcal{C}(Z), \psi_0\}}{\partial \psi}.$$

We remind the reader that when we introduce a coarsening model with parameter ψ, the nuisance tangent space is given by $\Lambda = \Lambda_\eta \oplus \Lambda_\psi$. Since estimating functions need to be associated with elements in $\Lambda^\perp$, we should only consider elements of Ξ that are also orthogonal to Λ_ψ; i.e., $\Pi[\Xi|\Lambda_\psi^\perp]$. Consequently, it will prove desirable that the space $\mathcal{G}_2 \subset \Lambda_2$ also contain Λ_ψ. For the restricted (class 2) estimators where $\mathcal{G}_2 = \Lambda_2$, this is automatically true, however, for the restricted (class 1) estimators, we will always include the elements $S_\psi\{\mathcal{C}, G_\mathcal{C}(Z)\}$ that span Λ_ψ as part of the vector $J_2\{\mathcal{C}, G_\mathcal{C}(Z)\}$ that spans $\mathcal{G}_2$ to ensure that $\Lambda_\psi \subset \mathcal{G}_2$.

Because the variance of an RAL estimator for β is the variance of its influence function, when we consider finding the optimal estimator within the restricted class of estimators (12.3), then we are looking for the estimator whose influence function has the smallest variance matrix. Recall that an influence function in addition to being orthogonal to the nuisance tangent space must also satisfy the property that

$$\begin{aligned} &E[\varphi\{\mathcal{C}, G_\mathcal{C}(Z)\}S_\beta^T\{\mathcal{C}, G_\mathcal{C}(Z)\}] \\ &= E[S_\beta\{\mathcal{C}, G_\mathcal{C}(Z)\}\varphi^T\{\mathcal{C}, G_\mathcal{C}(Z)\}] = I^{q\times q}, \end{aligned} \tag{12.7}$$

where $S_\beta\{\mathcal{C}, G_\mathcal{C}(Z)\}$ is the observed-data score vector with respect to β, which also equals the conditional expectation of the full-data score vector with respect to β given the observed data (i.e., $S_\beta\{\mathcal{C}, G_\mathcal{C}(Z)\} = E\{S_\beta^F(Z)|\mathcal{C}, G_\mathcal{C}(Z)\}$), and $\varphi\{\mathcal{C}, G_\mathcal{C}(Z)\}$ denotes an influence function. One can always normalize any element $\varphi^*\{\mathcal{C}, G_\mathcal{C}(Z)\} \in \Xi$ to ensure that it satisfies (12.7) by choosing

$$\varphi\{\mathcal{C}, G_\mathcal{C}(Z)\} = \Big(E[\varphi^*\{\mathcal{C}, G_\mathcal{C}(Z)\}S_\beta^T\{\mathcal{C}, G_\mathcal{C}(Z)\}]\Big)^{-1} \varphi^*\{\mathcal{C}, G_\mathcal{C}(Z)\}. \tag{12.8}$$

Also, because Ξ is a q-replicating linear subspace, the element $\varphi\{\mathcal{C}, G_\mathcal{C}(Z)\} \in \Xi$.

Let us define the subset of elements $\varphi\{\mathcal{C}, G_\mathcal{C}(Z)\} \in \Xi$ that satisfy the property (12.7) or, equivalently, the subset of elements defined by (12.8), as

$IF(\Xi)$. In Theorem 12.1 below, we derive the optimal influence function within $IF(\Xi)$; i.e., the element within $IF(\Xi)$ that has the smallest variance matrix.

Remark 3. If the coarsening probabilities are modeled using a parameter ψ that needs to be estimated, then influence functions of RAL estimators for β must be orthogonal to Λ_ψ. However, as we will show shortly, as long as $\Lambda_\psi \subset \mathcal{G}_2$, then the optimal influence function within $IF(\Xi)$ will also be orthogonal to Λ_ψ, as is desired. □

Theorem 12.1. Among the elements $\varphi\{\mathcal{C}, G_{\mathcal{C}}(Z)\} \in IF(\Xi)$, the one with the smallest variance matrix is given by the normalized version of the projection of $S_\beta\{\mathcal{C}, G_{\mathcal{C}}(Z)\}$ onto Ξ. Specifically, if we define $\varphi^*_{\text{opt}}\{\mathcal{C}, G_{\mathcal{C}}(Z)\} = \Pi[S_\beta\{\mathcal{C}, G_{\mathcal{C}}(Z)\}|\Xi]$, then the element in $IF(\Xi)$ with the smallest variance matrix is given by

$$\varphi_{\text{opt}}\{\mathcal{C}, G_{\mathcal{C}}(Z)\} = \left(E[\varphi^*_{\text{opt}}\{\mathcal{C}, G_{\mathcal{C}}(Z)\}S^T_\beta\{\mathcal{C}, G_{\mathcal{C}}(Z)\}]\right)^{-1} \varphi^*_{\text{opt}}\{\mathcal{C}, G_{\mathcal{C}}(Z)\}. \tag{12.9}$$

Proof. Because Ξ is a closed linear subspace, then, by the projection theorem for Hilbert spaces, there exists a unique projection $\varphi^*_{\text{opt}}\{\mathcal{C}, G_{\mathcal{C}}(Z)\} = \Pi[S_\beta\{\mathcal{C}, G_{\mathcal{C}}(Z)\}|\Xi]$ such that the residual $\left[S_\beta\{\mathcal{C}, G_{\mathcal{C}}(Z)\} - \varphi^*_{\text{opt}}\{\mathcal{C}, G_{\mathcal{C}}(Z)\}\right]$ is orthogonal to every element in Ξ. Consider any element $\varphi\{\mathcal{C}, G_{\mathcal{C}}(Z)\} \in IF(\Xi)$. Because $IF(\Xi) \subset \Xi$ and, since both $\varphi\{\mathcal{C}, G_{\mathcal{C}}(Z)\}$ and $\varphi_{\text{opt}}\{\mathcal{C}, G_{\mathcal{C}}(Z)\}$ belong to $IF(\Xi)$, this implies that $\varphi(\cdot) - \varphi_{\text{opt}}(\cdot) \in \Xi$. Also, because Ξ is a q-replicating linear space (see Remark 1), by Theorem 3.3 we obtain that

$$E[\{S_\beta(\cdot) - \varphi^*_{\text{opt}}(\cdot)\}\{\varphi(\cdot) - \varphi_{\text{opt}}(\cdot)\}^T] = 0^{q\times q}.$$

Because elements of $IF(\Xi)$ must satisfy (12.7), this implies that

$$E\{S_\beta(\cdot)\varphi^T(\cdot)\} = E\{S_\beta(\cdot)\varphi^T_{\text{opt}}(\cdot)\} = I^{q\times q},$$

and hence

$$E[\varphi^*_{\text{opt}}(\cdot)\{\varphi(\cdot) - \varphi_{\text{opt}}(\cdot)\}^T] = 0.$$

Premultiplying by the constant matrix $\left[E\{\varphi^*_{\text{opt}}(\cdot)S^T_\beta(\cdot)\}\right]^{-1}$, we obtain

$$E[\varphi_{\text{opt}}(\cdot)\{\varphi(\cdot) - \varphi_{\text{opt}}(\cdot)\}^T] = 0.$$

Consequently,

$$\begin{aligned} E\{\varphi(\cdot)\varphi^T(\cdot)\} &= E[\{\varphi_{\text{opt}}(\cdot) + \varphi(\cdot) - \varphi_{\text{opt}}(\cdot)\}\{\varphi_{\text{opt}}(\cdot) + \varphi(\cdot) - \varphi_{\text{opt}}(\cdot)\}^T] \\ &= E\{\varphi_{\text{opt}}(\cdot)\varphi^T_{\text{opt}}(\cdot)\} + E[\{\varphi(\cdot) - \varphi_{\text{opt}}(\cdot)\}\{\varphi(\cdot) - \varphi_{\text{opt}}(\cdot)\}^T] + 0. \end{aligned}$$

Since $E[\{\varphi(\cdot) - \varphi_{\text{opt}}(\cdot)\}\{\varphi(\cdot) - \varphi_{\text{opt}}(\cdot)\}^T]$ is a nonnegative definite matrix, this implies that

$$E\{\varphi(\cdot)\varphi^T(\cdot)\} \geq E\{\varphi_{\text{opt}}(\cdot)\varphi_{\text{opt}}^T(\cdot)\},$$

giving us the desired result. □

Corollary 1. The optimal element $\varphi_{\text{opt}}\{\mathcal{C}, G_{\mathcal{C}}(Z)\} \in IF(\Xi)$, derived in Theorem 12.1, is orthogonal to Λ_ψ and is an element of the space of observed-data influence functions (IF).

Proof. By construction, $S_\beta\{\mathcal{C}, G_{\mathcal{C}}(Z)\} - \varphi^*_{\text{opt}}\{\mathcal{C}, G_{\mathcal{C}}(Z)\}$ is orthogonal to Ξ, where Ξ is defined by (12.4). Hence, it must be orthogonal to $\mathcal{G}_2$. Also, by construction, $\Lambda_\psi \subset \mathcal{G}_2$. This implies that $S_\beta\{\mathcal{C}, G_{\mathcal{C}}(Z)\} - \varphi^*_{\text{opt}}\{\mathcal{C}, G_{\mathcal{C}}(Z)\}$ is orthogonal to Λ_ψ. In other words,

$$E\{(S_\beta - \varphi^*_{\text{opt}})^T h\} = 0 \text{ for all } h \in \Lambda_\psi.$$

But since $S_\beta\{\mathcal{C}, G_{\mathcal{C}}(Z)\}$ is orthogonal to Λ_ψ (see equation (11.7)), this implies that $\varphi^*_{\text{opt}}\{\mathcal{C}, G_{\mathcal{C}}(Z)\}$ must be orthogonal to Λ_ψ. Therefore $\varphi_{\text{opt}}\{\mathcal{C}, G_{\mathcal{C}}(Z)\}$, defined by (12.9), is also orthogonal to Λ_ψ. Hence,

$$\varphi_{\text{opt}}\{\mathcal{C}, G_{\mathcal{C}}(Z)\} \in \Pi[\Xi|\Lambda_\psi^\perp] \subset \Pi[\Lambda_\eta^\perp|\Lambda_\psi^\perp] = \Lambda^\perp,$$

where the last two relationships follow from (12.6) and (8.16), respectively. Therefore, $\varphi_{\text{opt}}\{\mathcal{C}, G_{\mathcal{C}}(Z)\}$ is an element orthogonal to the observed-data nuisance tangent space and by construction (see (12.9)) satisfies (12.7). Hence, $\varphi_{\text{opt}}\{\mathcal{C}, G_{\mathcal{C}}(Z)\}$ is an element of the space of observed-data influence functions (IF). □

12.2 Optimal Restricted (Class 1) Estimators

We first consider how we can use the result from Theorem 12.1 to obtain improved estimators by finding the optimal estimator within the restricted class of estimators (12.3) where $L_2\{\mathcal{C}, G_{\mathcal{C}}(Z)\} = A_2^{q\times t_2} J_2\{\mathcal{C}, G_{\mathcal{C}}(Z)\}$; i.e., the so-called (class 1) estimators. For this class of estimators, we will assume that the model for the coarsening probabilities has been correctly specified.

Examining the elements in (12.5), we note that $\Xi \subset \mathcal{H}$ is a finite-dimensional linear subspace that is spanned by the $(t_1 + t_2)$ vector of functions of the observed data, namely

$$\left\{ \frac{I(\mathcal{C} = \infty) m^{*^T}(Z, \beta_0)}{\varpi(\infty, Z)}, J_2^T(\cdot) \right\}^T.$$

Therefore, finding the projection onto this linear subspace of $\mathcal{H}$ is exactly the same as Example 2 of Chapter 2, which was solved by using equation (2.2). Applying this result to our problem, $\varphi^*_{\text{opt}}(\cdot)$ of Theorem 12.1 is obtained by

finding the constant matrices $(A^F_{\text{opt}})^{q\times t_1}$ and $(A_{2\text{opt}})^{q\times t_2}$ that solve the linear equation

$$E\left[\left\{S_\beta(\cdot)-\left(A^F_{\text{opt}},A_{2\text{opt}}\right)^{q\times(t_1+t_2)}\begin{bmatrix}\dfrac{I(\mathcal{C}=\infty)m^*(Z,\beta_0)}{\varpi(\infty,Z)}\\ J_2(\cdot)\end{bmatrix}^{(t_1+t_2)\times 1}\right\}\times\left\{\left[\frac{I(\mathcal{C}=\infty)m^{*^T}(Z,\beta_0)}{\varpi(\infty,Z)},J_2^T(\cdot)\right]^{1\times(t_1+t_2)}\right\}\right]=0^{q\times(t_1+t_2)}. \tag{12.10}$$

Before deriving the solution to equation (12.10), we first give some results through a series of lemmas that will simplify the equation.

Lemma 12.1.

$$E\left[S_\beta\{\mathcal{C},G_{\mathcal{C}}(Z)\}\frac{I(\mathcal{C}=\infty)m^{*^T}(Z,\beta_0)A^{F^T}}{\varpi(\infty,Z)}\right]$$
$$=E\left\{S^F_\beta(Z)m^{*^T}(Z,\beta_0)A^{F^T}\right\} \tag{12.11}$$
$$=-\left[A^FE\left\{\frac{\partial m^*(Z,\beta_0)}{\partial\beta^T}\right\}\right]^T. \tag{12.12}$$

Proof. We prove (12.11) using a series of iterated conditional expectations, where

$$\begin{aligned}
&E\left[S_\beta\{\mathcal{C},G_{\mathcal{C}}(Z)\}\frac{I(\mathcal{C}=\infty)m^{*^T}(Z,\beta_0)A^{F^T}}{\varpi(\infty,Z)}\right]\\
&=E\left[E\left\{S^F_\beta(Z)|\mathcal{C},G_{\mathcal{C}}(Z)\right\}\frac{I(\mathcal{C}=\infty)m^{*^T}(Z,\beta_0)A^{F^T}}{\varpi(\infty,Z)}\right]\\
&=E\left[E\left\{S^F_\beta(Z)\frac{I(\mathcal{C}=\infty)m^{*^T}(Z,\beta_0)A^{F^T}}{\varpi(\infty,Z)}\middle|\mathcal{C},G_{\mathcal{C}}(Z)\right\}\right]\\
&=E\left\{S^F_\beta(Z)\frac{I(\mathcal{C}=\infty)m^{*^T}(Z,\beta_0)A^{F^T}}{\varpi(\infty,Z)}\right\}\\
&=E\left[E\left\{S^F_\beta(Z)\frac{I(\mathcal{C}=\infty)m^{*^T}(Z,\beta_0)A^{F^T}}{\varpi(\infty,Z)}\middle|Z\right\}\right]\\
&=E\left[S^F_\beta(Z)E\left\{\frac{I(\mathcal{C}=\infty)m^{*^T}(Z,\beta_0)A^{F^T}}{\varpi(\infty,Z)}\middle|Z\right\}\right]\\
&=E\left\{S^F_\beta(Z)m^{*^T}(Z,\beta_0)A^{F^T}\right\}.
\end{aligned}$$

Equation (12.12) follows from the usual expansion for m-estimators, where the influence function for the estimator that solves the equation

$$\sum_{i=1}^{n} A^F m^*(Z_i, \beta) = 0$$

has influence function

$$\varphi^F(Z) = -\left[A^F E\left\{\frac{\partial m^*(Z,\beta_0)}{\partial \beta^T}\right\}\right]^{-1} A^F m^*(Z,\beta_0).$$

Because $E\{\varphi^F(Z) S_\beta^{F^T}(Z)\} = I^{q\times q}$, this implies that

$$E\left\{A^F m^*(Z,\beta_0) S_\beta^{F^T}(Z)\right\} = -A^F E\left\{\frac{\partial m^*(Z,\beta_0)}{\partial \beta^T}\right\}.$$

The result in (12.12) now follows after taking the transpose of both sides of the equation above. □

Lemma 12.2.

$$E\left[S_\beta\{\mathcal{C}, G_{\mathcal{C}}(Z)\} J_2^T\{\mathcal{C}, G_{\mathcal{C}}(Z)\}\right] = 0^{q\times t_2}. \tag{12.13}$$

Proof. Using a series of iterated conditional expectations similar to the proof of Lemma 12.1, we obtain that

$$\begin{aligned} &E\left[S_\beta\{\mathcal{C}, G_{\mathcal{C}}(Z)\} J_2^T\{\mathcal{C}, G_{\mathcal{C}}(Z)\}\right] \\ &= E\left(S_\beta^F(Z) E\left[J_2^T\{\mathcal{C}, G_{\mathcal{C}}(Z)\}\middle| Z\right]\right). \end{aligned}$$

Because each of the elements of $J_2\{\mathcal{C}, G_{\mathcal{C}}(Z)\}$ is an element of Λ_2, this implies that

$$E\left[J_2\{\mathcal{C}, G_{\mathcal{C}}(Z)\}\middle| Z\right] = 0,$$

thus giving us the desired result. □

Lemma 12.3.

$$E\left\{\frac{I(\mathcal{C}=\infty)}{\varpi^2(\infty, Z)} m^*(Z,\beta_0) m^{*^T}(Z,\beta_0)\right\} = E\left\{\frac{m^*(Z,\beta_0) m^{*^T}(Z,\beta_0)}{\varpi(\infty, Z)}\right\}. \tag{12.14}$$

Proof. This follows easily by an iterated conditional expectation argument where we first compute the conditional expectation given Z. □

Using the results from Lemmas 12.1–12.3, we obtain the solution to (12.10) as

$$[A_{\text{opt}}^F, A_{2\text{opt}}] \begin{bmatrix} U_{11} & U_{12} \\ U_{12}^T & U_{22} \end{bmatrix} = [H_1, H_2], \tag{12.15}$$

where

$$\begin{aligned} U_{11} &= E\left\{ \frac{m^*(Z,\beta_0) m^{*^T}(Z,\beta_0)}{\varpi(\infty, Z)} \right\}^{t_1 \times t_1}, \\ U_{12} &= E\left\{ \frac{I(\mathcal{C}=\infty)}{\varpi(\infty,Z)} m^*(Z,\beta_0) J_2^T\{\mathcal{C}, G_{\mathcal{C}}(Z)\} \right\}^{t_1 \times t_2}, \\ U_{22} &= E\left\{ J_2\{\mathcal{C}, G_{\mathcal{C}}(Z)\} J_2^T\{\mathcal{C}, G_{\mathcal{C}}(Z)\} \right\}^{t_2 \times t_2}, \\ H_1 &= \left(- \left[E\left\{ \frac{\partial m^*(Z,\beta_0)}{\partial \beta^T} \right\} \right]^T \right)^{q \times t_1}, \\ H_2 &= 0^{q \times t_2}. \end{aligned} \tag{12.16}$$

Therefore, solving (12.15) yields

$$\begin{aligned} [A_{\text{opt}}^F, A_{2\text{opt}}] &= [H_1, 0] \begin{bmatrix} U_{11} & U_{12} \\ U_{12}^T & U_{22} \end{bmatrix}^{-1} \\ &= [H_1, 0] \begin{bmatrix} U^{11} & U^{12} \\ U^{12T} & U^{22} \end{bmatrix} \\ &= [H_1 U^{11}, H_1 U^{12}]. \end{aligned}$$

Using standard results for the inverse of partitioned symmetric matrices (see, for example, Rao, 1973, p.33), we obtain

$$A_{\text{opt}}^F = H_1^{q \times t_1} U^{11(t_1 \times t_1)} \tag{12.17}$$

and

$$A_{2\text{opt}} = H_1^{q \times t_1} U^{12(t_1 \times t_2)}, \tag{12.18}$$

where

$$U^{11} = \left(U_{11} - U_{12} U_{22}^{-1} U_{12}^T \right)^{-1} \tag{12.19}$$

and

$$U^{12} = -U^{11} U_{12} U_{22}^{-1}. \tag{12.20}$$

Thus we have shown that the optimal influence function $\varphi_{\text{opt}}\{\mathcal{C}, G_{\mathcal{C}}(Z)\}$ in $IF(\Xi)$, given by Theorem 12.1, is obtained by choosing $\varphi^*\{\mathcal{C}, G_{\mathcal{C}}(Z)\} \in \Xi$ to be

$$\varphi_{\text{opt}}^*\{\mathcal{C}, G_{\mathcal{C}}(Z)\} = \frac{I(\mathcal{C}=\infty) A_{\text{opt}}^F m^*(Z,\beta_0)}{\varpi(\infty, Z)} + A_{2\text{opt}} J_2\{\mathcal{C}, G_{\mathcal{C}}(Z)\}, \tag{12.21}$$

where A_{opt}^F and $A_{2\text{opt}}$ are defined by (12.17) and (12.18), respectively.

We also note the following interesting relationship.

Lemma 12.4. The projection of $\frac{I(\mathcal{C}=\infty)A_{\text{opt}}^F m^*(Z,\beta_0)}{\varpi(\infty,Z)}$ onto the space $\mathcal{G}_2$ (i.e., the space spanned by $J_2\{\mathcal{C}, G_{\mathcal{C}}(Z)\}$) is equal to

$$\Pi\left[\frac{I(\mathcal{C}=\infty)A_{\text{opt}}^F m^*(Z,\beta_0)}{\varpi(\infty,Z)}\middle|\mathcal{G}_2\right] = -A_{2\text{opt}}J_2\{\mathcal{C}, G_{\mathcal{C}}(Z)\}, \tag{12.22}$$

which implies that

$$\varphi_{\text{opt}}^*\{\mathcal{C}, G_{\mathcal{C}}(Z)\} = \frac{I(\mathcal{C}=\infty)A_{\text{opt}}^F m^*(Z,\beta_0)}{\varpi(\infty,Z)} + A_{2\text{opt}}J_2\{\mathcal{C}, G_{\mathcal{C}}(Z)\}$$

is orthogonal to $\mathcal{G}_2$; that is,

$$\varphi_{\text{opt}}^*\{\mathcal{C}, G_{\mathcal{C}}(Z)\} \perp \mathcal{G}_2. \tag{12.23}$$

Proof. Because $\mathcal{G}_2$ is a linear subspace spanned by $J_2\{\mathcal{C}, G_{\mathcal{C}}(Z)\}$, the projection (12.22) is given by

$$\begin{aligned}
&E\left[\frac{I(\mathcal{C}=\infty)A_{\text{opt}}^F m^*(Z,\beta_0)}{\varpi(\infty,Z)}J_2^T\{\mathcal{C}, G_{\mathcal{C}}(Z)\}\right] \\
&\times\left(E[J_2\{\mathcal{C}, G_{\mathcal{C}}(Z)\}J_2^T\{\mathcal{C}, G_{\mathcal{C}}(Z)\}]\right)^{-1} J_2\{\mathcal{C}, G_{\mathcal{C}}(Z)\} \\
&= A_{\text{opt}}^F U_{12}U_{22}^{-1}J_2\{\mathcal{C}, G_{\mathcal{C}}(Z)\} = H_1\left(U^{11}U_{12}U_{22}^{-1}\right)J_2\{\mathcal{C}, G_{\mathcal{C}}(Z)\} \\
&= -H_1U^{12}J_2\{\mathcal{C}, G_{\mathcal{C}}(Z)\} = -A_{2\text{opt}}J_2\{\mathcal{C}, G_{\mathcal{C}}(Z)\}. \quad \square
\end{aligned}$$

The optimal constant matrices A_{opt}^F and $A_{2\text{opt}}$ involve the quantities

$$H_1, U_{11}, U_{12}, U_{22},$$

which are all matrices whose elements are expectations. For practical applications, these must be estimated from the data. We propose the following empirical averages:

$$\hat{H}_1(\beta) = -\left[n^{-1}\sum_{i=1}^n \frac{I(\mathcal{C}_i=\infty)}{\varpi(\infty, Z_i, \hat{\psi}_n)}\left\{\frac{\partial m^*(Z_i,\beta)}{\partial\beta^T}\right\}\right]^T, \tag{12.24}$$

$$\hat{U}_{11}(\beta) = n^{-1}\sum_{i=1}^n \frac{I(\mathcal{C}_i=\infty)}{\varpi^2(\infty, Z_i, \hat{\psi}_n)}\left\{m^*(Z_i,\beta)m^{*T}(Z_i,\beta)\right\}, \tag{12.25}$$

$$\hat{U}_{12}(\beta) = n^{-1}\sum_{i=1}^n \frac{I(\mathcal{C}_i=\infty)}{\varpi(\infty, Z_i, \hat{\psi}_n)}\left[m^*(Z_i,\beta)J_2^T\{\mathcal{C}_i, G_{\mathcal{C}_i}(Z_i), \hat{\psi}_n\}\right], \tag{12.26}$$

and

$$\hat{U}_{22} = n^{-1} \sum_{i=1}^{n} J_2\{\mathcal{C}_i, G_{\mathcal{C}_i}(Z_i), \hat{\psi}_n\} J_2^T\{\mathcal{C}_i, G_{\mathcal{C}_i}(Z_i), \hat{\psi}_n\}. \tag{12.27}$$

Consequently, we would estimate

$$[\hat{A}^F_{\text{opt}}(\beta), \hat{A}_{2\text{opt}}(\beta)] = [\hat{H}_1(\beta)\hat{U}^{11}(\beta), \hat{H}_1(\beta)\hat{U}^{12}(\beta)], \tag{12.28}$$

where $\hat{U}^{11}(\beta)$ and $\hat{U}^{12}(\beta)$ are obtained by substituting the empirical estimates for U_{11}, U_{12}, and U_{22} into equations (12.19) and (12.20).

Deriving the Optimal Restricted (Class 1) AIPWCC Estimator

Using a sample of observed data $\{\mathcal{C}_i, G_{\mathcal{C}_i}(Z_i)\}, i = 1, \ldots, n$, we propose estimating β by solving the equation

$$\sum_{i=1}^{n} \left[\frac{I(\mathcal{C}_i = \infty)\hat{A}^F_{\text{opt}}(\beta) m^*(Z_i, \beta)}{\varpi(\infty, Z_i, \hat{\psi}_n)} + \hat{A}_{2\text{opt}}(\beta) J_2\{\mathcal{C}_i, G_{\mathcal{C}_i}(Z_i), \hat{\psi}_n\} \right] = 0. \tag{12.29}$$

We denote this estimator as $\hat{\beta}_{n\text{opt}}$ and now prove the fundamental result for restricted optimal estimators.

Theorem 12.2. Among the restricted class of AIPWCC estimators (12.3), the optimal estimator (i.e., the estimator with the smallest asymptotic variance matrix) is given by $\hat{\beta}_{n\text{opt}}$, the solution to (12.29).

Before sketching out the proof of Theorem 12.2, we give another equivalent representation for the class of influence functions $IF(\Xi)$ defined by (12.8) that will be useful.

Lemma 12.5. The class of influence functions $IF(\Xi)$ can also be defined by

$$\varphi\{\mathcal{C}, G_{\mathcal{C}}(Z)\} = \left[-A^F E\left\{ \frac{\partial m^*(Z, \beta_0)}{\partial \beta^T} \right\} \right]^{-1} \varphi^*\{\mathcal{C}, G_{\mathcal{C}}(Z)\}, \tag{12.30}$$

where $\varphi^*\{\mathcal{C}, G_{\mathcal{C}}(Z)\}$ is an element of Ξ defined by (12.5).

Proof. Using (12.12) of Lemma 12.1 and Lemma 12.2, we can show that

$$E[\varphi^*\{\mathcal{C}, G_{\mathcal{C}}(Z)\} S_\beta^T\{\mathcal{C}, G_{\mathcal{C}}(Z)\}] = -A^F E\left\{ \frac{\partial m^*(Z, \beta_0)}{\partial \beta^T} \right\}. \tag{12.31}$$

The lemma now follows by substituting the right-hand side of (12.31) into (12.8). □

Proof. Theorem 12.2
Since the asymptotic variance of an RAL estimator is the variance of its influence function, it suffices to consider the influence functions of the competing

estimators. In the same manner that we found the influence function of the estimator in (9.7) of Theorem 9.1, we can show that the influence function of the estimator for β that solves (12.3) is given by

$$\varphi\{\mathcal{C}, G_{\mathcal{C}}(Z)\} = \left[- A^F E \left\{ \frac{\partial m^*(Z, \beta_0)}{\partial \beta^T} \right\} \right]^{-1}$$
$$\times \left(\left[\frac{I(\mathcal{C} = \infty) A^F m^*(Z, \beta_0)}{\varpi(\infty, Z)} + A_2 J_2\{\mathcal{C}, G_{\mathcal{C}}(Z)\} \right] - \Pi\left[\left[\cdot \right] \middle| \Lambda_\psi \right] \right). \quad (12.32)$$

Because we constructed the space $\mathcal{G}_2$ so that $\Lambda_\psi \subset \mathcal{G}_2$, this implies that $\Pi[[\cdot]|\Lambda_\psi] \in \mathcal{G}_2$, which in turn implies that (12.32) is an element of Ξ. Therefore, as a consequence of Lemma 12.5, the influence function $\varphi\{\mathcal{C}, G_{\mathcal{C}}(Z)\}$ defined above is an element of $IF(\Xi)$. By Theorem 12.1, we know that

$$\text{var}[\varphi_{\text{opt}}\{\mathcal{C}, G_{\mathcal{C}}(Z)\} \leq \text{var}[\varphi\{\mathcal{C}, G_{\mathcal{C}}(Z)\}].$$

Hence, if we can show that the influence function of the estimator (12.29) is equal to $\varphi_{\text{opt}}\{\mathcal{C}, G_{\mathcal{C}}(Z)\}$, then we would complete the proof of the theorem.

An expansion of the estimating equation in (12.29) about $\beta = \beta_0$, keeping $\hat{\psi}_n$ fixed, yields

$$n^{1/2}(\hat{\beta}_{n\text{opt}} - \beta_0) = \left[- A^F_{\text{opt}} E \left\{ \frac{\partial m^*(Z, \beta_0)}{\partial \beta^T} \right\} \right]^{-1} \times$$
$$n^{-1/2} \sum_{i=1}^{n} \left[\frac{I(\mathcal{C}_i = \infty) \hat{A}^F_{\text{opt}}(\beta_0) m^*(Z_i, \beta_0)}{\varpi(\infty, Z_i, \hat{\psi}_n)} + \hat{A}_{2\text{opt}}(\beta_0) J_2\{\mathcal{C}_i, G_{\mathcal{C}_i}(Z_i), \hat{\psi}_n\} \right]$$
$$+ o_p(1). \quad (12.33)$$

We now show that estimating A^F_{opt} and $A_{2\text{opt}}$ only has a negligible effect on the asymptotic properties of the estimator by noting that (12.33) equals

$$n^{-1/2} \sum_{i=1}^{n} \left[\frac{I(\mathcal{C}_i = \infty) A^F_{\text{opt}} m^*(Z_i, \beta_0)}{\varpi(\infty, Z_i, \hat{\psi}_n)} + A_{2\text{opt}} J_2\{\mathcal{C}_i, G_{\mathcal{C}_i}(Z_i), \hat{\psi}_n\} \right] \quad (12.34)$$
$$+ n^{1/2}\{\hat{A}^F_{\text{opt}}(\beta_0) - A^F_{\text{opt}}\} n^{-1} \sum_{i=1}^{n} \frac{I(\mathcal{C}_i = \infty) m^*(Z_i, \beta_0)}{\varpi(\infty, Z_i, \hat{\psi}_n)} \quad (12.35)$$
$$+ n^{1/2}\{\hat{A}_{2\text{opt}}(\beta_0) - A_{2\text{opt}}\} n^{-1} \sum_{i=1}^{n} J_2\{\mathcal{C}_i, G_{\mathcal{C}_i}(Z_i), \hat{\psi}_n\}. \quad (12.36)$$

Under mild regularity conditions, $n^{1/2}\{\hat{A}^F_{\text{opt}}(\beta_0) - A^F_{\text{opt}}\}$ and $n^{1/2}\{\hat{A}_{2\text{opt}}(\beta_0) - A_{2\text{opt}}\}$ will be bounded in probability,

$$n^{-1} \sum_{i=1}^{n} \frac{I(\mathcal{C}_i = \infty) m^*(Z_i, \beta_0)}{\varpi(\infty, Z_i, \hat{\psi}_n)} \xrightarrow{P} E\left\{ \frac{I(\mathcal{C} = \infty) m^*(Z, \beta_0)}{\varpi(\infty, Z_i, \psi_0)} \right\}$$
$$= E\{m^*(Z, \beta_0)\} = 0,$$

and

$$n^{-1}\sum_{i=1}^{n} J_2\{\mathcal{C}_i, G_{\mathcal{C}_i}(Z_i), \hat{\psi}_n\} \xrightarrow{P} E[J_2\{\mathcal{C}, G_{\mathcal{C}}(Z), \psi_0\}] = 0.$$

Hence (12.35) and (12.36) will converge in probability to zero. Using Theorem 9.1, we can expand $\hat{\psi}_n$ about ψ_0 in (12.34) to obtain that (12.34) equals

$$\begin{aligned} n^{-1/2}\sum_{i=1}^{n}\Bigg(&\left[\frac{I(\mathcal{C}_i=\infty)A^F_{\text{opt}}m^*(Z_i,\beta_0)}{\varpi(\infty,Z_i,\psi_0)} + A_{2\text{opt}}J_2\{\mathcal{C}_i, G_{\mathcal{C}_i}(Z_i),\psi_0\}\right] \\ &- \Pi\left[\frac{I(\mathcal{C}_i=\infty)A^F_{\text{opt}}m^*(Z_i,\beta_0)}{\varpi(\infty,Z_i,\psi_0)} + A_{2\text{opt}}J_2\{\mathcal{C}_i, G_{\mathcal{C}_i}(Z_i),\psi_0\}\Big|\Lambda_\psi\right]\Bigg) \\ &+ o_p(1). \end{aligned} \tag{12.37}$$

Combining all the results from (12.33) through (12.37), we obtain that the influence function of $\hat{\beta}_{n\text{opt}}$, the solution to (12.29), is given by

$$\left[-A^F_{\text{opt}}E\left\{\frac{\partial m^*(Z,\beta_0)}{\partial\beta^T}\right\}\right]^{-1}\Big(\varphi^*_{\text{opt}}\{\mathcal{C}, G_{\mathcal{C}}(Z)\} - \Pi[\varphi^*_{\text{opt}}\{\mathcal{C}, G_{\mathcal{C}}(Z)\}|\Lambda_\psi]\Big), \tag{12.38}$$

where $\varphi^*_{\text{opt}}\{\mathcal{C}, G_{\mathcal{C}}(Z)\}$ is defined by (12.21). We proved that $\varphi^*_{\text{opt}}\{\mathcal{C}, G_{\mathcal{C}}(Z)\}$ is orthogonal to Λ_ψ in Corollary 1. We also demonstrated in (12.23) of Lemma 12.4 that $\varphi^*_{\text{opt}}\{\mathcal{C}, G_{\mathcal{C}}(Z)\}$ is orthogonal to $\mathcal{G}_2$. Since $\Lambda_\psi \subset \mathcal{G}_2$, this, too, implies that

$$\Pi[\varphi^*_{\text{opt}}\{\mathcal{C}, G_{\mathcal{C}}(Z)\}|\Lambda_\psi] = 0.$$

Therefore, the influence function (12.38) is equal to

$$\varphi_{\text{opt}}\{\mathcal{C}, G_{\mathcal{C}}(Z)\} = \left[-A^F_{\text{opt}}E\left\{\frac{\partial m^*(Z,\beta_0)}{\partial\beta^T}\right\}\right]^{-1}\varphi^*_{\text{opt}}\{\mathcal{C}, G_{\mathcal{C}}(Z)\}, \tag{12.39}$$

thus proving that the estimator that is the solution to (12.29) is the optimal restricted estimator. □

Estimating the Asymptotic Variance

Using the matrix relationships (12.16) and (12.17), we note that the leading term in (12.39) can be written as the symmetric matrix

$$\left[-A^F_{\text{opt}}E\left\{\frac{\partial m^*(Z,\beta_0)}{\partial\beta^T}\right\}\right]^{-1} = \left(H_1U^{11}H_1^T\right)^{-1}. \tag{12.40}$$

After a little algebra, we can also show that the covariance matrix of $\varphi^*_{\text{opt}}\{\mathcal{C}, G_{\mathcal{C}}(Z)\}$ is equal to

$$E\left\{\varphi^*_{\text{opt}}(\cdot)\varphi^{*^T}_{\text{opt}}(\cdot)\right\} = H_1U^{11}H_1^T. \tag{12.41}$$

Because the asymptotic variance of an RAL estimator is equal to the variance of its influence function, this means that the asymptotic variance of $\hat{\beta}_{n\text{opt}}$ is the variance of (12.39). Using (12.40) and (12.41), we obtain that the asymptotic variance is equal to

$$\left(H_1 U^{11} H_1^T\right)^{-1} \left(H_1 U^{11} H_1^T\right) \left(H_1 U^{11} H_1^T\right)^{-1} = \left(H_1 U^{11} H_1^T\right)^{-1}.$$

Consequently, a consistent estimator for the asymptotic variance of $\hat{\beta}_{n\text{opt}}$ is given by

$$\left\{\hat{H}_1(\hat{\beta}_{n\text{opt}}) \hat{U}^{11}(\hat{\beta}_{n\text{opt}}) \hat{H}_1^T(\hat{\beta}_{n\text{opt}})\right\}^{-1}, \tag{12.42}$$

where $\hat{H}_1(\beta)$ and $\hat{U}^{11}(\beta)$ were defined by (12.24) and (12.28).

Remark 4. The method we proposed for estimating the parameter β using restricted optimal (class 1) estimators only needs that the model for the coarsening probabilities be correctly specified. The appeal of this method is its simplicity. We did not have to use adaptive methods, where a simpler model $p_Z^*(z, \xi)$ had to be posited and an estimator for ξ had to be derived. Yet the resulting estimator is guaranteed to have the smallest asymptotic variance within the class of estimators considered. It would certainly be more efficient than the simple inverse probability weighted complete-case estimator (IPWCC), which discards information from data that are not completely observed. However, this estimator is not efficient. How close the variance of such an estimator will be to the semiparametric efficiency bound will depend on how close the optimal element $B_{\text{eff}}^F(Z) \in \Lambda^{F\perp}$ is to the subspace $\mathcal{G}^F$ and how close the optimal element in Λ_2, $\Pi[\frac{I(C=\infty)B_{\text{eff}}^F(Z)}{\varpi(\infty,Z)}|\Lambda_2]$, is to $\mathcal{G}_2$.

If the missingness were by design, then restricted optimal (class 1) estimators would be guaranteed (subject to regularity conditions) to yield consistent, asymptotically normal estimators for β. However, if the coarsening probabilities are modeled and not correctly specified, then such estimators will be biased. There is no double-robustness protection guaranteed for such estimators. Therefore, in the next section, we consider what we refer to as (class 2) estimators. These estimators, although more complicated, will result in double-robust estimators that avoid the necessity to solve complicated integral equations. □

Before discussing restricted optimal (class 2) estimators, we first illustrate how to derive a restricted optimal (class 1) estimator by considering a specific example.

12.3 Example of an Optimal Restricted (Class 1) Estimator

We return to the example introduced in Section 7.4 where the goal is to estimate the parameter β in the restricted moment model

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

In this example, we let Y be a univariate response variable and $X = (X^{(1)}, X^{(2)})^T$ be two univariate covariates. The second covariate $X^{(2)}$ is expensive to measure and therefore, by design, was only collected on a subsample of the n individuals in the study, whereas $X^{(1)}$ was collected on all n individuals. The subsample of individuals for which $X^{(2)}$ was collected was chosen at random with a prespecified probability that depended on Y and $X^{(1)}$. This is an example of two levels of missingness, where we denote the complete-case indicator for the i-th individual as R_i (unscripted); that is, if $R_i = 1$, then we observe $(Y_1, X_i^{(1)}, X_i^{(2)})$, whereas if $R_i = 0$, then we observe $(Y_i, X_i^{(1)})$. The probability of a complete case is denoted by

$$P(R_i = 1|Y_i, X_i) = P(R_i = 1|Y_i, X_i^{(1)}) = \pi(Y_i, X_i^{(1)}),$$

where $\pi(Y_i, X_i^{(1)})$ (unscripted) is a known function of Y_i and $X_i^{(1)}$. Since the coarsening of the data was by design, the corresponding coarsening tangent space $\Lambda_\psi = 0$. Therefore, we need not worry that the vector J_2 that spans $\mathcal{G}_2$ contains the score vector S_ψ for this example.

To define the restricted class of estimators, we must first choose finite-dimensional subsets $\mathcal{G}^F \subset \Lambda^{F\perp} \subset \mathcal{H}^F$ and $\mathcal{G}_2 \subset \Lambda_2 \subset \mathcal{H}$. We remind the reader that the space $\Lambda^{F\perp}$ consists of elements

$$\left\{ h(X, \beta_0)^{q\times 1}\{Y - \mu(X, \beta_0)\} \text{ for arbitrary } q\text{-dimensional functions } h(\cdot) \text{ of } X \right\}$$

and the space Λ_2 consists of elements

$$\left\{ f(Y, X^{(1)})^{q\times 1}\{R - \pi(Y, X^{(1)})\} \text{ for arbitrary } q\text{-dimensional functions of } (Y, X^{(1)}) \right\}.$$

Recall that X, by itself, refers to $(X^{(1)}, X^{(2)})$.

To define $\mathcal{G}^F$, we must choose a t_1-dimensional function of the full-data (Y, X), say $J^F(Y, X, \beta) = \{J_1^F(Y, X, \beta), \ldots, J_{t_1}^F(Y, X, \beta)\}^T$, that spans $\mathcal{G}^F$, where the elements $J_j^F(Y, X, \beta_0) \in \Lambda^{F\perp^{(1)}}, j = 1, \ldots, t_1$, and where $\Lambda^{F\perp^{(1)}}$ denotes the linear subspace in $\mathcal{H}^{F^{(1)}}$ spanned by the first element of the q-dimensional vector that makes up $\Lambda^{F\perp}$; i.e., $\Lambda^{F\perp} = \{\Lambda^{F\perp^{(1)}}\}^q$.

Suppose, for example, we were considering a log-linear model, where

$$\mu(X,\beta) = \exp(\beta_1 + \beta_2 X^{(1)} + \beta_3 X^{(2)}).$$

If we also believed the variance as a function of X was homoscedastic, then the optimal full-data estimating function would be chosen to be

$$m(Y,X,\beta) = D^T(X,\beta)V^{-1}(X)\{Y - \exp(\beta_1 + \beta_2 X^{(1)} + \beta_3 X^{(2)})\},$$

where $D(X,\beta) = \partial\mu(X,\beta)/\partial\beta^T$. Therefore, for this example, we would choose

$$m(Y,X,\beta) = (1, X^{(1)}, X^{(2)})^T \exp(\beta^T X^*)\{Y - \exp(\beta^T X^*)\},$$

where $X^* = (1, X^{(1)}, X^{(2)})^T$ and $\beta = (\beta_1, \beta_2, \beta_3)^T$, as our optimal full-data estimating function if the data were not coarsened. However, this choice for $m(Y,X,\beta)$ may not be optimal with coarsened data. Therefore, we consider restricted optimal estimators, where we might choose

$$J^F(Y,X) = f^F(X)\exp(\beta_0^T X^*)\{Y - \exp(\beta_0^T X^*)\},$$

where $f^F(X) = (1, X^{(1)}, X^{(2)}, X^{(1)^2}, X^{(2)^2}, X^{(1)}X^{(2)})^T$. Such a set of basis functions allows for a quadratic relationship in $X^{(1)}$ and $X^{(2)}$. We also define the t_1-dimensional vector of estimating functions as

$$m^*(Y,X,\beta) = f^F(X)\exp(\beta^T X^*)\{Y - \exp(\beta^T X^*)\}. \tag{12.43}$$

To define $\mathcal{G}_2$, we must choose a t_2-dimensional function of the observed data $(R, Y, X^{(1)}, RX^{(2)})$, say $J_2(\cdot) = \{J_{21}(\cdot), \ldots, J_{2t_2}(\cdot)\}^T$, that spans $\mathcal{G}_2$, where the elements $J_{2j}(\cdot) \in \Lambda_2^{(1)}, j = 1, \ldots, t_2$, and where $\Lambda_2^{(1)}$ denotes the linear subspace in $\mathcal{H}^{(1)}$ spanned by the first element of the q-dimensional vector that makes up Λ_2. For example, we might choose

$$J_2(\cdot) = f_2(Y, X^{(1)})\{R - \pi(Y, X^{(1)})\}, \tag{12.44}$$

where $f_2(Y, X^{(1)}) = (1, Y, X^{(1)}, Y^2, X^{(1)^2}, YX^{(1)})^T$. Such a set of basis functions allows for a quadratic relationship in Y and $X^{(1)}$.

Therefore, the class of restricted estimators that will be considered are solutions to the estimating equations (12.3), namely

$$\sum_{i=1}^{n}\left[\frac{R_i A^F f^F(X_i)\exp(\beta^T X_i^*)\{Y_i - \exp(\beta^T X_i^*)\}}{\pi(Y_i, X_i^{(1)})} + \{R_i - \pi(Y_i, X_i^{(1)})\}A_2 f_2(Y_i, X_i^{(1)})\right] = 0, \tag{12.45}$$

for an arbitrary $q \times t_1$ constant matrix A^F and an arbitrary $q \times t_2$ constant matrix A_2. For this illustration, $q = 3$ and $t_1 = t_2 = 6$.

Finding the optimal estimator within this restricted class and deriving the asymptotic variance are now just a matter of plugging into the formulas derived in the previous section.

Taking the partial derivative of (12.43) with respect to β yields

$$E\left\{\frac{\partial m^*(Y,X,\beta_0)}{\partial \beta^T}\right\} = -f^F(X)\exp(\beta_0^T X^*)D(X,\beta_0)$$
$$= f^F(X)\exp(2\beta_0^T X^*){X^*}^T.$$

Therefore, we obtain a consistent estimator for $H_1(\beta)$ by using

$$\hat{H}_1(\beta) = n^{-1}\sum_{i=1}^{n}\frac{R_i}{\pi(Y_i, X_i^{(1)})}X_i^*\exp(2\beta^T X_i^*)f^{F^T}(X_i). \tag{12.46}$$

Next we use equations (12.25)–(12.27) to compute

$$\hat{U}_{11}(\beta) = n^{-1}\sum_{i=1}^{n}\frac{R_i\{Y_i - \exp(\beta^T X_i^*)\}^2}{\pi^2(Y_i, X_i^{(1)})}f^F(X_i)f^{F^T}(X_i), \tag{12.47}$$

$$\hat{U}_{12}(\beta) = n^{-1}\sum_{i=1}^{n}\frac{R_i\{R_i - \pi(Y_i, X_i^{(1)})\}}{\pi(Y_i, X_i^{(1)})}\{Y_i - \exp(\beta^T X_i^*)\}f^F(X_i)f_2^T(Y_i, X_i^{(1)}), \tag{12.48}$$

and

$$\hat{U}_{22} = n^{-1}\sum_{i=1}^{n}\{R_i - \pi(Y_i, X_i^{(1)})\}^2 f_2(Y_i, X_i^{(1)})f_2^T(Y_i, X_i^{(1)}). \tag{12.49}$$

From these, we can compute

$$\hat{U}^{11}(\beta) = \{\hat{U}_{11}(\beta) - \hat{U}_{12}(\beta)\hat{U}_{22}^{-1}(\beta)\hat{U}_{12}^T(\beta)\}^{-1}, \tag{12.50}$$
$$\hat{U}^{12}(\beta) = -\hat{U}^{11}(\beta)\hat{U}_{12}(\beta)\hat{U}_{22}^{-1}(\beta), \tag{12.51}$$
$$\hat{A}_{\text{opt}}^F(\beta) = \hat{H}_1(\beta)\hat{U}^{11}(\beta), \tag{12.52}$$

and

$$\hat{A}_{2\text{opt}}(\beta) = \hat{H}_1(\beta)\hat{U}^{12}(\beta). \tag{12.53}$$

Therefore, the optimal restricted estimator is the solution to the estimating equation

$$\sum_{i=1}^{n}\Bigg[\frac{R_i\hat{A}_{\text{opt}}^F(\beta)f^F(X_i)\exp(\beta^T X_i^*)\{Y_i - \exp(\beta^T X_i^*)\}}{\pi(Y_i, X_i^{(1)})}$$
$$+ \{R_i - \pi(Y_i, X_i^{(1)})\}\hat{A}_{2\text{opt}}(\beta)f_2(Y_i, X_i^{(1)})\Bigg] = 0, \tag{12.54}$$

which we denote by $\hat{\beta}_{n\text{opt}}$. Moreover, the asymptotic variance for $\hat{\beta}_{n\text{opt}}$ can be estimated using

$$\left\{ \hat{H}_1(\hat{\beta}_{n\text{opt}}) \hat{U}^{11}(\hat{\beta}_{n\text{opt}}) \hat{H}_1^T(\hat{\beta}_{n\text{opt}}) \right\}^{-1},$$

where $\hat{H}_1(\beta)$ and $\hat{U}^{11}(\beta)$ were defined by (12.46) and (12.50), respectively.

Modeling the Missingness Probabilities

In the example above, it was assumed that the missing values of $X^{(2)}$ were by design, where the investigator had control of the missingness probabilities. We now consider how the methods would be modified if these probabilities were not known and had to be estimated from the data. We might, for instance, consider the logistic regression model where

$$P(R = 1|Y, X^{(1)}) = \pi(Y, X^{(1)}, \psi) = \frac{\exp(\psi_0 + \psi_1 Y + \psi_2 X^{(1)})}{1 + \exp(\psi_0 + \psi_1 Y + \psi_2 X^{(1)})}. \quad (12.55)$$

The maximum likelihood estimator for $\psi = (\psi_0, \psi_1, \psi_2)^T$ is obtained by maximizing the likelihood given by (8.10); specifically,

$$\prod_{i=1}^{n} \left[\frac{\exp\{(\psi_0 + \psi_1 Y_i + \psi_2 X_i^{(1)}) R_i\}}{1 + \exp(\psi_0 + \psi_1 Y_i + \psi_2 X_i^{(1)})} \right],$$

and the resulting MLE is denoted by $\hat{\psi}_n = (\hat{\psi}_{0n}, \hat{\psi}_{1n} \hat{\psi}_{2n})^T$. Using standard results for the logistic regression model, we note that the score vector $S_\psi(\cdot)$ is given by

$$S_\psi(R, Y, X^{(1)}, \psi) = (1, Y, X^{(1)})^T \{R - \pi(Y, X^{(1)}, \psi)\}, \quad (12.56)$$

where $\pi(Y, X^{(1)}, \psi)$ is the probability of a complete case defined in (12.55). We also note that the score equation, evaluated at the MLE, is equal to zero; that is,

$$\sum_{i=1}^{n} S_\psi(R_i, Y_i, X_i^{(1)}, \hat{\psi}_n) = 0. \quad (12.57)$$

As mentioned in Remark 3 of this chapter, we should choose the score vector $S_\psi(\cdot)$ as part of a set of basis functions $J_2(\cdot)$ that spans the space $\mathcal{G}_2$. Examining the set of functions $J_2(\cdot)$ given by (12.44) for the example above, we note that indeed $S_\psi(\cdot)$ makes up the first three elements of $J_2(\cdot)$.

The estimator for β would then be identical to that given in equation (12.54) except that we would substitute $\pi(Y_i, X_i^{(1)}, \hat{\psi}_n)$ for $\pi(Y, X^{(1)})$ in the equation (12.54) itself and when evaluating all the quantities in equations (12.46)–(12.53).

Remark 5. We point out that the second term in the estimating equation (12.54) (i.e., the augmented term) can be written as

$$\hat{A}_{2\text{opt}} \sum_{i=1}^{n} f_2(Y_i, X_i^{(1)})\{R_i - \pi(Y_i, X_i^{(1)}, \hat{\psi}_n)\}.$$

Since the first three elements of the vector

$$\sum_{i=1}^{n} f_2(Y_i, X_i^{(1)})\{R_i - \pi(Y_i, X_i^{(1)}, \hat{\psi}_n)\} \tag{12.58}$$

are $\sum_{i=1}^{n} S_\psi(R_i, Y_i, X_i^{(1)}, \hat{\psi}_n)$, then as a consequence of (12.57), this means that the first three elements of the vector (12.58) are equal to zero. This observation may result in some modest savings in computation.

12.4 Optimal Restricted (Class 2) Estimators

(Class 2) restricted estimators are AIPWCC estimators where we restrict attention to a finite-dimensional linear subspace $\mathcal{G}^F \in \Lambda^{F\perp}$ spanned by the vector $J^F(Z)$, as we did for (class 1) restricted estimators, but where $\mathcal{G}_2 = \Lambda_2$. By so doing, we will show that the resulting optimal estimator for β within this class will have an influence function that is an element of the class of double-robust influence functions $(IF)_{\text{DR}}$; see Definition 2 of Chapter 10. Toward that end, we define the linear space Ξ similar to what we did in (12.4); that is,

$$\Xi = \left\{ \frac{I(\mathcal{C} = \infty)\mathcal{G}^F}{\varpi(\infty, Z)} \oplus \Lambda_2 \right\}, \tag{12.59}$$

and Ξ consists of the elements

$$\frac{I(\mathcal{C} = \infty)A^F m^*(Z, \beta_0)}{\varpi(\infty, Z)} + L_2\{\mathcal{C}, G_{\mathcal{C}}(Z)\} \tag{12.60}$$

in $\mathcal{H}$ for any constant matrix $A^{F^{q\times t_1}}$ and $L_2 \in \Lambda_2$, where $m^*(Z, \beta)$ is a $t_1 \times 1$ vector of estimating functions such that $m^*(Z, \beta_0) = J^F(Z)$. We denote an element within this class as $\varphi^*\{\mathcal{C}, G_{\mathcal{C}}(Z)\}$ and define $IF(\Xi)$ to be the elements within this class that satisfy (12.7). The elements within the class $IF(\Xi)$ are defined as $\varphi\{\mathcal{C}, G_{\mathcal{C}}(Z)\}$ and are given by (12.8).

We now prove the following key theorem for (class 2) restricted estimators.

Theorem 12.3. Among the elements $\varphi\{\mathcal{C}, G_{\mathcal{C}}(Z)\} \in IF(\Xi)$, the one with the smallest variance matrix is given by

$$\varphi_{\text{opt}}\{\mathcal{C}, G_{\mathcal{C}}(Z)\} = \mathcal{J}\{\varphi_{\text{opt}}^F(Z)\}, \tag{12.61}$$

where the linear operator $\mathcal{J}(\cdot)$ is given by Definition 1 of Chapter 10; that is,

$$\mathcal{J}\{\varphi^F_{\text{opt}}(Z)\} = \frac{I(\mathcal{C}=\infty)\varphi^F_{\text{opt}}(Z)}{\varpi(\infty, Z)} - \Pi\left[\frac{I(\mathcal{C}=\infty)\varphi^F_{\text{opt}}(Z)}{\varpi(\infty, Z)}\middle|\Lambda_2\right],$$

$$\varphi^F_{\text{opt}}(Z) = [E\{\varphi^{*F}_{\text{opt}}(Z)S^{F^T}_\beta(Z)\}]^{-1}\varphi^{*F}_{\text{opt}}(Z), \tag{12.62}$$

and $\varphi^{*F}_{\text{opt}}(Z)$ is the unique element in $\mathcal{G}^F$ that solves the equation

$$\Pi[S^F_\beta(Z) - \mathcal{M}^{-1}\{\varphi^{*F}_{\text{opt}}(Z)\}|\mathcal{G}^F] = 0, \tag{12.63}$$

where $\mathcal{M}^{-1}$ is the inverse of the linear operator $\mathcal{M}$ given by Definition 5 of Chapter 10. (The inverse operator $\mathcal{M}^{-1}$ exists and is unique; see Lemma 10.5.)

Proof. We first will prove that $IF(\Xi)$ consists of the class of elements

$$\frac{I(\mathcal{C}=\infty)\varphi^F(Z)}{\varpi(\infty, Z)} + L^*_2\{\mathcal{C}, G_\mathcal{C}(Z)\}, \tag{12.64}$$

where

$$\varphi^F(Z) = [E\{\varphi^{*F}(Z)S^{F^T}_\beta(Z)\}]^{-1}\varphi^{*F}(Z),$$

$\varphi^{*F}(Z) \in \mathcal{G}^F$, and $L^*_2\{\mathcal{C}, G_\mathcal{C}(Z)\} \in \Lambda_2$.

Because the elements in $IF(\Xi)$ are defined by (12.8), (12.64) will be true if we can show that

$$E[\varphi^*\{\mathcal{C}, G_\mathcal{C}(Z)\}S^T_\beta\{\mathcal{C}, G_\mathcal{C}(Z)\}] = E\{\varphi^{*F}(Z)S^{F^T}_\beta(Z)\}, \tag{12.65}$$

where $\varphi^*\{\mathcal{C}, G_\mathcal{C}(Z)\} \in \Xi$ is equal to

$$\varphi^*\{\mathcal{C}, G_\mathcal{C}(Z)\} = \frac{I(\mathcal{C}=\infty)\varphi^{*F}(Z)}{\varpi(\infty, Z)} + L_2\{\mathcal{C}, G_\mathcal{C}(Z)\}.$$

Because $S_\beta\{\mathcal{C}, G_\mathcal{C}(Z)\} = E\{S^F_\beta(Z)|\mathcal{C}, G_\mathcal{C}(Z)\}$, we use a series of iterated conditional expectations to obtain

$$\begin{aligned}
E[\varphi^*\{\mathcal{C}, G_\mathcal{C}(Z)\}S^T_\beta\{\mathcal{C}, G_\mathcal{C}(Z)\}] &= E(E[\varphi^*\{\mathcal{C}, G_\mathcal{C}(Z)\}S^{F^T}_\beta(Z)|\mathcal{C}, G_\mathcal{C}(Z)]) \\
&= E[\varphi^*\{\mathcal{C}, G_\mathcal{C}(Z)\}S^{F^T}_\beta(Z)] \\
&= E(E[\varphi^*\{\mathcal{C}, G_\mathcal{C}(Z)\}S^{F^T}_\beta(Z)|Z]) \\
&= E(E[\varphi^*\{\mathcal{C}, G_\mathcal{C}(Z)\}|Z]S^{F^T}_\beta(Z)) \\
&= E\{\varphi^{*F}(Z)S^{F^T}_\beta(Z)\},
\end{aligned}$$

thus proving (12.65).

By proving (12.64), we have demonstrated that all the elements in $IF(\Xi)$ can be written as

$$\varphi\{\mathcal{C}, G_{\mathcal{C}}(Z)\} = \frac{I(\mathcal{C}=\infty)\varphi^F(Z)}{\varpi(\infty, Z)} + L_2\{\mathcal{C}, G_{\mathcal{C}}(Z)\}, \tag{12.66}$$

where $\varphi^F(Z)$ is a full-data influence function and $L_2 \in \Lambda_2$. Because of Theorem 10.1, we know that, for a fixed $\varphi^F(Z)$, the optimal element (smallest variance matrix) among the class of elements (12.66) is given by $\mathcal{J}\{\varphi^F(Z)\}$. Therefore, we only need to restrict attention to those elements of $IF(\Xi)$ that are in the class

$$\left\{\mathcal{J}\{\varphi^F(Z)\} : \varphi^F(Z) = [E\{\varphi^{*F}(Z)S_\beta^{F^T}(Z)\}]^{-1}\varphi^{*F}(Z)\right\},$$

where $\varphi^{*F}(Z) \in \mathcal{G}^F$, if the goal is to find the optimal estimator in $IF(\Xi)$. Equivalently, we can restrict the search to the elements in the linear subspace $\mathcal{J}(\mathcal{G}^F) \subset \Xi$ that satisfy (12.7).

In Theorem 10.6, we proved that $\mathcal{J}\{h^F(Z)\} = \mathcal{L}[\mathcal{M}^{-1}\{h^F(Z)\}]$, where the linear operator $\mathcal{L}$ was defined by Definition 4 of Chapter 10 and, we remind the reader, is given by

$$\mathcal{L}\{h^F(Z)\} = E\{h^F(Z)|\mathcal{C}, G_{\mathcal{C}}(Z)\}.$$

Therefore, the linear space $\mathcal{J}(\mathcal{G}^F) = \mathcal{L}\{\mathcal{M}^{-1}(\mathcal{G}^F)\}$.

Using the same proof as for Theorem 12.1, we can also prove that among the elements $\varphi\{\mathcal{C}, G_{\mathcal{C}}(Z)\} \in \mathcal{J}(\mathcal{G}^F) = \mathcal{L}\{\mathcal{M}^{-1}(\mathcal{G}^F)\}$, the one with the smallest variance matrix is the normalized version of the projection of the observed-data score vector onto $\mathcal{J}(\mathcal{G}^F)$. Specifically,

$$\varphi_{\text{opt}}\{\mathcal{C}, G_{\mathcal{C}}(Z)\} = \left(E[\varphi^*_{\text{opt}}\{\mathcal{C}, G_{\mathcal{C}}(Z)\}S_\beta^T\{\mathcal{C}, G_{\mathcal{C}}(Z)\}]\right)^{-1}\varphi^*_{\text{opt}}\{\mathcal{C}, G_{\mathcal{C}}(Z)\}, \tag{12.67}$$

where

$$\varphi^*_{\text{opt}}\{\mathcal{C}, G_{\mathcal{C}}(Z)\} = \Pi[S_\beta\{\mathcal{C}, G_{\mathcal{C}}(Z)\}|\mathcal{L}\{\mathcal{M}^{-1}(\mathcal{G}^F)\}]. \tag{12.68}$$

We now show how to derive this projection.

Because of the projection theorem for Hilbert spaces, the projection of $S_\beta\{\mathcal{C}, G_{\mathcal{C}}(Z)\}$ onto the closed linear space $\mathcal{L}\{\mathcal{M}^{-1}(\mathcal{G}^F)\}$ is the unique element $\varphi^{*F}_{\text{opt}}(Z) \in \mathcal{G}^F$ that satisfies

$$E\left\{\left(S_\beta\{\mathcal{C}, G_{\mathcal{C}}(Z)\} - \mathcal{L}[\mathcal{M}^{-1}\{\varphi^{*F}_{\text{opt}}(Z)\}]\right)^T\mathcal{L}[\mathcal{M}^{-1}\{\varphi^{*F}(Z)\}]\right\} = 0, \tag{12.69}$$

for all $\varphi^{*F}(Z) \in \mathcal{G}^F$. Recalling that $S_\beta\{\mathcal{C}, G_{\mathcal{C}}(Z)\} = E\{S_\beta^F(Z)|\mathcal{C}, G_{\mathcal{C}}(Z)\}$ and $\mathcal{L}[\mathcal{M}^{-1}\{\varphi^{*F}(Z)\}] = E\{\mathcal{M}^{-1}(\varphi^{*F})|\mathcal{C}, G_{\mathcal{C}}(Z)\}$, we use a series of iterated conditional expectations to write (12.69) as

$$
\begin{aligned}
0 &= E\Big(E[\{S_\beta^F - \mathcal{M}^{-1}(\varphi_{\text{opt}}^{*F})\}^T \mathcal{L}\{\mathcal{M}^{-1}(\varphi^{*F})\}|\mathcal{C}, G_{\mathcal{C}}(Z)]\Big) \\
&= E\Big[\{S_\beta^F - \mathcal{M}^{-1}(\varphi_{\text{opt}}^{*F})\}^T \mathcal{L}\{\mathcal{M}^{-1}(\varphi^{*F})\}\Big] \\
&= E\Big(E[\{S_\beta^F - \mathcal{M}^{-1}(\varphi_{\text{opt}}^{*F})\}^T \mathcal{L}\{\mathcal{M}^{-1}(\varphi^{*F})\}|Z]\Big) \\
&= E\Big(\{S_\beta^F - \mathcal{M}^{-1}(\varphi_{\text{opt}}^{*F})\}^T E[\mathcal{L}\{\mathcal{M}^{-1}(\varphi^{*F})\}|Z]\Big) \\
&= E\Big[\{S_\beta^F - \mathcal{M}^{-1}(\varphi_{\text{opt}}^{*F})\}^T \mathcal{M}\{\mathcal{M}^{-1}(\varphi^{*F})\}\Big] \\
&= E\Big[\{S_\beta^F - \mathcal{M}^{-1}(\varphi_{\text{opt}}^{*F})\}^T \varphi^{*F}\Big]. \qquad (12.70)
\end{aligned}
$$

Therefore, (12.70) being true for all $\varphi^{*F} \in \mathcal{G}^F$ implies that $S_\beta^F - \mathcal{M}^{-1}(\varphi_{\text{opt}}^{*F})$ must be orthogonal to $\mathcal{G}^F$. Since the projection exists and is unique, this implies that there must exist a unique element $\varphi_{\text{opt}}^{*F} \in \mathcal{G}^F$ such that (12.70) holds for all $\varphi^{*F} \in \mathcal{G}^F$, or equivalently

$$\Pi[S_\beta^F(Z) - \mathcal{M}^{-1}\{\varphi_{\text{opt}}^{*F}(Z)\}|\mathcal{G}^F] = 0. \qquad (12.71)$$

Consequently,

$$\varphi_{\text{opt}}^{*}\{\mathcal{C}, G_{\mathcal{C}}(Z)\} = \mathcal{L}[\mathcal{M}^{-1}\{\varphi_{\text{opt}}^{*F}(Z)\}] = \mathcal{J}\{\varphi_{\text{opt}}^{*F}(Z)\},$$

where $\varphi_{\text{opt}}^{*F}(Z)$ satisfies (12.71), or (12.63) of the theorem. The proof is complete if we can show that

$$E[\varphi_{\text{opt}}^{*}\{\mathcal{C}, G_{\mathcal{C}}(Z)\}S_\beta^T\{\mathcal{C}, G_{\mathcal{C}}(Z)\}]$$

of equation (12.67), where

$$\varphi_{\text{opt}}^{*}\{\mathcal{C}, G_{\mathcal{C}}(Z)\} = \mathcal{L}[\mathcal{M}^{-1}\{\varphi_{\text{opt}}^{*F}(Z)\}],$$

is the same as $E\{\varphi_{\text{opt}}^{*F}(Z)S_\beta^{F^T}(Z)\}$. This can be shown by using the same iterated expectations argument that led to (12.70). □

A corollary that is an immediate consequence of Theorem 12.3 by taking $\mathcal{G}^F = \Lambda^{F\perp}$ is given as follows.

Corollary 2. Among all influence functions in

$$\Xi = \left\{\frac{I(\mathcal{C} = \infty)\Lambda^{F\perp}}{\varpi(\infty, Z)} \oplus \Lambda_2\right\}$$

(that is, elements of Ξ that satisfy (12.7)), the one with the smallest variance matrix is obtained by choosing

$$\varphi^*_{\text{opt}}\{\mathcal{C}, G_{\mathcal{C}}(Z)\} = \mathcal{J}\{\varphi^{*F}_{\text{opt}}(Z)\},$$

where $\varphi^{*F}_{\text{opt}}(Z)$ is the unique element in $\Lambda^{F\perp}$ that satisfies

$$\Pi[S^F_\beta(Z) - \mathcal{M}^{-1}\{\varphi^{*F}_{\text{opt}}(Z)\}|\Lambda^{F\perp}] = 0$$

or equivalently solves the equation

$$\Pi[\mathcal{M}^{-1}\{\varphi^{*F}_{\text{opt}}(Z)\}|\Lambda^{F\perp}] = \Pi[S^F_\beta(Z)|\Lambda^{F\perp}] = S^F_{\text{eff}}(Z).$$

Remark 6. Since the space Ξ in the corollary is the same as $\Lambda^\perp_\eta$, then the result above is an alternative proof of Theorem 11.1, which was used to derive the optimal influence function among all AIPWCC estimators for β. □

Returning to the restricted (class 2) estimators, we obtain the following corollary.

Corollary 3. Let $m^*(Z,\beta)$ be a $t_1 \times 1$ ($t_1 > q$) vector of estimating functions such that $m^*(Z,\beta_0) = J^F(Z)$ spans the linear space $\mathcal{G}^F \subset \Lambda^{F\perp}$. Then, among the class of influence functions $IF(\Xi)$, where Ξ is defined by (12.59), the optimal element (smallest variance matrix) is given by $\varphi^F_{\text{opt}}(Z)$, defined by (12.62), the normalized version of

$$\varphi^{*F}_{\text{opt}} = A^{F^{q\times t_1}}_{\text{opt}} m^*(Z,\beta_0),$$

where

$$A^{F^{q\times t_1}}_{\text{opt}} = -\left[E\left\{\frac{\partial m^*(Z,\beta_0)}{\partial \beta^T}\right\}\right]^T \left(E[\mathcal{M}^{-1}\{m^*(Z,\beta_0)\}m^{*^T}(Z,\beta_0)]\right)^{-1}. \tag{12.72}$$

Proof. Using (12.63), we are looking for $\varphi^{*F}_{\text{opt}} = A^{F^{q\times t_1}}_{\text{opt}} m^*(Z,\beta_0)$ such that

$$\Pi\left[S^F_\beta(Z) - \mathcal{M}^{-1}\{A^{F^{q\times t_1}}_{\text{opt}} m^*(Z,\beta_0)\}\middle|\mathcal{G}^F\right] = 0. \tag{12.73}$$

Because $\mathcal{G}^F$ is spanned by $m^*(Z,\beta_0)$, the standard results for projecting onto a finite-dimensional linear space yield

$$\Pi[h^F(Z)|\mathcal{G}^F] = E(h^F m^{*^T})\{E(m^* m^{*^T})\}^{-1} m^*(Z,\beta_0). \tag{12.74}$$

Using (12.74), we write equation (12.73) as

$$E[\{S^F_\beta - A^F_{\text{opt}}\mathcal{M}^{-1}(m^*)\}m^{*^T}]\{E(m^* m^{*^T})\}^{-1} m^*(Z,\beta_0) = 0. \tag{12.75}$$

Because $m^*(Z,\beta_0)$ is made up of t_1 linearly independent elements, this implies that the variance matrix $E(m^* m^{*^T})$ is positive definite and hence has a unique inverse. Consequently, equation (12.75) is true if and only if

$$E[\{S_\beta^F - A_{\text{opt}}^F \mathcal{M}^{-1}(m^*)\}m^{*^T}] = 0$$

or when

$$A_{\text{opt}}^F E\{\mathcal{M}^{-1}(m^*)m^{*^T}\} = E(S_\beta^F m^{*^T}). \tag{12.76}$$

Using the result from equation (12.12) of Lemma 12.1, we obtain that

$$E\{S_\beta^F(Z)m^{*^T}(Z,\beta_0)\} = -\left[E\left\{\frac{\partial m^*(Z,\beta_0)}{\partial \beta^T}\right\}\right]^T.$$

Substituting this last result into equation (12.76) and solving for A_{opt}^F leads to (12.72), thus proving the corollary. □

The results of Corollary 3 are especially useful when the inverse operator $\mathcal{M}^{-1}(\cdot)$ can be derived explicitly such as the case when there are two levels of coarsening or when the coarsening is monotone. The following algorithm can now be used to derive improved adaptive double-robust estimators for β:

1. If the coarsening of the data is not by design, we develop a model for the coarsening probabilities, say

   $$P(\mathcal{C} = r|Z) = \varpi\{r, G_r(Z), \psi\},$$

   and estimate ψ by maximizing

   $$\prod_{i=1}^n \varpi\{\mathcal{C}_i, G_{\mathcal{C}_i}(Z_i), \psi\}.$$

   We denote this estimator by $\hat{\psi}_n$.
2. We posit a simpler parametric model for the distribution of Z using $p_Z^*(z,\xi)$ and estimate ξ by maximizing the likelihood

   $$\prod_{i=1}^n p^*_{G_{r_i}(Z_i)}(g_{r_i},\xi)$$

   for a realization of the data $(r_i, g_{r_i}),\ i = 1,\ldots,n$, where

   $$p^*_{G_r(Z)}(g_r,\xi) = \int_{\{z:G_r(z)=g_r\}} p_Z^*(z,\xi)d\nu_Z(z).$$

   We denote the estimator by $\hat{\xi}_n^*$.
3. We consider a $t_1 \times 1$ vector $m^*(Z,\beta)$ of estimating functions, where $t_1 > q$. These may include the q-dimensional efficient full-data estimating function as a subset of $m^*(Z,\beta)$.

4. We compute $A^F_{\text{opt}}(\beta, \hat{\psi}_n, \hat{\xi}^*_n)$ by using

$$A^F_{\text{opt}}(\beta, , \psi, \xi) = -\left[E\left\{\frac{\partial m^*(Z,\beta)}{\partial \beta^T}, \xi\right\}\right]^T \times \left(E[\mathcal{M}^{-1}\{m^*(Z,\beta), \psi, \xi\} m^{*^T}(Z,\beta), \xi]\right)^{-1},$$

where we emphasize that $\mathcal{M}^{-1}$ is computed as a function of ψ and ξ and expectations of functions of Z are computed as a function of ξ.

5. We compute

$$L_2^{t_1 \times 1}\{\mathcal{C}, G_{\mathcal{C}}(Z), \beta, \psi, \xi\} = -\Pi\left[\frac{I(\mathcal{C} = \infty) m^*(Z,\beta)}{\varpi(\infty, Z, \psi)} \middle| \Lambda_2, \psi, \xi\right].$$

6. The improved double-robust estimator for β is given as the solution to the AIPWCC estimating equation

$$\sum_{i=1}^{n} A^F_{\text{opt}}(\beta, \hat{\psi}_n, \hat{\xi}^*_n)\left[\frac{I(\mathcal{C}_i = \infty) m^*(Z_i,\beta)}{\varpi(\infty, Z_i, \hat{\psi}_n)} + L_2\{\mathcal{C}_i, G_{\mathcal{C}_i}(Z_i), \beta, \hat{\psi}_n, \hat{\xi}^*_n\}\right] = 0. \tag{12.77}$$

Since this is an AIPWCC estimator, the asymptotic variance can be estimated using the sandwich variance estimator given by (9.19).

Logistic Regression Example Revisited

In Section 10.2, we developed a double-robust estimator for the parameters in a logistic regression model when one of the covariates was missing for some individuals. Specifically, we considered the model

$$P(Y = 1|X) = \frac{\exp(\beta^T X^*)}{1 + \exp(\beta^T X^*)},$$

where $X = (X_1^T, X_2)^T$, $X^* = (1, X_1^T, X_2)^T$, (Y, X_1) were always observed, whereas the covariate X_2 may be missing for some of the individuals in a study. A complete-case indicator was denoted by R, and it was assumed that

$$P(R = 1|Y, X) = \pi(Y, X_1, \psi) = \frac{\exp(\psi_0 + \psi_1 Y + \psi_2^T X_1)}{1 + \exp(\psi_0 + \psi_1 Y + \psi_2^T X_1)}.$$

The estimator $\hat{\psi}_n$ is obtained by maximizing the likelihood

$$\prod_{i=1}^{n} \frac{\exp\{(\psi_0 + \psi_1 Y_i + \psi_2^T X_{1i}) R_i\}}{1 + \exp(\psi_0 + \psi_1 Y_i + \psi_2^T X_{1i})}.$$

We also posited a model for the full data (Y, X) by assuming that the conditional distribution of X given Y follows a multivariate normal distribution with a mean that depends on Y but with a variance matrix that is independent of Y. Let us denote the mean vector of X given $Y = 1$ and $Y = 0$ as μ_1 and μ_0, respectively, and the common covariance matrix as Σ. We also denote the mean vector of X_1 given $Y = 1$ and $Y = 0$ as μ_{11} and μ_{10}, respectively, and the mean of X_2 given $Y = 1$ and $Y = 0$ as μ_{21} and μ_{20}, respectively. Similarly, we denote the variance matrix of X_1 by Σ_{11}, the variance of the single covariate X_2 by Σ_{22}, and the covariance of X_1 and X_2 by Σ_{12}. The parameter ξ for this posited model can be represented by $\xi = (\mu_1, \mu_0, \Sigma, \tau)$, where τ denotes $P(Y = 1)$. Since Y is observed for everyone, the estimate for τ is obtained by the sample proportion

$$\hat{\tau}_n = n^{-1} \sum_{i=1}^{n} Y_i.$$

The estimates for μ_1, μ_0, and Σ are obtained by maximizing the observed-data likelihood

$$\prod_{i=1}^{n} \prod_{k=0}^{1} \Bigg[|\Sigma_{11}|^{-1/2} \exp\left\{ -\frac{1}{2}(X_{1i} - \mu_{1k})^T \Sigma^{-1} (X_{1i} - \mu_{1k}) \right\}^{(1-R_i)I(Y_i=k)}$$
$$\times |\Sigma|^{-1/2} \exp\left\{ -\frac{1}{2}(X_i - \mu_k)^T \Sigma^{-1} (X_i - \mu_k) \right\}^{R_i I(Y_i=k)} \Bigg].$$

With full data we know that the optimal estimating function is given by

$$m(Y, X, \beta) = X^* \left\{ Y - \frac{\exp(\beta^T X^*)}{1 + \exp(\beta^T X^*)} \right\}.$$

Since this may no longer be the optimal choice with coarsened data, we now consider an expanded set of estimating functions, namely $m^*(Y, X, \beta)$. For example, we might take

$$m^*(Y, X, \beta) = X^{**} \left\{ Y - \frac{\exp(\beta^T X^*)}{1 + \exp(\beta^T X^*)} \right\},$$

where X^{**} is a vector consisting of X^* together with all the squared terms and cross-product terms of X.

To find the optimal estimator using $m^*(Y, X, \beta)$ defined above, we must compute $A^F_{\text{opt}}(\beta, \psi, \xi)$ in equation (12.72), as described in step 4 of the algorithm. Toward that end, we first note that

$$-\frac{\partial m^*(Y, X, \beta)}{\partial \beta^T} = X^{**} \frac{\exp(\beta^T X^*)}{\{1 + \exp(\beta^T X^*)\}^2} X^{*T}.$$

Also, with two levels of missingness,

$$\mathcal{M}^{-1}\{m^*(Y,X,\beta),\psi,\xi\} = \{\pi(Y,X_1,\psi)\}^{-1}m^*(Y,X,\beta) - \frac{1-\pi(Y,X_1,\psi)}{\pi(Y,X_1,\psi)}E\{m^*(Y,X,\beta)|Y,X_1,\xi\}.$$

Finally, with two levels of missingness, we use the results from Theorem 10.2 to obtain

$$\begin{aligned} L_2\{\mathcal{C},G_{\mathcal{C}}(Z),\beta,\psi,\xi\} &= -\Pi\left[\frac{I(\mathcal{C}=\infty)m^*(Z,\beta)}{\varpi(\infty,Z,\psi)}\middle|\Lambda_2,\psi,\xi\right] \\ &= -\left\{\frac{R-\pi(Y,X_1,\psi)}{\pi(Y,X_1,\psi)}\right\}E\{m^*(Y,X,\beta)|Y,X_1,\xi\}. \end{aligned}$$

Remark 7. Because Y is a binary indicator and the distribution of X given Y is multivariate normal, it would be easy to simulate full data (Y,X) from such a joint distribution. Such simulated data can then be used to estimate unconditional expectations such as $E\left\{-\frac{\partial m^*(Y,X,\beta)}{\partial\beta^T},\xi\right\}$, which for this example is

$$E\left[X^{**}\frac{\exp(\beta^T X^*)}{\{1+\exp(\beta^T X^*)\}^2}X^{*^T},\xi\right],$$

using Monte Carlo methods. Similarly, because the conditional distribution of X_2 given X_1, Y is normally distributed with mean

$$\mu_{21}Y+\mu_{20}(1-Y)+\Sigma_{12}\Sigma_{22}^{-1}\{X_1-\mu_{11}Y-\mu_{10}(1-Y)\}$$

and variance $\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{12}^T$, Monte Carlo methods can be used when computing conditional expectations of functions of (Y,X) given (Y,X_1), as was necessary to compute $\mathcal{M}^{-1}$ or $L_2(\cdot)$. □

We are finally in a position to define all the elements making up equation (12.77), which we can use to derive the estimator for β.

12.5 Recap and Review of Notation

- In this chapter, we considered a restricted class of AIPWCC estimators where observed-data estimating functions were chosen from the linear subspace $\Pi[\Xi|\Lambda_\psi^\perp] \subset \Lambda^\perp$, where

$$\Xi = \left\{\frac{I(\mathcal{C}=\infty)\mathcal{G}^F}{\varpi(\infty,Z)} \oplus \mathcal{G}_2\right\},$$

  where $\mathcal{G}^F$ was a linear subspace contained in $\Lambda^{F\perp}$, and $\mathcal{G}_2$ was a linear subspace contained in the augmentation space Λ_2.
- We considered two classes of restricted estimators:

  - (Class 1) were defined by letting both $\mathcal{G}^F \subset \Lambda^{F\perp}$ and $\mathcal{G}_2 \subset \Lambda_2$ be finite-dimensional linear spaces spanned by $J^{F^{t_1\times 1}}(Z)$ and $J_2^{t_2\times 1}\{\mathcal{C}, G_{\mathcal{C}}(Z)\}$, respectively, where $t_1 > q$ and where the elements of the coarsening model score vector $S_\psi\{\mathcal{C}, G_{\mathcal{C}}(Z)\}$ were elements contained in $J_2^{t_2\times 1}\{\mathcal{C}, G_{\mathcal{C}}(Z)\}$.
  - (Class 2) were defined by only letting $\mathcal{G}^F \subset \Lambda^{F\perp}$ be a finite-dimensional linear space spanned by $J^{F^{t_1\times 1}}(Z)$, with $t_1 > q$, but took $\mathcal{G}_2 = \Lambda_2$.
- The optimal influence functions within the class Ξ were derived for both classes and shown to be orthogonal to Λ_ψ. This then allowed us to derive optimal restricted AIPWCC estimators within these two classes.
  - The (class 1) restricted optimal estimators were the easiest to compute but not double robust.
  - The (class 2) restricted optimal estimators resulted in double-robust estimators. They were, however, computationally more intensive but not as difficult to compute as the locally efficient estimators of Chapter 11.

12.6 Exercises for Chapter 12

1. In Section 11.2, we outlined the steps that would be necessary to obtain a locally efficient estimator for β for the restricted moment model

$$E(Y|X) = \mu(X, \beta)$$

when the data are monotonically coarsened. This methodology led to the integral equation (11.42), which in general is very difficult if not impossible to solve. Only consider the case when Y is a univariate random variable. For this same problem, outline the steps that would be necessary to obtain the optimal restricted (class 2) estimator for β. For this exercise, take

$$m^*(Y, X, \beta) = f^{t_1\times 1}(X, \beta)\{Y - \mu(X, \beta)\},$$

where $f^{t_1\times 1}(X, \beta)$ is a $t_1 \times 1$ vector of linearly independent functions of X and β and $t_1 > q$. Assume that you can estimate the parameter ψ in the coarsening model and the parameter ξ in the posited model $p_Z^*(z, \xi)$.

13

Double-Robust Estimator of the Average Causal Treatment Effect

Statistical inference generally focuses on the associational relationships between variables in a population. Data that are collected are assumed to be realizations of iid random vectors $Z_1, \ldots, Z_n$ where a single observation Z is distributed according to some density in the model $p_Z(z, \theta)$, where θ denotes parameters that describe important features of the relationships of the variables of interest.

However, one may be interested in causal relationships. That is, does "A" cause "B"? For example, does a treatment intervention or exposure at one point in time have a causal effect on subsequent response? In order to formulate such a question from a statistical perspective, we will take the point of view advocated by Neyman (1923), Rubin (1974), Robins (1986), and Holland (1986), who considered potential outcomes. We will illustrate that the semiparametric theory developed in this book can be used to aid us in finding efficient semiparametric estimators of the average causal treatment effect under certain assumptions. The methods we will discuss only consider the simplest case of point exposure; that is, exposure or treatment of individuals at only one point in time. A much more complex and elegant theory has been developed by Robins and colleagues that provides methods for studying the effect of time-dependent treatments on response. We refer the reader to the book by van der Laan and Robins (2003) for more details and references.

13.1 Point Exposure Studies

We shall denote the possible treatments or exposures that can be given or experienced by an individual by the random variable A. For simplicity, we will assume that there are two possible treatments that we wish to compare and thus A will be a binary variable taking on the values 0 or 1. For example, A may denote whether an individual with hypertension is treated with a statin drug ($A = 1$) or not ($A = 0$). The response variable will be denoted by Y, say, change in blood pressure after three months. Hence, in a typical study of

this treatment, we consider a population of patients with hypertension, say individuals whose diastolic blood pressure is greater than or equal to 140, and identify a sample of such patients, some of which receive the statin drug ($A = 1$) and others who do not ($A = 0$). This sample of individuals is followed for three months, and the change in blood pressure Y is measured.

Ultimately, we are interested in establishing a causal link between treatment and response. That is, does treatment with the statin drug reduce blood pressure after three months as compared with no treatment? The data that are available from such a study may be summarized by $Z_i = (Y_i, A_i, X_i), i = 1, \ldots, n$, where for the i-th individual Y_i denotes the response, A_i the treatment received, and X_i other covariates that have been measured prior to treatment (i.e., baseline covariates).

In a typical associational analysis, we might define population parameters $\mu_1 = E(Y|A = 1)$, $\mu_0 = E(Y|A = 0)$, and $\Delta = \mu_1 - \mu_0$. That is, Δ denotes the difference in mean response for individuals receiving treatment 1 and the mean response for individuals receiving treatment 0. Without any additional assumptions, we can estimate Δ simply as the difference of the treatment-specific sample average of response, namely $\hat{\Delta} = \hat{\mu}_1 - \hat{\mu}_0$, where

$$\hat{\mu}_1 = n_1^{-1} \sum_{i=1}^{n} A_i Y_i, \quad \hat{\mu}_0 = n_0^{-1} \sum_{i=1}^{n} (1 - A_i) Y_i,$$

and $n_1 = \sum A_i$, $n_0 = \sum (1 - A_i)$ denote the treatment-specific sample sizes.

Typically, such an associational analysis does not answer the causal question of interest. If the treatments were not assigned to the patients at random, then one can easily imagine that individuals who receive the statin drugs may be inherently different from those who do not. They may be wealthier, younger, smoke less, etc. Consequently, the associational parameter $\Delta = \mu_1 - \mu_0$ may reflect these inherent differences as well as any effect due to treatment. In the study of epidemiology, such factors are referred to as confounders, as they may confound the relationship between treatment and response.

Thus, we have argued that statistical associations may not be adequate to describe causal effects. Therefore, how might we describe causal effects? The point of view we will adopt is that proposed by Neyman (1923) and Rubin (1974), where causal effects are defined through potential outcomes or counterfactual random variables. Specifically, for each level of the treatment $A = a$, we will assume that there exists a potential outcome $Y^*(a)$, where $Y^*(a)$ denotes the response of a randomly selected individual had that individual, possibly contrary to fact, been given treatment $A = a$. In our illustration, we only include two treatments and hence we define the potential outcomes $Y^*(1)$ and $Y^*(0)$. Again, we emphasize that these are referred to as potential outcomes or counterfactual random variables, as it is impossible to observe both $Y^*(1)$ and $Y^*(0)$ simultaneously. Nonetheless, using the notion of potential outcomes, we would define the causal treatment effect by $Y^*(1) - Y^*(0)$.

This definition of causal treatment effect is at the subject-specific level and, as we pointed out, is impossible to measure. However, it may be possible, under certain assumptions, to estimate the population-level causal treatment effect, i.e., the expected value of the subject-specific treatment effect denoted as

$$\delta = E\{Y^*(1) - Y^*(0)\} = E\{Y^*(1)\} - E\{Y^*(0)\}.$$

The parameter δ is referred to as the average causal treatment effect.

Since the average causal treatment effect is defined from parameters describing the potential outcomes, which are not directly observable, the question is whether this parameter can be deduced from parameters describing the distribution of the observable random variables $Z = (Y, A, X)$.

Using the ideas that were developed for missing-data problems, we consider the full data to be the variables $\{Y^*(1), Y^*(0), A, X\}$, which involve the potential outcomes as well as the treatment assignment and baseline covariates, and the observed data to be (Y, A, X). We now make the reasonable assumption that

$$Y = AY^*(1) + (1 - A)Y^*(0); \tag{13.1}$$

that is, the observed response Y is equal to $Y^*(1)$ if the subject was given treatment $A = 1$ and is equal to $Y^*(0)$ if the subject was given treatment $A = 0$.

Remark 1. Rubin (1978a) refers to the assumption (13.1) as the Stable Unit Treatment Value Assumption, or SUTVA. Although this assumption may seem straightforward at first, there are some philosophical subtleties that need to be considered in order to fully accept. For one thing, there must not be any interference in the response from other subjects. That is, the observed response for the i-th individual in the sample should not be affected by the response of the other individuals in the sample. Thus, for example, this assumption may not be reasonable in a vaccine intervention trial for an infectious disease, where the response of an individual is clearly affected by the response of others in the study. That is, whether or not an individual contracts an infectious disease will depend, to some extent, on whether and how many other individuals in the population are infected. From here on, we will assume the SUTVA assumption but caution that the plausibility of this assumption needs to be evaluated on a case-by-case basis. □

Because of assumption (13.1), we see that the observed data are a many-to-one transformation of the full data. We also note that the treatment assignment indicator A plays a role similar to that of the missingness indicator in missing-data problems. This analogy to missing-data problems will be useful as we develop the theory that enables us to estimate the average causal treatment effect.

13.2 Randomization and Causality

Intuitively, it has been accepted that the use of a randomized intervention study will result in an unbiased estimate of the average treatment effect with causal interpretations. This is because patients are assigned to treatment interventions according to a random mechanism that is independent of all other factors. Therefore, individuals in the two treatment groups are similar, on average, with respect to all characteristics except for the treatment intervention to which they were assigned. Consequently, differences in response between the two randomized groups can be reasonably attributed to the effect of treatment and not other extraneous factors.

We now formalize this notion through the use of potential outcomes. Specifically, we will show that the observed treatment difference in a randomized intervention study is an unbiased estimator of the average causal treatment effect δ.

Together with the SUTVA assumption (13.1), we also make the assumption that

$$A \perp\!\!\!\perp \{Y^*(1), Y^*(0)\}, \tag{13.2}$$

where "$\perp\!\!\!\perp$" denotes independence. This assumption is plausible since the response of an individual to one treatment or the other should be independent of treatment assignment because treatment was assigned according to some random mechanism.

Remark 2. The assumption that treatment assignment is independent of the potential outcomes should not be confused with treatment assignment being independent of the observed response. That is,

$$A \perp\!\!\!\perp \{Y^*(1), Y^*(0)\} \text{ does not imply that } A \perp\!\!\!\perp Y = AY^*(1) + (1-A)Y^*(0).$$

Generally, we reserve the notion that $A \perp\!\!\!\perp Y$ to denote the null hypothesis that there is no treatment effect, whereas if there is a treatment effect, then A is not independent of Y. □

In a randomized study, the associational treatment effect $\Delta = E(Y|A = 1) - E(Y|A = 0)$ is equal to the average causal treatment effect $\delta = E\{Y^*(1)\} - E\{Y^*(0)\}$, as we now demonstrate:

$$E(Y|A=1) = E\{AY^*(1) + (1-A)Y^*(0)|A=1\} = E\{Y^*(1)|A=1\} \tag{13.3}$$

$$= E\{Y^*(1)\}, \tag{13.4}$$

where (13.3) follows from the SUTVA assumption and (13.4) follows from assumption (13.2). Similarly, we can show that $E(Y|A = 0) = E\{Y^*(0)\}$. Consequently, the difference in the sample's average response between treatments $n_1^{-1}\sum_{i=1}^n A_i Y_i - n_0^{-1}\sum_{i=1}^n (1-A_i)Y_i$, which is an unbiased estimator

for Δ, the associational treatment effect, is also an unbiased estimator for the average causal treatment effect δ in a randomized intervention study.

As we mentioned earlier, the treatment indicator A serves a role analogous to the missingness indicator R, which was used to denote missing data with two levels of missingness. That is, when $A = 1$, we observe $Y^*(1)$, in which case $Y^*(0)$ is missing, and when $A = 0$, we observe $Y^*(0)$ and $Y^*(1)$ is missing. The assumption (13.2), which is induced because of randomization, is similar to missing completely at random; that is, the probability that A is equal to 1 or 0 is independent of all the data $\{Y^*(1), Y^*(0), X\}$.

13.3 Observational Studies

In an observational study, individuals are not assigned to treatment by an experimental design but rather by choice. It may be that data from an observational study are easier or cheaper to collect, or it may be that conducting a randomized study is infeasible or unethical. Clearly, if we want to evaluate the effect of smoking on some health outcome, it would be unethical to randomize individuals and force them to smoke or not smoke.

For simplicity, and without loss of generality, we again consider only two treatments. In an observational study, individuals who receive treatment $A = 1$ may not be prognostically comparable with those who receive treatment $A = 0$. Therefore, it may no longer be reasonable to assume that $A \perp\!\!\!\perp \{Y^*(1), Y^*(0)\}$. However, if pretreatment (baseline) prognostic factors X can be identified that affect treatment choice and, in addition, are themselves prognostic (i.e., related to clinical outcome), then it may be reasonable to assume

$$A \perp\!\!\!\perp \{Y^*(1), Y^*(0)\}|X; \tag{13.5}$$

that is, treatment assignment is independent of the potential outcomes given X. Such variables X are referred to in epidemiology as confounders, and assumption (13.5) is sometimes referred to as the assumption of "no unmeasured confounders." Rubin (1978a) also refers to this assumption as the strong ignorability assumption.

Remark 3. The assumption of no unmeasured confounders is key to being able to estimate the average causal treatment effect in an observational study. Presumably, when a patient or his or her physician is faced with a binary treatment choice of whether to treat the patient with treatment $A = 0$ or $A = 1$, they do not know what the patient's potential outcome will be. Consequently, treatment choice is made based on variables and characteristics of the patient prior to or at the time of treatment. If such factors are captured in the database, then the assumption of no unmeasured confounders may be reasonable. Of course, there may be factors that influence treatment decisions that are not captured in the data that are collected. If such factors are also correlated with response, then the assumption (13.5) is no longer tenable. □

We now give an argument to show that the average causal treatment effect can be identified through the distribution of the observable data (Y, A, X) if assumptions (13.1) and (13.5) hold. This follows because

$$\begin{aligned} E\{Y^*(1)\} &= E_X[E\{Y^*(1)|X\}] \\ &= E_X[E\{Y^*(1)|A=1,X\}] \qquad (13.6) \\ &= E_X\{E(Y|A=1,X)\}, \qquad (13.7) \end{aligned}$$

where (13.6) follows from the assumption of no unmeasured confounders and (13.7) follows from the SUTVA assumption. Similarly, we can show that $E\{Y^*(0)\} = E_X\{E(Y|A=0,X)\}$. Hence the average causal treatment effect is equal to

$$\delta = E_X\{E(Y|A=1,X) - E(Y|A=0,X)\}, \qquad (13.8)$$

which only involves the distribution of (Y, A, X).

Remark 4. It is important to note that the outer expectation of (13.6) and (13.7) is with respect to the marginal distribution of X and not the conditional distribution of X given $A = 1$. We must also be a little careful to make sure that, in equations (13.6) and (13.7), we are not conditioning on a null event. One way to ensure that we are not is to assume that both $P(A = 1|X = x)$ and $P(A = 0|X = x)$ are bounded away from zero for all x in the support of X. We will discuss this assumption in greater detail when we introduce the propensity score in the next section. □

13.4 Estimating the Average Causal Treatment Effect

Regression Modeling

We now consider the estimation of the average causal treatment effect from a sample of observed data (Y_i, A_i, X_i), $i = 1, \ldots, n$. The first approach, which we refer to as regression modeling, is motivated by equation (13.8). Here, we consider a restricted moment model for the conditional expectation of Y given (A, X) in terms of a finite-dimensional parameter, say ξ. That is,

$$E(Y|A, X) = \mu(A, X, \xi). \qquad (13.9)$$

The regression model could be as complicated as is deemed necessary by the data analyst to get a good fit. For example, we might consider the linear model with an interaction term; that is,

$$\mu(A, X, \xi) = \xi_0 + \xi_1 A + \xi_2 X + \xi_3 AX,$$

or, if the response variable Y is positive, we might consider the corresponding log-linear model

$$\mu(A, X, \xi) = \exp(\xi_0 + \xi_1 A + \xi_2 X + \xi_3 AX). \tag{13.10}$$

The parameter ξ can be estimated using the generalized estimating equations developed in Section 4.6. For example, we may solve the estimating equation

$$\sum_{i=1}^{n} \frac{\partial \mu(A_i, X_i, \xi)}{\partial \xi} V^{-1}(A_i, X_i)\{Y_i - \mu(A_i, X_i, \xi)\} = 0, \tag{13.11}$$

where $V(A_i, X_i) = \text{var}(Y_i|A_i, X_i)$, to obtain the estimator $\hat{\xi}_n$.

If the conditional expectations $E(Y|A = 1, X)$ and $E(Y|A = 0, X)$ were known, then a natural consistent and unbiased estimator of the average causal treatment effect δ, given by (13.8), would be obtained using the empirical average

$$n^{-1} \sum_{i=1}^{n} \{E(Y|A = 1, X_i) - E(Y|A = 0, X_i)\}. \tag{13.12}$$

Under the assumption that the restricted moment model is correct, a natural estimator for $E(Y|A = 1, X) - E(Y|A = 0, X)$ would then be $\mu(1, X, \hat{\xi}_n) - \mu(0, X, \hat{\xi}_n)$. Substituting this into (13.12) yields the estimator for the average causal treatment effect,

$$\hat{\delta}_n = n^{-1} \sum_{i=1}^{n} \{\mu(1, X_i, \hat{\xi}_n) - \mu(0, X_i, \hat{\xi}_n)\}. \tag{13.13}$$

So, for example, if we posited the log-linear model (13.10), then the estimator for the average causal treatment effect would be given by

$$\hat{\delta}_n = n^{-1} \sum_{i=1}^{n} \left[\exp\{\hat{\xi}_{0n} + \hat{\xi}_{1n} + (\hat{\xi}_{2n} + \hat{\xi}_{3n})X_i\} - \exp(\hat{\xi}_{0n} + \hat{\xi}_{2n} X_i) \right].$$

The consistency and asymptotic normality of the estimator $\hat{\delta}_n$ can be obtained in a straightforward fashion by deriving its influence function. We leave this as an exercise for the reader to derive. We do note, however, that these asymptotic properties are based on the assumption that the restricted moment model $E(Y|A, X) = \mu(A, X, \xi)$ is correctly specified.

13.5 Coarsened-Data Semiparametric Estimators

We now consider how to obtain estimators for the average causal treatment effect by casting the problem as a coarsened-data semiparametric model. Toward that end, we denote the full data to be $\{Y^*(1), Y^*(0), X, A\}$, where $Y^*(1)$ and $Y^*(0)$ denote the potential outcomes for treatment 1 and treatment 0, respectively, A denotes the treatment assignment, and X denotes the

vector of baseline covariates. The joint density of the full data can be written as

$$\begin{aligned} p\{y^*(1), y^*(0), x, a\} &= p\{a|y^*(1), y^*(0), x\} p\{y^*(1), y^*(0), x\} \\ &= p(a|x) p\{y^*(1), y^*(0), x\}, \end{aligned} \tag{13.14}$$

where (13.14) follows from the strong ignorability assumption (13.5). We will put no restrictions on the joint density $p\{y^*(1), y^*(0), x\}$ of $\{Y^*(1), Y^*(0), X\}$ (i.e., a nonparametric model). Consequently, using the same logic as in Section 5.3, we argue that there is only one full-data influence function of RAL estimators for $\delta = E\{Y^*(1) - Y^*(0)\}$. Letting δ_0 denote the true value of δ, the full-data influence function is given by

$$\varphi^F\{Y^*(1), Y^*(0), X\} = \{Y^*(1) - Y^*(0) - \delta_0\}, \tag{13.15}$$

which, of course, is the influence function for the full-data estimator

$$\hat{\delta}_n^F = n^{-1} \sum_{i=1}^{n} \{Y_i^*(1) - Y_i^*(0)\}.$$

Although the joint distribution of $\{Y^*(1), Y^*(0), X\}$ can be arbitrary, we will assume that $P(A = 1|X)$ can be modeled as $\pi(X, \psi)$ using a finite number of parameters ψ. Because treatment assignment A is a binary indicator, the conditional density of A given X is

$$p(a|x) = \pi(x, \psi)^a \{1 - \pi(x, \psi)\}^{1-a}.$$

The function $P(A = 1|X)$ is defined as the propensity score, as it reflects the propensity that an individual will receive one treatment or the other as a function of the baseline covariates X. The propensity score was first introduced by Rosenbaum and Rubin (1983), and the properties have been studied extensively by Rosenbaum and Rubin; see, for example, Rosenbaum and Rubin (1984, 1985), Rosenbaum (1984, 1987), and Rubin (1997). Although the model for the propensity score is not used in defining the full-data estimator for the average causal treatment effect δ, it will play a crucial role when we consider observed-data estimators for δ.

In contrast with the full data $\{Y^*(1), Y^*(0), X, A\}$, the observed data are given as the many-to-one transformation of the full data, namely (Y, X, A), where $Y = AY^*(1) + (1 - A)Y^*(0)$. As such, this is an example of coarsened data, where A plays a role similar to the coarsening variable $\mathcal{C}$, or, more specifically, to the complete-case indicator R defined in the previous chapters.

Remark 5. Throughout the chapters on missing and coarsened data, we always assumed that, with positive probability, the coarsening variable $\mathcal{C}$ could take on the value ∞ to denote the case when the full data were observed. For this problem, we never get to observe the full data $\{Y^*(1), Y^*(0), X\}$. We either observe $Y^*(1)$ when $A = 1$ or $Y^*(0)$ when $A = 0$. Nonetheless, we will see a similarity in the semiparametric estimators developed for this problem and those for the missing-data problems developed previously. □

Observed-Data Influence Functions

As with all semiparametric models, the key to finding estimators is to derive the nuisance tangent space and the space orthogonal to the nuisance tangent space, which, in turn, are used to derive the space of influence functions.

We first consider the case when the propensity score $P(A = 1|X)$ is known to us. This may be the case, for example, if we designed a randomized study where treatment was assigned at random with known probabilities that could depend on pretreatment baseline covariates. Using arguments identical to those in Chapters 7 and 8, we can show that the space of observed-data influence functions of RAL estimators for δ corresponds to the space $\mathcal{K}^{-1}\{(IF)^F\}$, where $(IF)^F$ denotes the space of full-data influence functions and $\mathcal{K} : \mathcal{H} \rightarrow \mathcal{H}^F$, given by Definition 1 of Chapter 7, is the many-to-one linear mapping from the observed-data Hilbert space to the full-data Hilbert space, where, for any $h \in \mathcal{H}$, $\mathcal{K}(h) = E\{h(Y, A, X)|Y^*(1), Y^*(0), X\}$. Consequently, $\mathcal{K}^{-1}\{(IF)^F\}$ denotes all the functions $\varphi(Y, A, X)$ such that

$$E\{\varphi(Y, A, X)|Y^*(1), Y^*(0), X\} = \varphi^F\{Y^*(1), Y^*(0), X\}$$

for any $\varphi^F\{Y^*(1), Y^*(0), X\}$ corresponding to a full-data influence function.

In our problem, there is only one full-data influence function given by (13.15). Hence, the class of observed-data influence functions corresponds to functions $\mathcal{K}^{-1}\{Y^*(1) - Y^*(0) - \delta_0\}$, which, by Lemma 7.4, are equal to

$$h(Y, A, X) + \Lambda_2,$$

where $h(Y, A, X)$ is any function that satisfies the relationship

$$E\{h(Y, A, X)|Y^*(1), Y^*(0), X\} = \{Y^*(1) - Y^*(0) - \delta_0\}, \tag{13.16}$$

and Λ_2 is the linear subspace in $\mathcal{H}$ consisting of elements $L_2(Y, A, X)$ such that

$$E\{L_2(Y, A, X)|Y^*(1), Y^*(0), X\} = 0, \tag{13.17}$$

also referred to as the augmentation space.

A function $h(\cdot)$ satisfying equation (13.16) is motivated by the inverse propensity weighted full-data influence function; namely,

$$h(Y, A, X) = \left\{ \frac{AY}{\pi(X)} - \frac{(1 - A)Y}{1 - \pi(X)} - \delta_0 \right\}. \tag{13.18}$$

To verify this, consider the conditional expectation of the first term on the right-hand side of (13.18); that is,

$$E\left\{\frac{AY}{\pi(X)}\middle|Y^*(1),Y^*(0),X\right\} = E\left\{\frac{AY^*(1)}{\pi(X)}\middle|Y^*(1),Y^*(0),X\right\}$$
$$= \frac{Y^*(1)}{\pi(X)}E\{A|Y^*(1),Y^*(0),X\}$$
$$= \frac{Y^*(1)}{\pi(X)}E(A|X) \tag{13.19}$$
$$= \frac{Y^*(1)}{\pi(X)}\pi(X) = Y^*(1), \tag{13.20}$$

where (13.19) follows from the strong ignorability assumption. Also, in order for (13.20) to hold, we must not be dividing 0 by 0. Therefore, we will need the additional assumption that the propensity score $P(A = 1|X) = \pi(X)$ is strictly greater than zero almost everywhere. Similarly, we can show that $E\left\{\frac{(1-A)Y}{1-\pi(X)}|Y^*(1),Y^*(0),X\right\} = Y^*(0)$ as long as $1 - \pi(X)$ is strictly greater than zero almost everywhere. Therefore, we have proved that the function $h(\cdot)$ defined in (13.18) satisfies the relationship (13.16) as long as the propensity score $0 < \pi(x) < 1$, for all x in the support of X.

To derive the augmentation space Λ_2, we must find all functions $L_2(\cdot)$ of Y, A, X that satisfy (13.17). Because A is a binary indicator, any function of Y, A, X can be written as

$$L_2(Y,A,X) = AL_{21}(Y,X) + (1-A)L_{20}(Y,X) \tag{13.21}$$

for arbitrary functions $L_{21}(\cdot)$ and $L_{20}(\cdot)$ of Y, X. Hence the conditional expectation of $L_2(\cdot)$, given $\{Y^*(1), Y^*(0), X\}$, is

$$E\{L_2(Y,A,X)|Y^*(1),Y^*(0),X\}$$
$$= E[AL_{21}\{Y^*(1),X\} + (1-A)L_{20}\{Y^*(0),X\}|Y^*(1),Y^*(0),X] \tag{13.22}$$
$$= \pi(X)L_{21}\{Y^*(1),X\} + \{1-\pi(X)\}L_{20}\{Y^*(0),X\}, \tag{13.23}$$

where (13.22) follows from the SUTVA assumption and (13.23) follows from the strong ignorability assumption. Therefore, in order for (13.17) to hold, we need

$$L_{20}\{Y^*(0),X\} = -\left\{\frac{\pi(X)}{1-\pi(X)}\right\}L_{21}\{Y^*(1),X\}. \tag{13.24}$$

Since both the left- and right-hand sides of (13.24) denote functions of $Y^*(1), Y^*(0), X$, the only way that equation (13.24) can hold is if both $L_{20}(\cdot)$ and $L_{21}(\cdot)$ are functions of X alone. In that case, any element of Λ_2 must satisfy

$$L_{20}(X) = -\left\{\frac{\pi(X)}{1-\pi(X)}\right\}L_{21}(X). \tag{13.25}$$

Substituting the relationship given by (13.25) into (13.21), we conclude that the space Λ_2 consists of functions

$$\left\{\frac{A-\pi(X)}{1-\pi(X)}\right\} L_{21}(X) \text{ for arbitrary functions } L_{21}(X).$$

Since $1-\pi(X)$ is strictly greater than zero almost everywhere and $L_{21}(X)$ is an arbitrary function of X, we can write Λ_2 as

$$\Lambda_2 = \left\{ \{A-\pi(X)\}h_2(X) \text{ for arbitrary } h_2(X) \right\}. \tag{13.26}$$

Thus, we have demonstrated that the class of all influence functions of RAL estimators for δ, when the propensity score $\pi(X)$ is known, is given by

$$\left\{\frac{AY}{\pi(X)} - \frac{(1-A)Y}{1-\pi(X)} - \delta_0\right\} + \Lambda_2, \tag{13.27}$$

where Λ_2 is defined by (13.26). The optimal influence function in this class is the one with the smallest variance or, equivalently, the element with the smallest norm. This is obtained by choosing the element $\{A-\pi(X)\}h_2^0(X)$ to be minus the projection of $\left\{\frac{AY}{\pi(X)} - \frac{(1-A)Y}{1-\pi(X)} - \delta_0\right\}$ onto Λ_2.

We remind the reader that the projection of some arbitrary element $\varphi(Y,A,X)$ onto Λ_2 is obtained by finding the unique element $h_2^0(X)$ such that

$$E\left(\left[\varphi(Y,A,X) - \{A-\pi(X)\}h_2^0(X)\right]\{A-\pi(X)\}h_2(X)\right) = 0$$

for all functions $h_2(X)$. We denote this as $\Pi[\varphi(Y,A,X)|\Lambda_2] = \{A-\pi(X)\}h_2^0(X)$. As indicated earlier, any function $\varphi(\cdot)$ of Y,A,X can be written as $\varphi(Y,A,X) = A\varphi_1(Y,X) + (1-A)\varphi_0(Y,X)$. Because projections are linear operators,

$$\Pi[\varphi(Y,A,X)|\Lambda_2] = \Pi[A\varphi_1(Y,X)|\Lambda_2] + \Pi[(1-A)\varphi_0(Y,X)|\Lambda_2].$$

We now show how to derive these projections in the following theorem.

Theorem 13.1. The projection

$$\Pi[A\varphi_1(Y,X)|\Lambda_2] = \{A-\pi(X)\}h_1^0(X), \tag{13.28}$$

where $h_1^0(X) = E\{\varphi_1(Y,X)|A=1,X\}$, and

$$\Pi[(1-A)\varphi_0(Y,X)|\Lambda_2] = \{A-\pi(X)\}h_0^0(X), \tag{13.29}$$

where $h_0^0(X) = -E\{\varphi_0(Y,X)|A=0,X\}$.

Proof. We first consider (13.28). By definition, $\Pi[A\varphi_1(Y,X)|\Lambda_2]$ is defined as the function $\{A-\pi(X)\}h_1^0(X)$ such that

$$E\left(\left[A\varphi_1(Y,X) - \{A-\pi(X)\}h_1^0(X)\right]\{A-\pi(X)\}h(X)\right) = 0,$$

or equivalently

$$E\Big[A\{A-\pi(X)\}h(X)\varphi_1(Y,X)-\{A-\pi(X)\}^2h_1^0(X)h(X)\Big]=0, \quad (13.30)$$

for all $h(X)$. By a simple conditioning argument, where we first condition on X, we can show that the second term on the left-hand side of (13.30) is equal to

$$E\Big[\{A-\pi(X)\}^2h_1^0(X)h(X)\Big]=E\Big[\pi(X)\{1-\pi(X)\}h_1^0(X)h(X)\Big]. \quad (13.31)$$

The first term on the left-hand side of (13.30) is also computed through a series of iterated conditional expectations; namely,

$$\begin{aligned}
&E\Big[A\{A-\pi(X)\}h(X)\varphi_1(Y,X)\Big]\\
&=E\bigg(E\Big[A\{A-\pi(X)\}h(X)\varphi_1(Y,X)\Big|A,X\Big]\bigg)\\
&=E\Big[A\{A-\pi(X)\}h(X)E\{\varphi_1(Y,X)|A=1,X\}\Big]\\
&=E\bigg(E\Big[A\{A-\pi(X)\}h(X)E\{\varphi_1(Y,X)|A=1,X\}\Big|X\Big]\bigg)\\
&=E\Big[\pi(X)\{1-\pi(X)\}h(X)E\{\varphi_1(Y,X)|A=1,X\}\Big]. \qquad (13.32)
\end{aligned}$$

Substituting (13.31) and (13.32) into (13.30), we obtain that

$$E\bigg(\pi(X)\{1-\pi(X)\}\Big[E\{\varphi_1(Y,X)|A=1,X\}-h_1^0(X)\Big]h(X)\bigg)=0 \quad (13.33)$$

for all $h(X)$. Since $\pi(X)$ and $1-\pi(X)$ are both bounded away from zero almost surely, and because $E\{\varphi_1(Y,X)|A=1,X\}-h_1^0(X)$ is a function of X, then (13.33) can only be true for all $h(X)$ if and only if $E\{\varphi_1(Y,X)|A=1,X\}-h_1^0(X)$ is identically equal to zero; i.e., $h_1^0(X)$ must equal $E\{\varphi_1(Y,X)|A=1,X\}$, thus proving (13.28). The proof of (13.29) follows similarly. □

Returning to the class of influence functions of RAL estimators for δ given by (13.27), the efficient influence function in this class is given by

$$\bigg\{\frac{AY}{\pi(X)}-\frac{(1-A)Y}{1-\pi(X)}-\delta_0\bigg\}-\Pi\bigg[\frac{AY}{\pi(X)}-\frac{(1-A)Y}{1-\pi(X)}-\delta_0\Big|\Lambda_2\bigg],$$

which by Theorem 13.1 is equal to

$$\begin{aligned}
&\bigg\{\frac{AY}{\pi(X)}-\frac{\{A-\pi(X)\}E(Y|A=1,X)}{\pi(X)}\\
&-\frac{(1-A)Y}{1-\pi(X)}-\frac{\{A-\pi(X)\}E(Y|A=0,X)}{1-\pi(X)}-\delta_0\bigg\}. \qquad (13.34)
\end{aligned}$$

Motivated by the efficient influence function (13.34), an estimator for δ would be obtained as the solution to the estimating equation

$$\sum_{i=1}^{n}\left\{\frac{A_iY_i}{\pi(X_i)}-\frac{\{A_i-\pi(X_i)\}\mu(1,X_i)}{\pi(X_i)}\right.$$
$$\left.-\frac{(1-A_i)Y_i}{1-\pi(X_i)}-\frac{\{A_i-\pi(X_i)\}\mu(0,X_i)}{1-\pi(X_i)}-\delta\right\}=0,$$

or, more specifically,

$$\hat{\delta}_n = n^{-1}\sum_{i=1}^{n}\left\{\frac{A_iY_i}{\pi(X_i)}-\frac{\{A_i-\pi(X_i)\}\mu(1,X_i)}{\pi(X_i)}\right.$$
$$\left.-\frac{(1-A_i)Y_i}{1-\pi(X_i)}-\frac{\{A_i-\pi(X_i)\}\mu(0,X_i)}{1-\pi(X_i)}\right\}, \tag{13.35}$$

where $\mu(A,X)=E(Y|A,X)$. Of course, $\mu(A,X)$ is not known to us but can be estimated by positing a model where $E(Y|A,X)=\mu(A,X,\xi)$ as we did in (13.9). The parameter ξ can be estimated by solving the estimating equation (13.11), and then $\mu(1,X_i,\hat{\xi}_n)$ and $\mu(0,X_i,\hat{\xi}_n)$ can be substituted into (13.35) to obtain a locally efficient estimator for δ.

The development above assumed that the propensity score $\pi(X)$ was known to us. In observational studies, this will not be the case. Consequently, we must posit a model for the propensity score, say assuming that

$$P(A=1|X)=\pi(X,\psi). \tag{13.36}$$

Since A is binary, the logistic regression model is often used; i.e.,

$$\pi(X,\psi)=\frac{\exp(\psi_0+\psi_1^TX)}{1+\exp(\psi_0+\psi_1^TX)}.$$

The estimator for ψ would be obtained using maximum likelihood; i.e., $\hat{\psi}_n$ is the value of ψ that maximizes the likelihood

$$\prod_{i=1}^{n}\pi(X_i,\psi)^{A_i}\{1-\pi(X_i,\psi)\}^{(1-A_i)}.$$

Consequently, in an observational study, the locally efficient semiparametric estimator for the average causal treatment effect would be given by

$$\hat{\delta}_n = n^{-1}\sum_{i=1}^{n}\left\{\frac{A_iY_i}{\pi(X_i,\hat{\psi}_n)}-\frac{\{A_i-\pi(X_i,\hat{\psi}_n)\}\mu(1,X_i,\hat{\xi}_n)}{\pi(X_i,\hat{\psi}_n)}\right.$$
$$\left.-\frac{(1-A_i)Y_i}{1-\pi(X_i,\hat{\psi}_n)}-\frac{\{A_i-\pi(X_i,\hat{\psi}_n)\}\mu(0,X_i,\hat{\xi}_n)}{1-\pi(X_i,\hat{\psi}_n)}\right\}. \tag{13.37}$$

Double Robustness

We complete this chapter by demonstrating that the locally efficient semiparametric estimator for δ, given by (13.37), is doubly robust. That is, under the SUTVA and strong ignorability assumptions, (13.37) is a consistent estimator for δ if either the model for the propensity score $\pi(X, \psi)$ or the regression model $E(Y|A, X) = \mu(A, X, \xi)$ is correct. We use the conventions that $\hat{\psi}_n \xrightarrow{p} \psi^*$ and $\hat{\xi}_n \xrightarrow{p} \xi^*$ to denote that, under suitable regularity conditions, these estimators will converge whether the corresponding model is correct or not and denote, by letting $\psi^* = \psi_0$ or $\xi^* = \xi_0$, the case when these estimators converge to the truth; i.e., that the corresponding model is correctly specified.

Because the estimator (13.37) is a sample average, it is easy to show that it will converge in probability to

$$E\Bigg\{\frac{AY}{\pi(X,\psi^*)} - \frac{\{A-\pi(X,\psi^*)\}\mu(1,X,\xi^*)}{\pi(X,\psi^*)} - \frac{(1-A)Y}{1-\pi(X,\psi^*)} - \frac{\{A-\pi(X,\psi^*)\}\mu(0,X,\xi^*)}{1-\pi(X,\psi^*)}\Bigg\}. \tag{13.38}$$

By the SUTVA assumption,

$$\frac{AY}{\pi(X,\psi^*)} = \frac{AY^*(1)}{\pi(X,\psi^*)} = Y^*(1) + \frac{\{A-\pi(X,\psi^*)\}Y^*(1)}{\pi(X,\psi^*)}, \tag{13.39}$$

and

$$\frac{(1-A)Y}{1-\pi(X,\psi^*)} = \frac{(1-A)Y^*(0)}{1-\pi(X,\psi^*)} = Y^*(0) - \frac{\{A-\pi(X,\psi^*)\}Y^*(0)}{1-\pi(X,\psi^*)}. \tag{13.40}$$

Substituting (13.39) and (13.40) into (13.38) yields

$$E\{Y^*(1) - Y^*(0)\} \tag{13.41}$$

$$+E\left[\frac{\{A-\pi(X,\psi^*)\}\{Y^*(1)-\mu(1,X,\xi^*)\}}{\pi(X,\psi^*)}\right] \tag{13.42}$$

$$+E\left[\frac{\{A-\pi(X,\psi^*)\}\{Y^*(0)-\mu(0,X,\xi^*)\}}{1-\pi(X,\psi^*)}\right]. \tag{13.43}$$

Since $E\{Y^*(1)-Y^*(0)\}$ of (13.41) is equal to δ, the proof of double robustness will follow if we can show that the expectations in (13.42) and (13.43) are equal to zero if either $\psi^* = \psi_0$ or $\xi^* = \xi_0$.

Let us first assume that the propensity score is correctly specified; i.e., $\psi^* = \psi_0$. We compute the expectation of (13.42) by iterated conditional expectations, where we first condition on $Y^*(1)$ and X to obtain that (13.42) equals

$$E\left(\frac{[E\{A|Y^*(1),X\} - \pi(X,\psi_0)]\{Y^*(1)-\mu(1,X,\xi^*)\}}{\pi(X,\psi_0)}\right).$$

Because of the strong ignorability assumption, $E\{A|Y^*(1), X\} = E(A|X) = \pi(X, \psi_0)$, thus allowing us to conclude that (13.42) equals zero when $\psi^* = \psi_0$. A similar argument can also be used to show that (13.43) equals zero when $\psi^* = \psi_0$.

Now consider the case where the regression model $E(Y|A, X)$ is correctly specified; i.e., $\xi^* = \xi_0$. Again, we compute the expectation of (13.42) by iterated conditional expectations, but now we first condition on A and X to obtain that (13.42) equals

$$E\left(\frac{\{A - \pi(X, \psi^*)\}[E\{Y^*(1)|A, X\} - \mu(1, X, \xi_0)]}{\pi(X, \psi^*)}\right).$$

Because of SUTVA, $\mu(1, X, \xi_0) = E(Y|A = 1, X) = E\{Y^*(1)|A = 1, X\}$, and because of the strong ignorability assumption,

$$E\{Y^*(1)|X\} = E\{Y^*(1)|A, X\} = E\{Y^*(1)|A = 1, X\} = \mu(1, X, \xi_0),$$

thus allowing us to conclude that (13.42) equals zero when $\xi^* = \xi_0$. A similar argument can also be used to show that (13.43) equals zero when $\xi^* = \xi_0$.

The local semiparametric efficiency together with the double-robustness property makes the estimator (13.37) desirable as compared with, say, the regression estimator (13.13), whose consistency is completely dependent on correctly modeling the regression relationship $E(Y|A, X)$.

Remark 6. The connection of the propensity score to causality has been studied carefully by Rosenbaum, Rubin, and others; see, for example, Rosenbaum and Rubin (1983, 1984, 1985), Rosenbaum (1984, 1987), and Rubin (1997). Different methods for estimating the average causal treatment effect using propensity scores have been advocated. These include stratification, matching, and inverse propensity weighting. The locally efficient estimator derived above is sometimes referred to as the augmented inverse propensity weighted estimator. This estimator was compared with other commonly used estimators of the average causal treatment effect in a series of numerical simulations by Lunceford and Davidian (2004), who generally found that this estimator performs the best across a wide variety of scenarios. □

13.6 Exercises for Chapter 13

1. Derive the influence function for the regression estimator $\hat{\delta}_n$ given by equation (13.13).
2. Derive the influence function for the augmented inverse propensity weighted estimator $\hat{\delta}_n$ given by equation (13.37).

14

Multiple Imputation: A Frequentist Perspective

A popular approach for dealing with missing data is the use of multiple imputation, which was first introduced by Rubin (1978b). Although most of this book has focused on semiparametric models, where the model includes infinite-dimensional nuisance parameters, this chapter will only consider finite-dimensional parametric models, as in Chapter 3. Because of its importance in missing-data problems, we conclude with a discussion of this methodology.

Imputation methods, where we replace missing values by some "best guess" and then analyze the data as if complete, have a great deal of intuitive appeal. However, unless one is careful about how the imputation is carried out and how the subsequent inference is made, imputation methods may lead to biased estimates with estimated confidence intervals that are too narrow. Rubin's proposal for multiple imputation allowed the use of this intuitive idea in a manner that results in correct inference. Rubin's justification is based on a Bayesian paradigm. In this chapter, we will consider the statistical properties of multiple-imputation estimators from a frequentist point of view using large sample theory. Much of the development is based on two papers, by Wang and Robins (1998) and Robins and Wang (2000). Although most of the machinery developed in the previous chapters that led to AIPWCC estimators will not be used here, we feel that this topic is of sufficient importance to warrant study in its own right.

As we have all along, we will consider a full-data model, where the full data are denoted by $Z_1, \ldots, Z_n$ assumed iid with density $p_Z(z, \beta)$, where β is a finite-dimensional parameter, say q-dimensional. The observed (coarsened) data will be assumed coarsened at random and denoted by

$$\{\mathcal{C}_i, G_{\mathcal{C}_i}(Z_i)\}, i = 1, \ldots, n.$$

Remark 1. When data are coarsened at random and the full-data parameter β is finite-dimensional, then β can be estimated using maximum likelihood. The coarsened-data likelihood was derived in (7.10), where we showed that β could be estimated by maximizing

$$\prod_{i=1}^{n} p_{G_{r_i}(Z_i)}(g_{r_i}, \beta), \tag{14.1}$$

where

$$p_{G_r(Z)}(g_r, \beta) = \int\limits_{\{z: G_r(z) = g_r\}} p_Z(z, \beta) d\nu_Z(z).$$

However, even for parametric models, maximizing coarsened-data likelihoods may be difficult to implement. Multiple imputation is an attempt to use simpler methods that are easier to implement. □

We will assume throughout that the full-data maximum likelihood estimator for β can be derived easily using standard software. We will also assume that the standard large-sample properties of maximum likelihood estimators apply to the full-data model. For example, denoting the full-data score vector by

$$S^F(z, \beta) = \frac{\partial \log p_Z(z, \beta)}{\partial \beta},$$

the MLE $\hat{\beta}_n^F$ is obtained by solving the likelihood equation

$$\sum_{i=1}^{n} S^F(Z_i, \beta) = 0.$$

Remark 2. On notation
In this chapter, we only consider the parameter β with no additional nuisance parameters. Therefore, the full-data score vector will be denoted by $S^F(Z, \beta)$ (without the subscript β used in the previous chapters). As usual, when we use the notation $S^F(Z, \beta)$, we are considering a q-dimensional vector of functions of Z and β. If the score vector is evaluated at the truth, $\beta = \beta_0$, then this will often be denoted by $S^F(Z) = S^F(Z, \beta_0)$. A similar convention will be used when we consider the observed-data score vector, which will be denoted as $S\{\mathcal{C}, G_{\mathcal{C}}(Z), \beta\}$, and, at the truth, as $S\{\mathcal{C}, G_{\mathcal{C}}(Z), \beta_0\}$ or $S\{\mathcal{C}, G_{\mathcal{C}}(Z)\}$. □

Under suitable regularity conditions, which will be assumed throughout,

$$n^{1/2}(\hat{\beta}_n^F - \beta_0) \xrightarrow{\mathcal{D}} N\Big(0, V_{\text{eff}}^F(\beta_0)\Big),$$

where $\{V_{\text{eff}}^F(\beta_0)\}^{-1}$ is the full-data information matrix, which we denote by $I^F(\beta_0) = E\{S^F(Z)S^{F^T}(Z)\}$. That is,

$$V_{\text{eff}}^F(\beta_0) = \left\{I^F(\beta_0)\right\}^{-1}.$$

We present an illustrative example where multiple-imputation methods may be used.

Example 1. Surrogate marker problem
Consider the logistic regression model used to model the probability of a dichotomous response as a function of the covariates X. Letting $Y = \{1, 0\}$ denote a binary response variable, we assume

$$P(Y = 1|X) = \frac{\exp(\theta^T X)}{1 + \exp(\theta^T X)}. \tag{14.2}$$

With full data $(Y_i, X_i), i = 1, \ldots, n$, the maximum likelihood estimator for θ can be obtained in a straightforward fashion using standard software. We wish to consider the problem where the variable X, or a subset of X, is expensive to obtain. For instance, X may represent some biological marker, derived from stored plasma collected on everyone, that is expensive to measure. However, another variable, W, is identified as a cheaper surrogate for X. That is, W is correlated with X and satisfies the surrogacy assumption; namely,

$$P(Y = 1|X, W) = P(Y = 1|X) = \frac{\exp(\theta^T X)}{1 + \exp(\theta^T X)}.$$

Let us also assume that (X, W) follows a multivariate normal distribution

$$\begin{pmatrix} X \\ W \end{pmatrix} \sim N\left[\begin{pmatrix} \mu_X \\ \mu_W \end{pmatrix}, \begin{bmatrix} \sigma_{XX} & \sigma_{XW} \\ \sigma_{XW}^T & \sigma_{WW} \end{bmatrix}\right].$$

In an attempt to save costs, the inexpensive surrogate variable W will be collected on everyone in the sample, as will the response variable Y. However, the expensive variable X will be obtained on only a validation subsample of individuals in the study using stored blood chosen at random with a prespecified probability of selection that may depend on (Y, W) (i.e., missing at random). Letting R denote the indicator of a complete-case, then the observed data are denoted as

$$(R_i, Y_i, W_i, R_i X_i), i = 1, \ldots, n.$$

Although the primary focus is the parameter θ, since the model is finite-dimensional, we will not differentiate between the parameters of interest and the nuisance parameters. Hence

$$\beta = (\theta^T, \mu_X^T, \mu_W^T, \sigma_{XX}, \sigma_{XW}, \sigma_{WW})^T.$$

This example will be used for illustration later. □

With observed data, the parameter β can be estimated efficiently by maximizing the observed-data likelihood (14.1). That is, we would find the solution to the observed-data likelihood equation

$$\sum_{i=1}^{n} S\{\mathcal{C}_i, G_{\mathcal{C}_i}(Z_i), \beta\} = 0, \tag{14.3}$$

where the observed-data score vector is given by

$$S\{\mathcal{C}_i, G_{\mathcal{C}_i}(Z_i), \beta\} = E\{S^F(Z_i, \beta)|\mathcal{C}_i, G_{\mathcal{C}_i}(Z_i), \beta\}. \tag{14.4}$$

Remark 3. It is important for some of the subsequent developments that we introduce the following notation. When we write

$$E\{S^F(Z_i, \beta')|\mathcal{C}_i, G_{\mathcal{C}_i}(Z_i), \beta''\},$$

where we allow β', β'' to be arbitrary values not necessarily at the truth or equal to each other, we mean specifically that

$$\begin{aligned} E\{S^F(Z_i, \beta')|\mathcal{C}_i = r_i, G_{\mathcal{C}_i}(Z_i) = g_{r_i}, \beta''\} \\ = \frac{\int\limits_{\{z:G_{r_i}(z)=g_{r_i}\}} S^F(z, \beta')p_Z(z, \beta'')d\nu_Z(z)}{\int\limits_{\{z:G_{r_i}(z)=g_{r_i}\}} p_Z(z, \beta'')d\nu_Z(z)}. \end{aligned} \tag{14.5}$$

Unless otherwise stated, expectations and conditional expectations will be evaluated with parameter values at the truth. Consequently, the observed-data score vector, evaluated at the truth, is

$$S\{\mathcal{C}, G_{\mathcal{C}}(Z)\} = E\{S^F(Z)|\mathcal{C}, G_{\mathcal{C}}(Z)\} = E\{S^F(Z, \beta_0)|\mathcal{C}, G_{\mathcal{C}}(Z), \beta_0\}. \square$$

Therefore, when we derive the observed-data MLE by solving (14.3), we are specifically looking for $\hat{\beta}_n$ that satisfies

$$\sum_{i=1}^{n} E\{S^F(Z_i, \hat{\beta}_n)|\mathcal{C}_i, G_{\mathcal{C}_i}(Z_i), \hat{\beta}_n\} = 0.$$

$$\uparrow \qquad\qquad \uparrow$$

notice that these two must be equal

Again, under suitable regularity conditions,

$$n^{1/2}(\hat{\beta}_n - \beta_0) \xrightarrow{\mathcal{D}} N\Big(0, V_{\text{eff}}(\beta_0)\Big),$$

where $\{V_{\text{eff}}(\beta_0)\}^{-1}$ is the observed-data information matrix, denoted by

$$I(\beta_0) = E\left[S\{\mathcal{C}_i, G_{\mathcal{C}}(Z_i)\}S^T\{\mathcal{C}_i, G_{\mathcal{C}_i}(Z_i)\}\right].$$

14.1 Full- Versus Observed-Data Information Matrix

Because coarsened data $G_r(Z)$ represent a many-to-one transformation of the full data Z (i.e., a data reduction), it seems intuitively clear that the asymptotic variance of the full-data MLE $\hat{\beta}_n^F$ will be smaller than the asymptotic

variance of the observed-data MLE $\hat{\beta}_n$. In this section, we give a formal proof of this result.

Let $\text{var}\{S^F(Z)\} = E\{S^F(Z)S^{F^T}(Z)\}$ denote that variance matrix of $S^F(Z)$. Then the asymptotic variance of the full-data MLE $\hat{\beta}_n^F$ is given by $V_{\text{eff}}^F(\beta_0) = [\text{var}\{S^F(Z)\}]^{-1}$. Similarly, the asymptotic variance of the observed-data MLE $\hat{\beta}_n$ is $V_{\text{eff}}(\beta_0) = (\text{var}[S\{\mathcal{C}, G_{\mathcal{C}}(Z)\}])^{-1}$.

We now give two results regarding the full-data and observed-data estimators for β.

Theorem 14.1. The observed-data information matrix is smaller than or equal to the full-data information matrix; that is,

$$\text{var}[S\{\mathcal{C}, G_{\mathcal{C}}(Z)\}] \leq \text{var}\{S^F(Z)\},$$

where the notation "$\leq$" means that $\text{var}\{S^F(Z)\} - \text{var}[S\{\mathcal{C}, G_{\mathcal{C}}(Z)\}]$ is nonnegative definite.

Proof. By the law of conditional variance,

$$\begin{aligned}\text{var}\ \{S^F(Z)\} &= \text{var}\left[E\left\{S^F(Z)|\mathcal{C}, G_{\mathcal{C}}(Z)\right\}\right] + E\left[\text{var}\left\{S^F(Z)|\mathcal{C}, G_{\mathcal{C}}(Z)\right\}\right] \\ &= \text{var}\left[S\{\mathcal{C}, G_{\mathcal{C}}(Z)\}\right] + E\left[\text{var}\left\{S^F(Z)|\mathcal{C}, G_{\mathcal{C}}(Z)\right\}\right]. \qquad (14.6)\end{aligned}$$

This implies that

$$\text{var}\{S^F(Z)\} - \text{var}[S\{\mathcal{C}, G_{\mathcal{C}}(Z)\}]$$

is nonnegative definite. □

Theorem 14.2. The asymptotic variance of the full-data MLE is smaller than or equal to the asymptotic variance of the observed-data MLE; that is,

$$V_{\text{eff}}^F(\beta_0) \leq V_{\text{eff}}(\beta_0). \qquad (14.7)$$

Proof. This follows from results about influence functions; namely, if we define

$$\varphi_{\text{eff}}^F(Z) = \left[E\left\{S^F(Z)S^{F^T}(Z)\right\}\right]^{-1} S^F(Z)$$

and

$$\varphi_{\text{eff}}\{\mathcal{C}, G_{\mathcal{C}}(Z)\} = \left(E\left[S\{\mathcal{C}, G_{\mathcal{C}}(Z)\}S^T\{\mathcal{C}, G_{\mathcal{C}}(Z)\}\right]\right)^{-1} S\{\mathcal{C}, G_{\mathcal{C}}(Z)\},$$

then

$$\varphi_{\text{eff}}\{\mathcal{C}, G_{\mathcal{C}}(Z)\} = \varphi_{\text{eff}}^F(Z) + \left[\varphi_{\text{eff}}\{\mathcal{C}, G_{\mathcal{C}}(Z)\} - \varphi_{\text{eff}}^F(Z)\right].$$

We also note that

$$\begin{aligned}E\left[S^F(Z)S^T\{\mathcal{C}, G_{\mathcal{C}}(Z)\}\right] &= E\left(E\left[S^F(Z)S^T\{\mathcal{C}, G_{\mathcal{C}}(Z)\}|\mathcal{C}, G_{\mathcal{C}}(Z)\right]\right) \\ &= E\left[S\{\mathcal{C}, G_{\mathcal{C}}(Z)\}S^T\{\mathcal{C}, G_{\mathcal{C}}(Z)\}\right]. \qquad (14.8)\end{aligned}$$

Equation (14.8) can be used to show that

$$E(\varphi^F_{\text{eff}}(Z)\left[\varphi_{\text{eff}}\{\mathcal{C}, G_{\mathcal{C}}(Z)\} - \varphi^F_{\text{eff}}(Z)\right]^T) = 0,$$

which, in turn, can be used to show that

$$\text{var}[\varphi_{\text{eff}}\{\mathcal{C}, G_{\mathcal{C}}(Z)\}] = \text{var}\{\varphi^F_{\text{eff}}(Z)\} + \text{var}(\left[\varphi_{\text{eff}}\{\mathcal{C}, G_{\mathcal{C}}(Z)\} - \varphi^F_{\text{eff}}(Z)\right])$$

or

$$\text{var}[\varphi_{\text{eff}}\{\mathcal{C}, G_{\mathcal{C}}(Z)\}] - \text{var}\{\varphi^F_{\text{eff}}(Z)\}$$

is nonnegative definite. The proof is complete upon noticing that

$$\text{var}[\varphi_{\text{eff}}\{\mathcal{C}, G_{\mathcal{C}}(Z)\}] = (\text{var}[S\{\mathcal{C}, G_{\mathcal{C}}(Z)\}])^{-1} = V_{\text{eff}}(\beta_0)$$

and

$$\text{var}\{\varphi^F_{\text{eff}}(Z)\} = [\text{var}\left\{S^F(Z)\right\}]^{-1} = V^F_{\text{eff}}(\beta_0).$$

□

For many problems, working with the observed (coarsened) data likelihood may be difficult, with no readily available software. For such instances, it may be useful to find methods where we can use the simpler full-data inference to analyze coarsened data. This is what motivated much of the inverse weighted methodology discussed in the previous chapters. **Multiple imputation** is another such methodology popularized by Rubin (1978b, 1987). See also the excellent overview paper by Rubin (1996).

14.2 Multiple Imputation

Multiple imputation is implemented as follows. For each observed-data value $\{\mathcal{C}_i, G_{\mathcal{C}_i}(Z_i)\}$, we sample at random from the conditional distribution

$$p_{Z|\mathcal{C}, G_{\mathcal{C}}(Z)}\{z|\mathcal{C}_i, G_{\mathcal{C}_i}(Z_i)\}$$

m times to obtain random quantities

$$Z_{ij}, j = 1, \ldots, m, i = 1, \ldots, n.$$

These are the imputation of the full data from the observed (coarsened) data. The "j"-th set of imputed full data, denoted by $Z_{ij}, i = 1, \ldots, n$, is used to obtain the j-th estimator $\hat{\beta}^*_{nj}$ by solving the full-data likelihood equation:

$$\sum_{i=1}^{n} S^F(Z_{ij}, \hat{\beta}^*_{nj}) = 0. \tag{14.9}$$

That is, we use the j-th imputed data set and treat these data as if they were full data to obtain the full-data MLE $\hat{\beta}^*_{nj}$.

The proposed multiple-imputation estimator is

$$\hat{\beta}^*_n = m^{-1} \sum_{j=1}^{m} \hat{\beta}^*_{nj}. \tag{14.10}$$

Rubin (1987) argues that, under appropriate conditions, this estimator is consistent and asymptotically normal. That is,

$$n^{1/2}(\hat{\beta}^*_n - \beta_0) \xrightarrow{\mathcal{D}} N(0, \Sigma^*),$$

where the asymptotic variance Σ^* can be estimated by $\hat{\Sigma}^* =$

$$m^{-1} \underbrace{\sum_{j=1}^{m} \left\{ n^{-1} \sum_{i=1}^{n} -\frac{\partial S^F(Z_{ij}, \hat{\beta}^*_{nj})}{\partial \beta^T} \right\}^{-1}}$$

$$+ \left(\frac{m+1}{m} \right) n \underbrace{\frac{\sum_{j=1}^{m} \left(\hat{\beta}^*_{nj} - \hat{\beta}^*_n \right) \left(\hat{\beta}^*_{nj} - \hat{\beta}^*_n \right)^T}{m-1}}.$$

We note that the first term in the sum is an average of the estimators of the full-data asymptotic variance using the inverse of the full-data observed information matrix over the imputed full-data sets and the second term is the sample variance of the imputation estimators multiplied by a finite "m" correction factor.

Remark 4. If we knew the true value, β_0, then we could generate a random variable $Z_{ij}(\beta_0)$ whose distribution has density $p_Z(z, \beta_0)$ by first generating a random $\{\mathcal{C}_i, G_{\mathcal{C}_i}(Z_i)\}$ from the distribution with density $p_{\mathcal{C}, G_{\mathcal{C}}(Z)}(r, g_r, \beta_0)$ and then generating a random $Z_{ij}(\beta_0)$ from the conditional distribution with conditional density

$$p_{Z|\mathcal{C}, G_{\mathcal{C}}(Z)}\{z|\mathcal{C}_i, G_{\mathcal{C}_i}(Z_i), \beta_0\}.$$

This is essentially the logic that motivates multiple imputation. Assuming the model is correct, the observed data $\{\mathcal{C}_i, G_{\mathcal{C}_i}(Z_i)\}$ are generated, by nature, from the density $p_{\mathcal{C}, G_{\mathcal{C}}(Z)}(r, g_r, \beta_0)$. However, since we don't know the true value of β_0, the Z_{ij} must be generated from the conditional density

$$p_{Z|\mathcal{C}, G_{\mathcal{C}}(Z)}\{z|\mathcal{C}_i, G_{\mathcal{C}_i}(Z_i), \beta\},$$

where β must be estimated from the data in some fashion. Such imputed values are referred to as $Z_{ij}(\beta)$. □

We will consider two cases.

(a) In what we will denote as *frequentist* imputation, an initial estimator $\hat{\beta}_n^I$ is obtained from the coarsened data. The imputation $Z_{ij}(\hat{\beta}_n^I)$ are obtained by sampling from

$$p_{Z|\mathcal{C},G_{\mathcal{C}}(Z)}\{z|\mathcal{C}_i, G_{\mathcal{C}_i}(Z_i), \hat{\beta}_n^I\}.$$

The resulting estimator is referred to by Wang and Robins (1998) as a type B estimator. Rubin (1987) also refers to this type of imputation as improper imputation.

(b) For *Bayesian* imputation (the approach advocated by Rubin), the Z_{ij} are generated from the predictive distribution

$$p_{Z|\mathcal{C},G_{\mathcal{C}}(Z)}\{z|\mathcal{C}_i, G_{\mathcal{C}_i}(Z_i)\}.$$

The predictive distribution is given by

$$\int p_{Z|\mathcal{C},G_{\mathcal{C}}(Z)}\{z|\mathcal{C}_i, G_{\mathcal{C}_i}(Z_i), \beta\} \underbrace{p_{\beta|\mathcal{C},G_{\mathcal{C}}(Z)}\{\beta|\mathcal{C}_i, G_{\mathcal{C}_i}(Z_i)\}d\nu_\beta(\beta)}_{\text{This is the posterior distribution of } \beta \text{ given the observed data}}.$$

In order to implement the Bayesian approach, which Rubin refers to as proper imputation, one needs to specify a prior distribution for β, from which the posterior distribution $p_{\beta|\mathcal{C},G_{\mathcal{C}}(Z)}\{\beta|\mathcal{C}_i, G_{\mathcal{C}_i}(Z_i)\}$ is computed using Bayes's rule. Then

(i) we randomly select $\beta^{(j)}$ from the posterior distribution and,
(ii) conditional on $\beta^{(j)}$, we sample from $p_{Z|\mathcal{C},G_{\mathcal{C}}(Z)}\{z|\mathcal{C}_i, G_{\mathcal{C}_i}(Z_i), \beta^{(j)}\}$ to obtain $Z_{ij}(\beta^{(j)})$.

A multiple-imputation estimator using this approach is referred to by Wang and Robins (1998) as a type A estimator.

14.3 Asymptotic Properties of the Multiple-Imputation Estimator

In what follows, we will derive the asymptotic properties of the multiple-imputation estimator. We first will consider the "frequentist" approach, which imputes from the conditional distribution fixing the parameter β at some initial estimator $\hat{\beta}_n^I$. Some discussion of Bayesian proper imputation will also be considered later.

In order to implement multiple imputation from a frequentist view, we must start with some initial estimator for β, say $\hat{\beta}_n^I$. For missing-data problems, an initial estimator can often be obtained using only the complete cases (i.e., $\{i : \mathcal{C}_i = \infty\}$). For example, when the data are missing completely at random, we can obtain a consistent asymptotically normal estimator using only

the complete cases since these are a representative sample (albeit smaller) from the population. If the data are missing at random (MAR), we might use an inverse probability weighted complete-case estimator as an initial value.

The initial estimator for β will be assumed to be an RAL estimator; that is,

$$n^{1/2}(\hat{\beta}_n^I - \beta_0) = n^{-1/2} \sum_{i=1}^{n} q\{\mathcal{C}_i, G_{\mathcal{C}_i}(Z_i)\} + o_p(1),$$

where $q\{\mathcal{C}_i, G_{\mathcal{C}_i}(Z_i)\}$ is the i-th influence function of the estimator $\hat{\beta}_n^I$. We remind the reader of the following properties of influence functions for RAL estimators:

(a) The efficient influence function, denoted by $\varphi_{\text{eff}}\{\mathcal{C}_i, G_{\mathcal{C}_i}(Z_i)\}$, equals

$$\left(E\left[S\{\mathcal{C}_i, G_{\mathcal{C}_i}(Z_i)\}S^T\{\mathcal{C}_i, G_{\mathcal{C}_i}(Z_i)\}\right]\right)^{-1} S\{\mathcal{C}_i, G_{\mathcal{C}_i}(Z_i)\}. \tag{14.11}$$

(b) For any influence function of an RAL estimator $q\{\mathcal{C}_i, G_{\mathcal{C}_i}(Z_i)\}$,

$$E\left[q\{\mathcal{C}_i, G_{\mathcal{C}_i}(Z_i)\}S^T\{\mathcal{C}_i, G_{\mathcal{C}_i}(Z_i)\}\right] = \underbrace{I^{q\times q}}_{\text{the identity matrix}}. \tag{14.12}$$

(c) The influence function $q\{\mathcal{C}_i, G_{\mathcal{C}_i}(Z_i)\}$ can be written as

$$q\{\mathcal{C}_i, G_{\mathcal{C}_i}(Z_i)\} = \varphi_{\text{eff}}\{\mathcal{C}_i, G_{\mathcal{C}_i}(Z_i)\} + h\{\mathcal{C}_i, G_{\mathcal{C}_i}(Z_i)\}, \tag{14.13}$$

where $h\{\mathcal{C}_i, G_{\mathcal{C}_i}(Z_i)\}$ has mean zero and

$$E\left[h\{\mathcal{C}_i, G_{\mathcal{C}_i}(Z_i)\}S^T\{\mathcal{C}_i, G_{\mathcal{C}_i}(Z_i)\}\right] = 0^{q\times q}.$$

Hence

$$\begin{aligned}\text{var}\left[q\{\mathcal{C}_i, G_{\mathcal{C}_i}(Z_i)\}\right] &= \text{var}\left[\varphi_{\text{eff}}\{\mathcal{C}_i, G_{\mathcal{C}_i}(Z_i)\}\right] + \text{var}\left[h\{\mathcal{C}_i, G_{\mathcal{C}_i}, (Z_i)\}\right] \\ &= \left(E\left[S\{\mathcal{C}_i, G_{\mathcal{C}_i}(Z_i)\}S^T\{\mathcal{C}_i, G_{\mathcal{C}_i}(Z_i)\}\right]\right)^{-1} + \text{var}\left[h\{\mathcal{C}_i, G_{\mathcal{C}_i}(Z_i)\}\right].\end{aligned} \tag{14.14}$$

In studying the asymptotic properties of $n^{1/2}(\hat{\beta}_n^* - \beta_0)$, we will consider the following questions.

1. Is the distribution asymptotically normal?
2. What is its asymptotic variance?
3. Can we estimate the asymptotic variance? How well does Rubin's variance estimator work?
4. What are the efficiency properties of the multiple-imputation estimator?

We begin the study of the asymptotic properties by first introducing some preliminary lemmas. Rigorous proofs of these lemmas require careful analysis

and the use of empirical processes that are beyond the scope of this book. We will, however, provide some heuristic justification of the results.

As we have emphasized repeatedly throughout this book, the key to the asymptotic properties of RAL estimators is being able to derive their influence function. We will derive the influence function of the multiple-imputation estimator through a series of approximations. We begin by giving the first such approximation.

Lemma 14.1. The multiple-imputation estimator $\hat{\beta}_n^*$ defined by (14.9) and (14.10) using the imputed values $Z_{ij}(\hat{\beta}_n^I), j = 1, \ldots, m,\ i = 1, \ldots, n$, where $\hat{\beta}_n^I$ denotes some initial estimator, can be approximated as

$$n^{1/2}(\hat{\beta}_n^* - \beta_0) = n^{-1/2} \sum_{i=1}^{n} \left\{ I^F(\beta_0) \right\}^{-1} \left[m^{-1} \sum_{j=1}^{m} S^F\{Z_{ij}(\hat{\beta}_n^I), \beta_0\} \right] + o_p(1), \tag{14.15}$$

where $I^F(\beta_0)$ is the full-data information matrix; namely,

$$E\left\{ \frac{-\partial S^F(Z, \beta_0)}{\partial \beta^T} \right\} = E\left\{ S^F(Z, \beta_0) S^{F^T}(Z, \beta_0) \right\}.$$

Proof. Heuristic sketch
The j-th imputation estimator $\hat{\beta}_{nj}^*$ is the solution to the equation

$$\sum_{i=1}^{n} S^F\{Z_{ij}(\hat{\beta}_n^I), \hat{\beta}_{nj}^*\} = 0,$$

which by the mean value expansion is equal to

$$\sum_{i=1}^{n} S^F\{Z_{ij}(\hat{\beta}_n^I), \beta_0\} + \left[\sum_{i=1}^{n} \frac{\partial S^F\{Z_{ij}(\hat{\beta}_n^I), \tilde{\beta}_{nj}\}}{\partial \beta^T} \right] \times \left(\hat{\beta}_{nj}^* - \beta_0 \right),$$

where $\tilde{\beta}_{nj}$ is an intermediate value between $\hat{\beta}_{nj}^*$ and β_0. Under suitable regularity conditions,

$$-n^{-1} \sum_{i=1}^{n} \frac{\partial S^F\{Z_{ij}(\hat{\beta}_n^I), \tilde{\beta}_{nj}\}}{\partial \beta^T} \xrightarrow{P} E\left[-\frac{\partial S^F\{Z_{ij}(\beta_0), \beta_0\}}{\partial \beta^T} \right], \tag{14.16}$$

which, by standard results from likelihood theory, is the information matrix, which is also equal to

$$E\left[S^F\{Z_{ij}(\beta_0), \beta_0\} S^{F^T}\{Z_{ij}(\beta_0), \beta_0\} \right] = I^F(\beta_0).$$

Therefore,

$$n^{1/2}\left(\hat{\beta}^*_{nj} - \beta_0\right) = n^{1/2}\sum_{i=1}^{n}\{I^F(\beta_0)\}^{-1}S^F\{Z_{ij}(\hat{\beta}^I_n), \beta_0\} + o_p(1), \quad (14.17)$$

which implies that

$$n^{1/2}(\hat{\beta}^*_n - \beta_0) = n^{-1/2}\sum_{i=1}^{n}\left\{I^F(\beta_0)\right\}^{-1}\left[m^{-1}\sum_{j=1}^{m}S^F\{Z_{ij}(\hat{\beta}^I_n), \beta_0\}\right] + o_p(1). \quad \square$$

As a consequence of Lemma 14.1, we have shown that $n^{1/2}(\hat{\beta}^*_n - \beta_0)$, up to order $o_p(1)$, is proportional to

$$n^{-1/2}\sum_{i=1}^{n}\left[m^{-1}\sum_{j=1}^{m}S^F\{Z_{ij}(\hat{\beta}^I_n), \beta_0\}\right]. \quad (14.18)$$

However, this does not give us the desired influence function for $\hat{\beta}^*_n$ because $Z_{ij}(\hat{\beta}^I_n)$ is evaluated at a random quantity $(\hat{\beta}^I_n)$ that involves data from all individuals and therefore (14.18) is not a sum of iid terms. In order to find the influence function, we write (14.18) as the sum of

$$n^{-1/2}\sum_{i=1}^{n}\left[m^{-1}\sum_{j=1}^{m}S^F\{Z_{ij}(\beta_0), \beta_0\}\right] \quad (14.19)$$

$+$

$$n^{-1/2}\sum_{i=1}^{n}\left[m^{-1}\sum_{j=1}^{m}S^F\{Z_{ij}(\hat{\beta}^I_n), \beta_0\} - m^{-1}\sum_{j=1}^{m}S^F\{Z_{ij}(\beta_0), \beta_0\}\right] \quad (14.20)$$

and then show how these two pieces can be expressed approximately as a sum of iid terms.

Remark 5. We remind the reader that, for a fixed β, $Z_{ij}(\beta), j = 1, \ldots, m,\ i = 1, \ldots, n$ are obtained through a two-stage process. For any i, nature (sampling from the population) provides us with the data $\{\mathcal{C}_i, G_{\mathcal{C}_i}(Z_i)\}$ derived from the distribution $p_{\mathcal{C},G_{\mathcal{C}}(Z)}(r, g_r, \beta_0)$. Then, for $j = 1, \ldots, m$, the data analyst draws m values at random from the conditional distribution $p_{Z|\mathcal{C},G_{\mathcal{C}}(Z)}\{z|\mathcal{C}_i, G_{\mathcal{C}_i}(Z_i), \beta\}$. Consequently, the vector $\{Z_{i1}(\beta), \ldots, Z_{im}(\beta)\}$ is made up of correlated random variables, but these random vectors across i are iid random vectors. Also, because of the way the data are generated, the marginal distribution of $Z_{ij}(\beta)$ is the same for all i and j. If $\beta = \beta_0$, then the marginal distribution of $Z_{ij}(\beta_0)$ has density $p_Z(z, \beta_0)$ (i.e., the density for the full data at the truth). However, if $\beta \neq \beta_0$, then the marginal density for $Z_{ij}(\beta)$ is more complex. $\square$

Based on the discussion in Remark 5, (14.19) is made up of a sum of n iid elements, where the i-th element of the sum is equal to

$$m^{-1}\sum_{j=1}^{m} S^F\{Z_{ij}(\beta_0),\beta_0\}, \tag{14.21}$$

$i = 1,\ldots,n$. In addition, the distribution of $S^F\{Z_{ij}(\beta_0),\beta_0\}$ is the same as the distribution of $S^F(Z_i,\beta_0)$, where Z_i is the full data. Therefore,

$$E[S^F\{Z_{ij}(\beta_0),\beta_0\}] = E\{S^F(Z_i,\beta_0)\} = 0 \text{ for all } i,j,$$

which implies that (14.21) has mean zero. Hence, (14.19) is a normalized sum of mean-zero iid random vectors that will converge to a normal distribution by the central limit theorem.

We now consider the approximation of (14.20) as a sum of iid random vectors. This is given by the following theorem.

Theorem 14.3. The expression (14.20) is equal to

$$n^{-1/2}\sum_{i=1}^{n}\left\{I^F(\beta_0) - I(\beta_0)\right\} q\{\mathcal{C}_i, G_{\mathcal{C}_i}(Z_i)\} + o_p(1), \tag{14.22}$$

where $I^F(\beta_0)$ is the full-data information matrix,

$$E\left\{\frac{-\partial S^F(Z,\beta_0)}{\partial \beta^T}\right\} = E\left\{S^F(Z,\beta_0)S^{F^T}(Z,\beta_0)\right\},$$

$I(\beta_0)$ is the observed-data information matrix,

$$E\left[-\frac{\partial S\{\mathcal{C},G_{\mathcal{C}}(Z),\beta_0\}}{\partial\beta^T}\right] = E\left[S\{\mathcal{C},G_{\mathcal{C}}(Z),\beta_0\}S^T\{\mathcal{C},G_{\mathcal{C}}(Z),\beta_0\}\right],$$

and $q\{\mathcal{C}_i, G_{\mathcal{C}_i}(Z_i)\}$ is the i-th influence function of the initial estimator $\hat{\beta}_n^I$.

Theorem 14.3 will be proved using a series of lemmas.

Lemma 14.2. Let $Z_{ij}(\beta)$ denote a random draw from the conditional distribution with conditional density $p_{Z|\mathcal{C},G_{\mathcal{C}}(Z)}\{z|\mathcal{C}_i, G_{\mathcal{C}_i}(Z_i),\beta\}$. If we define

$$\lambda(\beta,\beta_0) = E\left[S^F\{Z_{ij}(\beta),\beta_0\}\right], \tag{14.23}$$

then

$$\left.\frac{\partial\lambda(\beta,\beta_0)}{\partial\beta^T}\right|_{\beta=\beta_0} = I^F(\beta_0) - I(\beta_0).$$

Proof. By the law of conditional expectations, (14.23) equals

$$E\Big(E\left[S^F\{Z_{ij}(\beta),\beta_0\}|\mathcal{C}_i, G_{\mathcal{C}_i}(Z_i),\beta\right]\Big). \tag{14.24}$$

Because $Z_{ij}(\beta)$ is a random draw from the conditional distribution with conditional density

$$p_{Z|\mathcal{C},G_{\mathcal{C}}(Z)}\{z|\mathcal{C}_i, G_{\mathcal{C}_i}(Z_i), \beta\},$$

this means that the inner expectation of (14.24) is equal to

$$\int S^F(z,\beta_0) p_{Z|\mathcal{C},G_{\mathcal{C}}(Z)}\{z|\mathcal{C}_i, G_{\mathcal{C}_i}(Z_i), \underset{\substack{\uparrow \\ \text{note the value} \\ \beta \text{ here}}}{\beta}\} d\nu_{Z|\mathcal{C},G_{\mathcal{C}}(Z)}\{z|\mathcal{C}_i, G_{\mathcal{C}_i}(Z_i)\}.$$

The outer expectation of (14.24), however, involves the distribution of the observed data $\{\mathcal{C}_i, G_{\mathcal{C}}(Z_i)\}$, which are generated from the truth, β_0. Therefore, the overall expectation is given as

$$\lambda(\beta,\beta_0) = \int \left\{ \int S^F(z,\beta_0) p_{Z|\mathcal{C},G_{\mathcal{C}}(Z)}(z|r,g_r,\beta) d\nu_{Z|\mathcal{C},G_{\mathcal{C}}(Z)}(z|r,g_r) \right\} \times p_{\mathcal{C},G_{\mathcal{C}}(Z)}(r,g_r,\beta_0) d\nu_{\mathcal{C},G_{\mathcal{C}}(Z)}(r,g_r). \tag{14.25}$$

Because

$$\begin{aligned} &p_{Z|\mathcal{C},G_{\mathcal{C}}(Z)}(z|r,g_r,\beta) p_{\mathcal{C},G_{\mathcal{C}}(Z)}(r,g_r,\beta) d\nu_{Z|\mathcal{C},G_{\mathcal{C}}(Z)}(z|r,g_r) d\nu_{\mathcal{C},G_{\mathcal{C}}(Z)}(r,g_r) \\ &\quad = p_{\mathcal{C},Z}(r,z,\beta) d\nu_{\mathcal{C},Z}(r,z), \end{aligned}$$

we can write (14.25) as

$$\int S^F(z,\beta_0) \left\{ \frac{p_{\mathcal{C},Z}(r,z,\beta)}{p_{\mathcal{C},G_{\mathcal{C}}(Z)}(r,g_r,\beta)} \right\} p_{\mathcal{C},G_{\mathcal{C}}(Z)}(r,g_r,\beta_0) d\nu_{\mathcal{C},Z}(r,z). \tag{14.26}$$

Taking the derivative of (14.26) with respect to β and evaluating at β_0, we obtain

$$\begin{aligned} \left.\frac{\partial\lambda(\beta,\beta_0)}{\partial\beta^T}\right|_{\beta=\beta_0} &= \int S^F(z,\beta_0) \left.\frac{\partial}{\partial\beta^T}\left\{ \frac{p_{\mathcal{C},Z}(r,z,\beta)}{p_{\mathcal{C},G_{\mathcal{C}}(Z)}(r,g_r,\beta)} \right\}\right|_{\beta=\beta_0} \\ &\qquad \times p_{\mathcal{C},G_{\mathcal{C}}(Z)}(r,g_r,\beta_0) d\nu_{\mathcal{C},Z}(r,z) \\ &= \int S^T(z,\beta_0) \left\{ \frac{p_{\mathcal{C},Z}(r,z,\beta_0)}{p_{\mathcal{C},G_{\mathcal{C}}(Z)}(r,g_r,\beta_0)} \right\} \left.\frac{\partial \log}{\partial\beta^T}\left\{ \frac{p_{\mathcal{C},Z}(r,z,\beta)}{p_{\mathcal{C},G_{\mathcal{C}}(Z)}(r,g_r,\beta)} \right\}\right|_{\beta=\beta_0} \\ &\qquad \times p_{\mathcal{C},G_{\mathcal{C}}(Z)}(r,g_r,\beta_0) d\nu_{\mathcal{C},Z}(r,z). \end{aligned} \tag{14.27}$$

Because the data are coarsened at random,

$$\frac{p_{\mathcal{C},Z}(r,z,\beta)}{p_{\mathcal{C},G_{\mathcal{C}}(Z)}(r,g_r,\beta)} = \frac{p_Z(z,\beta)}{p_{G_r(Z)}(g_r,\beta)}. \tag{14.28}$$

Therefore

$$\left.\frac{\partial \log}{\partial\beta^T}\left\{ \frac{p_{\mathcal{C},Z}(r,z,\beta)}{p_{\mathcal{C},G_{\mathcal{C}}(Z)}(r,g_r,\beta)} \right\}\right|_{\beta=\beta_0} = \frac{\partial \log p_Z(z,\beta_0)}{\partial\beta^T} - \frac{\partial \log p_{G_r(Z)}(g_r,\beta_0)}{\partial\beta^T}. \tag{14.29}$$

When data are coarsened at random, we showed in Lemma 7.2 that

$$\frac{\partial \log p_{G_r(Z)}(g_r, \beta_0)}{\partial \beta^T} = E\left\{S^F(Z)|\mathcal{C} = r, G_{\mathcal{C}}(Z) = g_r\right\} = S(r, g_r).$$

Therefore, (14.29) is equal to

$$S^{F^T}(z, \beta_0) - S^T(r, g_r, \beta_0).$$

Substituting this last result into (14.27) and rearranging some terms yields

$$\begin{aligned}&\int S^F(z, \beta_0)\left\{S^{F^T}(z, \beta_0) - S^T(r, g_r, \beta_0)\right\} p_{\mathcal{C},Z}(r, z, \beta_0) d\nu_{\mathcal{C},Z}(r, z) \\ &= E\left\{S^F(Z, \beta_0) S^{F^T}(Z, \beta_0)\right\} - E\left[S^F(Z, \beta_0) S^T\{\mathcal{C}, G_{\mathcal{C}}(Z), \beta_0\}\right]. \quad (14.30)\end{aligned}$$

Using (14.8), we obtain that

$$E\left\{S^F(Z, \beta_0) S^T(\mathcal{C}, G_{\mathcal{C}}(Z), \beta_0)\right\} = I(\beta_0).$$

Hence, (14.30) equals $I^F(\beta_0) - I(\beta_0)$, proving Lemma 14.2. □

Remark 6. The result of Lemma 14.2 can be deduced, with slight modification, from equation (6) on page 480 of Oakes (1999). The term $I^F(\beta_0) - I(\beta_0)$ is also referred to as the "missing information," as this is the information that is lost due to the coarsening of the data. □

Before giving the next lemma, we first give a short description of the notion of "stochastic equicontinuity."

Stochastic Equicontinuity

Consider the centered stochastic process

$$W_n(\beta) = n^{-1/2} \sum_{i=1}^{n} \left[m^{-1} \sum_{j=1}^{m} S^F\{Z_{ij}(\beta), \beta_0\} - \lambda(\beta, \beta_0) \right]$$

as a process in β, where $\lambda(\beta, \beta_0) = E\left[S^F\{Z_{ij}(\beta), \beta_0\}\right]$.

Using the theory of empirical processes (see van der Vaart and Wellner, 1996), under suitable regularity conditions, the process $W_n(\beta)$ converges weakly to a tight Gaussian process. When this is the case, we have stochastic equicontinuity, where, for every $\varepsilon, \eta > 0$, there exists a $\delta > 0$ and an n_0 such that

$$P\left\{ \sup_{\beta', \beta'': \|\beta' - \beta''\| < \delta} \|W_n(\beta') - W_n(\beta'')\| \geqslant \varepsilon \right\} \leq \eta$$

for all $n > n_0$, where "$\|\cdot\|$" denotes the usual Euclidean norm or distance.

Lemma 14.3. If stochastic equicontinuity holds, then

$$W_n(\hat{\beta}_n^I) - W_n(\beta_0) = o_p(1)$$

whenever $\hat{\beta}_n^I$ is a consistent estimator for β_0.

Proof. Choose arbitrary $\varepsilon, \eta > 0$. Then, by stochastic equicontinuity, there exists a $\delta(\varepsilon, \eta)$ and $n(\varepsilon, \eta)$ such that

$$P\left\{\sup_{\beta', \beta'' : \|\beta' - \beta''\| < \delta(\varepsilon,\eta)} \|W_n(\beta') - W_n(\beta'')\| \geqslant \varepsilon\right\} \leq \eta/2$$

for all $n > n(\varepsilon, \eta)$.

By consistency, there exists for every δ and η an $n^*(\delta, \eta)$ such that

$$P(\|\hat{\beta}_n^I - \beta_0\| \geqslant \delta) \leq \eta/2 \text{ for all } n > n^*(\delta, \eta).$$

Note that the event

$$\left\{\|\hat{\beta}_n^I - \beta_0\| < \delta(\varepsilon,\eta) \text{ and } \sup_{\beta', \beta'' : \|\beta' - \beta''\| < \delta(\varepsilon,\eta)} \|W_n(\beta') - W_n(\beta'')\| < \varepsilon\right\}$$
$$\subseteq \left\{\|W_n(\hat{\beta}_n^I) - W_n(\beta_0)\| < \varepsilon\right\}.$$

By taking complements, we obtain

$$\begin{aligned}
&P\{\|W_n(\hat{\beta}_n^I) - W_n(\beta_0)\| \geqslant \varepsilon\} \\
&\leq P\left[\{\|\hat{\beta}_n^I - \beta_0\| \geqslant \delta(\varepsilon,\eta)\}\right. \\
&\qquad \left. \cup \left\{\sup_{\beta', \beta'' : \|\beta' - \beta''\| < \delta(\varepsilon,\eta)} \|W_n(\beta') - W_n(\beta'')\| \geqslant \varepsilon\right\}\right] \\
&\leq P\{\|\hat{\beta}_n^I - \beta_0\| \geqslant \delta(\varepsilon,\eta)\} \\
&\qquad + P\left\{\sup_{\beta', \beta'' : \|\beta' - \beta''\| < \delta(\varepsilon,\eta)} \|W_n(\beta') - W_n(\beta'')\| \geqslant \varepsilon\right\}.
\end{aligned}$$

By choosing $n > \max[n(\varepsilon, \eta), n^*\{\delta(\varepsilon, \eta), \eta\}]$, we obtain

$$P\{\|W_n(\hat{\beta}_n^I) - W_n(\beta_0)\| \geqslant \varepsilon\} \leq \eta/2 + \eta/2 = \eta. \quad \square$$

Lemma 14.4. The statistic (14.20), namely

$$n^{-1/2} \sum_{i=1}^{n} \left[m^{-1} \sum_{j=1}^{m} S^F\{Z_{ij}(\hat{\beta}_n^I), \beta_0\} - m^{-1} \sum_{j=1}^{m} S^F\{Z_{ij}(\beta_0), \beta_0\} \right],$$

is equal to

$$n^{1/2}\{\lambda(\hat{\beta}_n^I, \beta_0) - \lambda(\beta_0, \beta_0)\} + o_p(1).$$

Proof. This follows directly as a consequence of Lemma 14.3, which states that

$$W_n(\hat{\beta}_n^I) - W_n(\beta_0) = o_p(1),$$

after some simple rearrangement of terms. □

Lemma 14.5.

$$n^{1/2}\{\lambda(\hat{\beta}_n^I, \beta_0) - \lambda(\beta_0, \beta_0)\} = \left.\frac{\partial \lambda(\beta, \beta_0)}{\partial \beta^T}\right|_{\beta=\beta_0} n^{1/2}(\hat{\beta}_n^I - \beta_0) + o_p(1).$$

Proof. Follows from a standard delta method argument. □

We now return to the proof of Theorem 14.3.

Proof of Theorem 14.3
Lemmas 14.4 and 14.5 imply that (14.20) is equal to

$$\left.\frac{\partial \lambda(\beta, \beta_0)}{\partial \beta^T}\right|_{\beta=\beta_0} n^{1/2}(\hat{\beta}_n^I - \beta_0) + o_p(1).$$

The proof is complete because of Lemma 14.2, which states that

$$\left.\frac{\partial \lambda(\beta, \beta_0)}{\partial \beta^T}\right|_{\beta=\beta_0} = I^F(\beta_0) - I(\beta_0),$$

and the definition of the influence function of $\hat{\beta}_n^I$,

$$n^{1/2}(\hat{\beta}_n^I - \beta_0) = n^{-1/2} \sum_{i=1}^{n} q\{\mathcal{C}_i, G_{\mathcal{C}_i}(Z_i)\} + o_p(1). \quad \square$$

14.4 Asymptotic Distribution of the Multiple-Imputation Estimator

We are now in a position to derive the asymptotic distribution of $\hat{\beta}_n^*$, which we present in the following theorem:

Theorem 14.4. Letting $\hat{\beta}_n^*$ denote the multiple imputation estimator and denoting the i-th influence function of the initial estimator by (14.13), i.e.,

$$q\{\mathcal{C}_i, G_{\mathcal{C}_i}(Z_i)\} = \varphi_{\text{eff}}\{\mathcal{C}_i, G_{\mathcal{C}_i}(Z_i)\} + h\{\mathcal{C}_i, G_{\mathcal{C}_i}(Z_i)\},$$

where $\varphi_{\text{eff}}\{\mathcal{C}_i, G_{\mathcal{C}_i}(Z_i)\}$ is defined by (14.11), then

$$n^{1/2}(\hat{\beta}_n^* - \beta_0) \xrightarrow{\mathcal{D}} N\left(0, \Sigma^*\right),$$

where Σ^* is equal to

$$\{I(\beta_0)\}^{-1} + m^{-1}\left\{I^F(\beta_0)\right\}^{-1}\left\{I^F(\beta_0) - I(\beta_0)\right\}\left\{I^F(\beta_0)\right\}^{-1} + \left\{I^F(\beta_0)\right\}^{-1}\left\{I^F(\beta_0) - I(\beta_0)\right\} \text{ var}\left[h\{\mathcal{C}_i, G_{\mathcal{C}_i}(Z_i)\}\right] \left\{I^F(\beta_0) - I(\beta_0)\right\}\left\{I^F(\beta_0)\right\}^{-1}. \tag{14.31}$$

Proof. As a result of Lemma14.1, Theorem 14.3, and equations (14.19) and (14.20), we have shown that

$$n^{1/2}(\hat{\beta}_n^* - \beta_0) = n^{-1/2}\sum_{i=1}^{n}\left\{I^F(\beta_0)\right\}^{-1}\left(\left[m^{-1}\sum_{j=1}^{m} S^F\{Z_{ij}(\beta_0), \beta_0\}\right] + \left\{I^F(\beta_0) - I(\beta_0)\right\} q\{\mathcal{C}_i, G_{\mathcal{C}_i}(Z_i)\}\right) + o_p(1). \tag{14.32}$$

This is a key result, as we have identified the influence function for the multiple-imputation estimator as the i-th element in the summand in (14.32). Asymptotic normality follows from the central limit theorem. The asymptotic variance of $n^{1/2}(\hat{\beta}_n^* - \beta_0)$ is the variance of the influence function, which we will now derive in a series of steps.

Toward that end, we first compute

$$\text{var}\left[m^{-1}\sum_{j=1}^{m} S^F\{Z_{ij}(\beta_0), \beta_0\}\right]. \tag{14.33}$$

Computing (14.33)
Using the law for iterated conditional variance, (14.33) can be written as

$$E\left(\text{var}\left[m^{-1}\sum_{j=1}^{m} S^F\{Z_{ij}(\beta_0), \beta_0\}\middle|\mathcal{C}_i, G_{\mathcal{C}_i}(Z_i)\right]\right) \tag{14.34}$$

$$+ \text{var}\left(E\left[m^{-1}\sum_{j=1}^{m} S^F\{Z_{ij}(\beta_0), \beta_0\}\middle|\mathcal{C}_i, G_{\mathcal{C}_i}(Z_i)\right]\right). \tag{14.35}$$

Because, conditional on $\{\mathcal{C}_i, G_{\mathcal{C}_i}(Z_i)\}$, the $Z_{ij}(\beta_0), j = 1, \ldots, m$ are independent draws from the conditional distribution with conditional density $p_{Z|\mathcal{C},G_{\mathcal{C}}(Z)}\{z|\mathcal{C}_i, G_{\mathcal{C}_i}(Z_i), \beta_0\}$, this means that the conditional variance

$$\text{var}\left[m^{-1}\sum_{j=1}^{m} S^F\{Z_{ij}(\beta_0), \beta_0\}\middle|\mathcal{C}_i, G_{\mathcal{C}_i}(Z_i)\right]$$

is equal to

$$m^{-1} \text{ var } \left\{S^F(Z_i, \beta_0)|\mathcal{C}_i, G_{\mathcal{C}_i}(Z_i)\right\}.$$

Therefore, (14.34) is equal to

$$m^{-1} \; E\left[\text{var } \left\{S^F(Z_i, \beta_0)|\mathcal{C}_i, G_{\mathcal{C}_i}(Z_i)\right\}\right].$$

In equation (14.6), we showed

$$E\left[\text{var } \left\{S^F(Z_{ij}\beta_0)|\mathcal{C}_i, G_{\mathcal{C}_i}(Z_i)\right\}\right] = I^F(\beta_0) - I(\beta_0).$$

Consequently,

$$(14.34) = m^{-1}\left\{I^F(\beta_0) - I(\beta_0)\right\}. \tag{14.36}$$

Similar logic can be used to show

$$\begin{aligned} & E\left[m^{-1}\sum_{j=1}^{m} S^F\{Z_{ij}, (\beta_0), \beta_0\}\middle|\mathcal{C}_i, G_{\mathcal{C}_i}(Z_i)\right] \\ & = E\left\{S^F(Z_i, \beta_0)|\mathcal{C}_i, G_{\mathcal{C}_i}(Z_i)\right\} \\ & = S\{\mathcal{C}_i, G_{\mathcal{C}_i}(Z_i)\}. \end{aligned} \tag{14.37}$$

Therefore (14.35) is equal to

$$\text{var } [S\{\mathcal{C}_i, G_{\mathcal{C}_i}(Z_i)\}] = I(\beta_0). \tag{14.38}$$

Combining (14.36) and (14.38) gives us that

$$(14.33) = m^{-1}\left\{I^F(\beta_0) - I(\beta_0)\right\} + I(\beta_0). \tag{14.39}$$

Next we compute the covariance matrix

$$E\left(\left[m^{-1}\sum_{j=1}^{m} S^F\{Z_{ij}(\beta_0), \beta_0\}\right]\left[\left\{I^F(\beta_0) - I(\beta_0)\right\} q\{\mathcal{C}_i, G_{\mathcal{C}_i}(Z_i)\}\right]^T\right). \tag{14.40}$$

Computing (14.40)
Expression (14.40) can be written as

$$m^{-1}\sum_{j=1}^{m} E\left[S^F\{Z_{ij}(\beta_0), \beta_0\}q^T\{\mathcal{C}_i, G_{\mathcal{C}_i}(Z_i)\}\right]\left\{I^F(\beta_0) - I(\beta_0)\right\},$$

where

$$\begin{aligned} & E\left[S^F\{Z_{ij}(\beta_0), \beta_0\}q^T\{\mathcal{C}_i, G_{\mathcal{C}_i}(Z_i)\}\right] \\ & \quad = E\left(E\left[S^F\{Z_{ij}(\beta_0), \beta_0\}q^T\{\mathcal{C}_i, G_{\mathcal{C}_i}(Z_i)\}|\mathcal{C}_i, G_{\mathcal{C}_i}(Z_i)\right]\right) \\ & \quad E\Bigg(\underbrace{E\left[S^F\{Z_{ij}(\beta_0), \beta_0\}|\mathcal{C}_i, G_{\mathcal{C}_i}(Z_i)\right]}_{\substack{\text{we showed in (14.37) that} \\ \text{this equals } S\{\mathcal{C}_i, G_{\mathcal{C}_i}(Z_i)\}}} q^T\{\mathcal{C}_i, G_{\mathcal{C}_i}(Z_i)\}\Bigg) \\ & \quad = E\left[S\{\mathcal{C}_i, G_{\mathcal{C}_i}(Z_i)\}q^T\{\mathcal{C}_i, G_{\mathcal{C}_i}(Z_i)\}\right] \\ & \quad = I^{q\times q} \text{ (identity matrix) by (14.12).} \end{aligned}$$

Therefore

$$(14.40) = I^F(\beta_0) - I(\beta_0). \tag{14.41}$$

Finally,

$$\begin{aligned}
&\text{var}\left[\{I^F(\beta_0) - I(\beta_0)\}q\{\mathcal{C}_i, G_{\mathcal{C}_i}(Z_i)\}\right] \\
&= \left\{I^F(\beta_0) - I(\beta_0)\right\} \text{ var}\left[q\{\mathcal{C}_i, G_{\mathcal{C}_i}(Z_i)\}\right]\left\{I^F(\beta_0) - I(\beta_0)\right\} \qquad (14.42) \\
&= \{I^F(\beta_0) - I(\beta_0)\}\Big(I^{-1}(\beta_0) + \text{var}\left[h\{\mathcal{C}_i, G_{\mathcal{C}_i}(Z_i)\}\right]\Big)\{I^F(\beta_0) - I(\beta_0)\}, \qquad (14.43)
\end{aligned}$$

where (14.43) follows from (14.14).

Using (14.39), (14.41), and (14.43), and after some straightforward algebraic manipulation, we are able to derive the variance matrix for the influence function of $\hat{\beta}_n^*$ (i.e., the i-th summand in (14.32)) to be

$$\begin{aligned}
\Sigma^* = &\{I(\beta_0)\}^{-1} \\
&+ m^{-1}\left\{I^F(\beta_0)\right\}^{-1}\left\{I^F(\beta_0) - I(\beta_0)\right\}\left\{I^F(\beta_0)\right\}^{-1} \\
&+ \left\{I^F(\beta_0)\right\}^{-1}\left\{I^F - I(\beta_0)\right\}\text{var}[h\{\mathcal{C}_i, G_{\mathcal{C}_i}(Z_i)\}]\left\{I^F(\beta_0)\right. \\
&\left.-I(\beta_0)\right\}\left\{I^F(\beta_0)\right\}^{-1}.
\end{aligned}$$

This completes the proof of Theorem 14.4. □

Examining (14.32), we conclude that the influence function for the multiple-imputation estimator, with m imputation draws, is equal to

$$\{I^F(\beta_0)\}^{-1}\left(\left[m^{-1}\sum_{j=1}^{m} S^F\{Z_j(\beta_0), \beta_0\}\right] + \{I^F(\beta_0) - I(\beta_0)\}q\{\mathcal{C}, G_{\mathcal{C}}(Z)\}\right), \tag{14.44}$$

where $Z_j(\beta_0)$ denotes the j-th random draw from the conditional distribution of $p_{Z|\mathcal{C},G_{\mathcal{C}}(Z)}\{z|\mathcal{C}, G_{\mathcal{C}}(Z), \beta_0\}$. As a consequence of the law of large numbers, we would expect, under suitable regularity conditions, that

$$m^{-1}\sum_{j=1}^{m} S^F\{Z_j(\beta_0), \beta_0\} \xrightarrow{P} E\{S^F(Z)|\mathcal{C}, G_{\mathcal{C}}(Z)\} = S\{\mathcal{C}, G_{\mathcal{C}}(Z)\}$$

as $m \to \infty$. Specifically, in order to prove that

$$m^{-1}\sum_{j=1}^{m} S^F\{Z_j(\beta_0), \beta_0\} - S\{\mathcal{C}, G_{\mathcal{C}}(Z)\} \xrightarrow{P} 0,$$

it suffices to show that

$$E\left[m^{-1}\sum_{j=1}^{m} S^F\{Z_j(\beta_0),\beta_0\} - S\{\mathcal{C}, G_{\mathcal{C}}(Z)\}\right]^2 \xrightarrow{m\to\infty} 0. \tag{14.45}$$

Computing the expectation above by first conditioning on $\{\mathcal{C}, G_{\mathcal{C}}(Z)\}$, we obtain that

$$E\left[m^{-1}\sum_{j=1}^{m} S^F\{Z_j(\beta_0),\beta_0\} - S\{\mathcal{C}, G_{\mathcal{C}}(Z)\}\right]^2 = E\left[m^{-1}\mathrm{var}\left\{S^F(Z)|\mathcal{C}, G_{\mathcal{C}}(Z)\right\}\right].$$

So, for example, if the conditional variance $\mathrm{var}\left\{S^F(Z)|\mathcal{C}, G_{\mathcal{C}}(Z)\right\}$ is bounded almost surely, then (14.45) would hold. Consequently, as $m \to \infty$, the influence function of the multiple-imputation estimator (14.44) converges to

$$\{I^F(\beta_0)\}^{-1}\left[S\{\mathcal{C}, G_{\mathcal{C}}(Z)\} + \{I^F(\beta_0) - I(\beta_0)\}q\{\mathcal{C}, G_{\mathcal{C}}(Z)\}\right]. \tag{14.46}$$

As we will now demonstrate, the limit, as $m \to \infty$, of the multiple-imputation estimator is related to the expectation maximization (EM) algorithm. The EM algorithm, first introduced by Dempster, Laird, and Rubin (1977) and studied carefully by Wu (1983), is a numerical method for finding the MLE that is especially useful with missing-data problems. The EM algorithm is an iterative procedure where the expectation of the full-data log-likelihood is computed with respect to a current value of a parameter estimate and then an updated estimate for the parameter is obtained by maximizing the expected full-data log-likelihood. Hence, if $\hat{\beta}_n^{(j)}$ is the value of the estimator at the j-th iteration, then the $(j+1)$-th value of the estimator is obtained by solving the estimating equation

$$\sum_{i=1}^{n} E\left\{S^F(Z, \hat{\beta}_n^{(j+1)})|\mathcal{C}_i, G_{\mathcal{C}_i}(Z_i), \hat{\beta}_n^{(j)}\right\} = 0. \tag{14.47}$$

We now show the relationship between the EM algorithm and the multiple-imputation estimator in the following theorem.

Theorem 14.5. Let the one-step updated EM estimator, $\hat{\beta}_n^{\mathrm{EM}}$, be the solution to

$$\sum_{i=1}^{n} E\left\{S^F(Z, \hat{\beta}_n^{\mathrm{EM}})|\mathcal{C}_i, G_{\mathcal{C}_i}(Z_i), \hat{\beta}_n^{I}\right\} = 0, \tag{14.48}$$

where $\hat{\beta}_n^I$ is the initial estimator for β, whose i-th influence function is $q\{\mathcal{C}_i, G_{\mathcal{C}_i}(Z_i)\}$, which was used for imputation. The influence function of $\hat{\beta}_n^{\mathrm{EM}}$ is identically equal to (14.46), the limiting influence function for the multiple-imputation estimator as $m \to \infty$.

Proof. A simple expansion of $\hat{\beta}_n^{\mathrm{EM}}$ about β_0, keeping $\hat{\beta}_n^I$ fixed, in (14.48) yields

$$0 = n^{-1/2}\sum_{i=1}^{n} E\left\{ S^F(Z,\beta_0)|\mathcal{C}_i, G_{\mathcal{C}_i}(Z_i), \hat{\beta}_n^I \right\}$$
$$+ \left[n^{-1}\sum_{i=1}^{n} E\left\{ \frac{\partial S^F(Z,\beta_0)}{\partial \beta^T} \middle| \mathcal{C}_i, G_{\mathcal{C}_i}(Z_i), \beta_0 \right\} \right] n^{1/2}(\hat{\beta}_n^{\mathrm{EM}} - \beta_0) + o_p(1).$$

Because

$$n^{-1}\sum_{i=1}^{n} E\left\{ \frac{\partial S^F(Z,\beta_0)}{\partial \beta^T} \middle| \mathcal{C}_i, G_{\mathcal{C}_i}(Z_i), \beta_0 \right\}$$
$$\xrightarrow{P} E\left[E\left\{ \frac{\partial S^F(Z,\beta_0)}{\partial \beta^T} \middle| \mathcal{C}_i, G_{\mathcal{C}_i}(Z_i), \beta_0 \right\} \right]$$
$$= E\left\{ \frac{\partial S^F(Z,\beta_0)}{\partial \beta^T} \right\} = -I^F(\beta_0),$$

we obtain

$$n^{1/2}(\hat{\beta}_n^{\mathrm{EM}} - \beta_0) = \{I^F(\beta_0)\}^{-1} n^{-1/2}\sum_{i=1}^{n} E\left\{ S^F(Z,\beta_0)|\mathcal{C}_i, G_{\mathcal{C}_i}(Z_i), \hat{\beta}_n^I \right\} + o_p(1). \tag{14.49}$$

An expansion of $\hat{\beta}_n^I$ about β_0 on the right-hand side of (14.49) yields

$$n^{-1/2}\sum_{i=1}^{n} E\left\{ S^F(Z,\beta_0)|\mathcal{C}_i, G_{\mathcal{C}_i}(Z_i), \hat{\beta}_n^I \right\}$$
$$= n^{-1/2}\sum_{i=1}^{n} E\left\{ S^F(Z,\beta_0)|\mathcal{C}_i, G_{\mathcal{C}_i}(Z_i), \beta_0 \right\} \tag{14.50}$$
$$+ \left[n^{-1}\sum_{i=1}^{n} \frac{\partial}{\partial \beta^T} E\{S^F(Z,\beta_0)|\mathcal{C}_i, G_{\mathcal{C}_i}(Z_i), \beta\}\Big|_{\beta=\beta_0} \right] n^{1/2}(\hat{\beta}_n^I - \beta_0) + o_p(1).$$

The sample average is

$$\left[n^{-1}\sum_{i=1}^{n} \frac{\partial}{\partial \beta^T} E\{S^F(Z,\beta_0)|\mathcal{C}_i, G_{\mathcal{C}_i}(Z_i), \beta\}\Big|_{\beta=\beta_0} \right]$$
$$\xrightarrow{P} E\left[\frac{\partial}{\partial \beta^T} E\{S^F(Z,\beta_0)|\mathcal{C}_i, G_{\mathcal{C}_i}(Z_i), \beta\}\Big|_{\beta=\beta_0} \right], \tag{14.51}$$

where

$$\frac{\partial}{\partial \beta^T} E\{S^F(Z,\beta_0)|\mathcal{C}_i, G_{\mathcal{C}_i}(Z_i), \beta\}\Big|_{\beta=\beta_0}$$
$$= \frac{\partial}{\partial \beta^T}\left[\frac{\int_{z:G_{\mathcal{C}_i}(z)=G_{\mathcal{C}_i}(Z_i)} S^F(z,\beta_0) p(z,\beta) d\nu(z)}{\int_{z:G_{\mathcal{C}_i}(z)=G_{\mathcal{C}_i}(Z_i)} p(z,\beta) d\nu(z)} \right]\Bigg|_{\beta=\beta_0}. \tag{14.52}$$

Straightforward calculations yield that the derivative in (14.52) is equal to $\text{var}\{S^F(Z)|\mathcal{C}_i, G_{\mathcal{C}_i}(Z_i)\}$ and hence (14.51) is equal to

$$E\left[\text{var}\{S^F(Z)|\mathcal{C}_i, G_{\mathcal{C}_i}(Z_i)\}\right] = I^F(\beta_0) - I(\beta_0), \tag{14.53}$$

where the last equality follows from (14.6). Using (14.4), we obtain that (14.50) is equal to

$$n^{-1/2}\sum_{i=1}^{n} S\{\mathcal{C}_i, G_{\mathcal{C}_i}(Z_i)\}. \tag{14.54}$$

By the definition of an influence function,

$$n^{1/2}(\hat{\beta}_n^I - \beta_0) = n^{-1/2}\sum_{i=1}^{n} q\{\mathcal{C}_i, G_{\mathcal{C}_i}(Z_i)\} + o_p(1). \tag{14.55}$$

Combining equations (14.49)–(14.55), we conclude that the influence function of $\hat{\beta}_n^{\text{EM}}$ is equal to (14.46), thus concluding the proof. □

In (14.13), we expressed the influence function of $\hat{\beta}_n^I$ as the sum of

$$\{I(\beta_0)\}^{-1}S\{\mathcal{C}, G_{\mathcal{C}}(Z)\} + h\{\mathcal{C}, G_{\mathcal{C}}(Z)\}, \tag{14.56}$$

where $h\{\mathcal{C}, G_{\mathcal{C}}(Z)\}$ is orthogonal to $S\{\mathcal{C}, G_{\mathcal{C}}(Z)\}$; i.e., $E(hS^T) = 0^{q\times q}$. Substituting (14.56) for $q\{\mathcal{C}, G_{\mathcal{C}}(Z)\}$ in (14.46), we obtain another representation of the influence function for $\hat{\beta}_n^{\text{EM}}$ as

$$\varphi_{\text{eff}}\{\mathcal{C}, G_{\mathcal{C}}(Z)\} + \{I^F(\beta_0)\}^{-1}\{I^F(\beta_0) - I(\beta_0)\}h\{\mathcal{C}, G_{\mathcal{C}}(Z)\}, \tag{14.57}$$

where

$$\varphi_{\text{eff}}\{\mathcal{C}, G_{\mathcal{C}}(Z)\} = \{I(\beta_0)\}^{-1}S\{\mathcal{C}, G_{\mathcal{C}}(Z)\}$$

is the efficient observed-data influence function.

We note that the influence function in (14.57) is that for a one-step updated EM estimator that started with the initial estimator $\hat{\beta}_n^I$. Using similar logic, we would conclude that the EM algorithm after j iterations would yield an estimator $\hat{\beta}_n^{\text{EM}(j)}$ with influence function

$$\varphi_{\text{eff}}\{\mathcal{C}, G_{\mathcal{C}}(Z)\} + J^j h\{\mathcal{C}, G_{\mathcal{C}}(Z)\},$$

where J is the $q \times q$ matrix $\{I^F(\beta_0)\}^{-1}\{I^F(\beta_0) - I(\beta_0)\}$. For completeness, we now show that J^j will converge to zero (in the sense that all elements in the matrix will converge to zero) as j goes to infinity, thus demonstrating that the EM algorithm will converge to the efficient observed-data estimator. This proof is taken from Lemma A.1 of Wang and Robins (1998).

Proof. Both $I^F(\beta_0)$ and $\{I^F(\beta_0) - I(\beta_0)\}$ are symmetric positive-definite matrices. From Rao (1973, p. 41), we can express $I^F(\beta_0) = R^T R$ and $\{I^F(\beta_0) - I(\beta_0)\} = R^T \Lambda R$, where R is a nonsingular matrix and Λ is a diagonal matrix. Moreover, because

$$I(\beta_0) = I^F(\beta_0) - \{I^F(\beta_0) - I(\beta_0)\} = R^T(I^{q\times q} - \Lambda)R$$

is positive definite, this implies that all the elements on the diagonal of Λ must be strictly less than 1. Consequently,

$$J^j = (R)^{-1}\Lambda^j R$$

will converge to zero as $j \to \infty$. □

Remarks regarding the asymptotic variance

1. The asymptotic variance of the observed-data efficient estimator for β is $\{I(\beta_0)\}^{-1}$. Because (14.31) is greater than $\{I(\beta_0)\}^{-1}$, this implies that the multiple-imputation estimator is not fully efficient.
2. Even if the initial estimator $\hat{\beta}_n^I$ is efficient (i.e., $h\{\mathcal{C}_i, G_{\mathcal{C}_i}(Z_i)\} = 0$), the resulting multiple-imputation estimator loses some efficiency due to the contribution to the variance of the second term of (14.31), but this loss of efficiency vanishes as $m \to \infty$.
3. For any initial estimator $\hat{\beta}_n^I$, as the number of multiple-imputation draws "m" gets larger, the asymptotic variance decreases (from the contribution of the second term of (14.31)). Thus the resulting estimator becomes more efficient with increasing m.
4. We showed in Theorem 14.5 that as m goes to infinity, the resulting multiple-imputation estimator is equivalent to a one-step update of an EM algorithm. This suggests that one strategy is that after m imputations, we start the process again, now using $\hat{\beta}_n^*$ as the initial estimator. By continuing this process, we can iterate toward full efficiency. To implement such a strategy, m must be chosen sufficiently large to ensure that the new estimator is more efficient than the previous one. As we converge toward efficiency, m must get increasingly large.
5. Other algebraically equivalent representations of the asymptotic variance, which we will use later, are given by

$$\begin{aligned}&\left\{I^F(\beta_0)\right\}^{-1}\left[I(\beta_0) + m^{-1}\left\{I^F(\beta_0) - I(\beta_0)\right\} + 2\left\{I^F(\beta_0) - I(\beta_0)\right\}\right.\\&\quad\left. + \left\{I^F(\beta_0) - I(\beta_0)\right\}\mathrm{var}[q\{\mathcal{C}_i, G_{\mathcal{C}_i}(Z_i)\}]\left\{I^F(\beta_0) - I(\beta_0)\right\}\right]\\&\quad\left\{I^F(\beta_0)\right\}^{-1}\end{aligned} \tag{14.58}$$

and

$$\left\{I^F(\beta_0)\right\}^{-1} + \left(\frac{m+1}{m}\right)\left\{I^F(\beta_0)\right\}^{-1}\left\{I^F(\beta_0) - I(\beta_0)\right\}\left\{I^F(\beta_0)\right\}^{-1}$$
$$+ \left\{I^F(\beta_0)\right\}^{-1}\left\{I^F(\beta_0) - I(\beta_0)\right\}\mathrm{var}[q\{\mathcal{C}_i, G_{\mathcal{C}_i}(Z_i)\}]$$
$$\left\{I^F(\beta_0) - I(\beta_0)\right\}\left\{I^F(\beta_0)\right\}^{-1}. \tag{14.59}$$

14.5 Estimating the Asymptotic Variance

In this section, we will consider estimators for the asymptotic variance of the frequentist (type B) multiple-imputation estimator. Although Rubin (1987) refers to this type of imputation as improper and does not advocate using his intuitive variance estimator in such cases, my experience has been that, in practice, many statisticians do not distinguish between proper and improper imputation and will often use Rubin's variance formula. Therefore, we begin this section by studying the properties of Rubin's variance formula when used with frequentist multiple imputation.

Rubin suggested the following estimator for the asymptotic variance of $n^{1/2}(\hat{\beta}_n^* - \beta_0)$:

$$m^{-1}\sum_{j=1}^{m}\left[n^{-1}\sum_{i=1}^{n}\frac{-\partial S^F\{Z_{ij}(\hat{\beta}_n^I), \hat{\beta}_{nj}^*\}}{\partial \beta^T}\right]^{-1}$$
$$+ \left(\frac{m+1}{m}\right) n\sum_{j=1}^{m}\frac{(\hat{\beta}_{nj}^* - \hat{\beta}_n^*)(\hat{\beta}_{nj}^* - \hat{\beta}_n^*)^T}{m-1}. \tag{14.60}$$

It is easy to see that the first term in (14.60) converges in probability to

$$E\left\{-\frac{\partial S(Z,\beta_0)}{\partial \beta^T}\right\} = \{I^F(\beta_0)\}^{-1}.$$

Results regarding the second term of (14.60) are given in the following theorem.

Theorem 14.6.

$$E\left\{n\sum_{j=1}^{m}\left(\hat{\beta}_{nj}^* - \hat{\beta}_n^*\right)\left(\hat{\beta}_{nj}^* - \hat{\beta}_n^*\right)^T\right\} \xrightarrow{n\to\infty}$$
$$(m-1)\left\{I^F(\beta_0)\right\}^{-1}\left\{I^F(\beta_0) - I(\beta_0)\right\}\left\{I^F(\beta_0)\right\}^{-1}.$$

Proof. Examining the influence function of $n^{1/2}(\hat{\beta}_n^* - \beta_0)$, derived in (14.32), we conclude that

$$
\begin{aligned}
n^{1/2}(\hat{\beta}^*_{nj} - \hat{\beta}^*_n) &= n^{\frac{1}{2}}(\hat{\beta}^*_{nj} - \beta_0) - n^{1/2}(\hat{\beta}^*_n - \beta_0) \\
&= n^{-1/2} \sum_{i=1}^{n} \left\{ I^F(\beta_0) \right\}^{-1} \left[S^F\{Z_{ij}(\beta_0), \beta_0\} - \bar{S}^F_i(\beta_0) \right] \\
&\quad + o_p(1),
\end{aligned}
$$

where

$$
\bar{S}^F_i(\beta_0) = m^{-1} \sum_{j=1}^{m} S^F\{Z_{ij}(\beta_0), \beta_0\}.
$$

(Notice that the terms involving $q\{\mathcal{C}_i, G_{\mathcal{C}_i}(Z_i)\}$ in (14.32) cancel out.) Therefore,

$$
\begin{aligned}
& n(\hat{\beta}^*_{nj} - \hat{\beta}^*_n)(\hat{\beta}^*_{nj} - \hat{\beta}^*_n)^T = \\
& n^{-1} \left(\sum_{i=1}^{n} \left\{ I^F(\beta_0) \right\}^{-1} \left[S^F\{Z_{ij}(\beta_0), \beta_0\} - \bar{S}^F_i(\beta_0) \right] \right) \\
& \times \left(\sum_{i=1}^{n} \left[S^F \left\{ Z_{ij}(\beta_0), \beta_0 \right\} - \bar{S}^F_i(\beta_0) \right]^T \left\{ I^F(\beta_0) \right\}^{-1} \right) + o_p(1).
\end{aligned}
$$

Because the quantities inside the two sums above are independent across $i = 1, \ldots, n$, we obtain

$$
\begin{aligned}
& E\{n(\hat{\beta}^*_{nj} - \hat{\beta}^*_n)(\hat{\beta}^*_{nj} - \hat{\beta}^*_n)^T\} \to \\
& \quad E \left(\left\{ I^F(\beta_0) \right\}^{-1} \left[S^F\{Z_{ij}(\beta_0), \beta_0\} - \bar{S}^F_i(\beta_0) \right] \right. \\
& \quad \left. \times \left[S^F\{Z_{ij}(\beta_0), \beta_0\} - \bar{S}^F_i(\beta_0) \right]^T \left\{ I^F(\beta_0) \right\}^{-1} \right)
\end{aligned}
$$

and

$$
\begin{aligned}
& E \left\{ n \sum_{j=1}^{m} (\hat{\beta}^*_{nj} - \hat{\beta}^*_{nj})(\hat{\beta}^*_{nj} - \hat{\beta}^*_n)^T \right\} \to \\
& \quad \left\{ I^F(\beta_0) \right\}^{-1} E \left(\sum_{j=1}^{m} \left[S^F\{Z_{ij}(\beta_0), \beta_0\} - \bar{S}^F_i(\beta_0) \right] \right. \\
& \quad \left. \left[S^F\{Z_{ij}(\beta_0), \beta_0\} - \bar{S}^F_i(\beta_0) \right]^T \right) \times \left\{ I^F(\beta_0) \right\}^{-1}. \qquad (14.61)
\end{aligned}
$$

The expectation in (14.61) is evaluated by

$$E\left(\sum_{j=1}^{m}\left[S^F\{Z_{ij}(\beta_0),\beta_0\}-\bar{S}_i^F(\beta_0)\right]\left[S_\beta^F\{Z_{ij}(\beta_0),\beta_0\}-\bar{S}_i^F(\beta_0)\right]^T\right)$$
$$=E\sum_{j=1}^{m}\left[S^F\{Z_{ij}(\beta_0),\beta_0\}S^{F^T}\{Z_{ij}(\beta_0),\beta_0\}\right] \tag{14.62}$$
$$-\frac{1}{m}E\sum_{j=1}^{m}\sum_{j'=1}^{m}\left[S^F\{Z_{ij}(\beta_0),\beta_0\}S^{F^T}\{Z_{ij'}(\beta_0),\beta_0\}\right].$$

When $j=j'$,

$$E[S^F\{Z_{ij}(\beta_0),\beta_0\}S^{F^T}\{Z_{ij'}(\beta_0),\beta_0\}]=E\left[S^F(Z_i,\beta_0)S^{F^T}(Z_i,\beta_0)\right]=I^F(\beta_0),$$

whereas when $j\neq j'$,

$$\begin{aligned}&E\left[S^F\{Z_{ij}(\beta_0),\beta_0\}S^{F^T}\{Z_{ij'}(\beta_0),\beta_0\}\right]\\&=\mathrm{cov}\left[S^F\{Z_{ij}(\beta_0),\beta_0\},S^F\{Z_{ij'}(\beta_0),\beta_0\}\right]\\&=E\left(\mathrm{cov}\left[S^F\{Z_{ij}(\beta_0),\beta_0\},S^F\{Z_{ij'}(\beta_0),\beta_0\}|\mathcal{C}_i,G_{\mathcal{C}_i}(Z_i)\right]\right)\\&\quad+\mathrm{cov}\left(E\left[S^F\{Z_{ij}(\beta_0),\beta_0\}|\mathcal{C}_i,G_{\mathcal{C}_i}(Z_i)\right],E\left[S^F\{Z_{ij'}(\beta_0),\beta_0\}|\mathcal{C}_i,G_{\mathcal{C}_i}(Z_i)\right]\right).\end{aligned} \tag{14.63}$$

Because, conditional on $\mathcal{C}_i,G_{\mathcal{C}_i}(Z_i)$, the $Z_{ij}(\beta_0)$ are independent draws for different j, from the conditional density $p_{Z|\mathcal{C},G_{\mathcal{C}}(Z)}\{z|\mathcal{C}_i,G_{\mathcal{C}_i}(Z_i)\}$, this means that the first term of (14.63) (conditional covariance) is zero. Because $E[S^F\{Z_{ij}(\beta_0),\beta_0\}|\mathcal{C}_i,G_{\mathcal{C}_i}(Z_i)]=S\{\mathcal{C}_i,G_{\mathcal{C}_i}(Z_i)\}$ for all $j=1,\ldots,m$, then the second term of (14.63) is

$$\mathrm{var}[S\{\mathcal{C}_i,G_{\mathcal{C}_i}(Z_i)\}]=I(\beta_0).$$

Thus, from these results, we obtain that (14.62) equals

$$\begin{aligned}&mI^F(\beta_0)-\frac{1}{m}\left\{mI^F(\beta_0)+m(m-1)I(\beta_0)\right\}\\&=(m-1)\left\{I^F(\beta_0)-I(\beta_0)\right\}.\end{aligned} \tag{14.64}$$

Finally, using (14.61), we obtain

$$\begin{aligned}&E\left\{n\sum_{j=1}^{m}\left(\hat{\beta}_{nj}^*-\hat{\beta}_n^*\right)\left(\hat{\beta}_{nj}^*-\hat{\beta}_n^*\right)^T\right\}\rightarrow\\&\quad(m-1)\left\{I^F(\beta_0)\right\}^{-1}\left\{I^F(\beta_0)-I(\beta_0)\right\}\left\{I^F(\beta_0)\right\}^{-1}.\end{aligned} \tag{14.65}$$

This completes the proof of Theorem 14.6. □

Consequently, Rubin's estimator for the variance, (14.60), is an unbiased (asymptotically) estimator for

$$\{I^F(\beta_0)\}^{-1} + \left(\frac{m+1}{m}\right)\{I^F(\beta_0)\}^{-1}\{I^F(\beta_0) - I(\beta_0)\}\{I^F(\beta_0)\}^{-1}.$$

Comparing this with the asymptotic variance of $n^{1/2}(\hat{\beta}_n^* - \beta_0)$ given by (14.59), we see that Rubin's formula underestimates the asymptotic variance for the frequentist type B multiple-imputation estimator.

Remark 7. We wish to note that the first term in Rubin's variance estimator is indeed a consistent estimator for $\{I^F(\beta_0)\}^{-1}$; that is, it converges in probability as $n \to \infty$. The second term, however, namely

$$\left(\frac{m+1}{m}\right) n \sum_{j=1}^{m} \frac{(\hat{\beta}_{nj}^* - \hat{\beta}_n^*)(\hat{\beta}_{nj}^* - \hat{\beta}_n^*)^T}{m-1},$$

is an asymptotically unbiased estimator for

$$\left(\frac{m+1}{m}\right)\{I^F(\beta_0)\}^{-1}\{I^F(\beta_0) - I(\beta_0)\}\{I^F(\beta_0)\}^{-1}.$$

That is, the expectation converges as $n \to \infty$ for m fixed. However, this second term is not a consistent estimator but rather converges to a proper distribution. Consistency is only obtained by also letting $m \to \infty$. □

Nonetheless, a consistent estimator for the asymptotic variance of $n^{1/2}(\hat{\beta}_n^* - \beta_0)$ can be derived as follows.

Consistent Estimator for the Asymptotic Variance

Because $\hat{\beta}_n^I$ was assumed to be an RAL estimator for β with influence function $q\{\mathcal{C}, G_{\mathcal{C}}(Z)\}$, this implies that

$$n^{1/2}(\hat{\beta}_n^I - \beta_0) \to N\Big(0, \text{var}\left[q\{\mathcal{C}, G_{\mathcal{C}}(Z)\}\right]\Big).$$

Suppose we have a consistent estimator for the asymptotic variance of our initial estimator $\hat{\beta}_n^I$, which we denote as $\hat{\text{var}}[q\{\mathcal{C}, G_{\mathcal{C}}(Z)\}]$. Then, if we can construct consistent estimators for $I^F(\beta_0)$ and $I(\beta_0)$, we can substitute these into (14.59) to obtain a consistent estimator of the asymptotic variance of the multiple-imputation estimator.

As we already discussed in the context of Rubin's variance formula, a consistent estimator for $I^F(\beta_0)$ can be derived by

$$\hat{I}_n^F(\beta_0) = -m^{-1}\sum_{j=1}^{m}\underbrace{\left[n^{-1}\sum_{i=1}^{n}\frac{\partial S^F\{Z_{ij}(\hat{\beta}_n^I), \hat{\beta}_{nj}^*\}}{\partial \beta^T}\right]}_{\substack{\text{This is the observed information matrix, which is often derived for us in the } j\text{-th imputed full-data analysis}}}. \tag{14.66}$$

A consistent estimator for $I(\beta_0)$ can be obtained by

$$\hat{I}_n(\beta_0) = n^{-1}\sum_{i=1}^{n}\{m(m-1)\}^{-1}\sum\sum_{\substack{j,j'=1,\ldots,m\\ j\neq j'}}\Big[S^F\{Z_{ij}(\hat{\beta}_n^I),\hat{\beta}_{nj}^*\} \times S^{F^T}\{Z_{ij'}(\hat{\beta}_n^I),\hat{\beta}_{nj'}^*\}\Big]. \tag{14.67}$$

This follows directly from (14.63). Another estimator for $\{I^F(\beta_0) - I(\beta_0)\}$ is motivated from the relationship (14.6), which states that

$$I^F(\beta_0) - I(\beta_0) = E\bigg(\text{var}\left[S^F\{Z_{ij}(\beta_0),\beta_0\}|\mathcal{C}_i, G_{\mathcal{C}_i}(Z_i)\right]\bigg).$$

This suggests using

$$n^{-1}\sum_{i=1}^{n}(m-1)^{-1}\sum_{j=1}^{m}\left[S^F\{Z_{ij}(\hat{\beta}_n^I),\hat{\beta}_{nj}^*\} - \bar{S}_{\beta_i}^F(\hat{\beta}_n^*)\right]\left[S^F\{Z_{ij}(\hat{\beta}_n^I),\hat{\beta}_{nj}^*\} - \bar{S}_{\beta_i}^F(\beta_n^*)\right]^T \tag{14.68}$$

as an estimator for $I^F(\beta_0) - I(\beta_0)$.

14.6 Proper Imputation

In Bayesian imputation, the j-th imputation is obtained by sampling from $p_{Z|\mathcal{C},G_\mathcal{C}(Z)}\{z|\mathcal{C}_i, G_{\mathcal{C}_i}(Z_i), \beta^{(j)}\}$, where $\beta^{(j)}$ itself is sampled from some distribution. Rubin (1978b) suggests sampling $\beta^{(j)}$ from the posterior distribution $p\{\beta|\mathcal{C}_i, G_{\mathcal{C}_i}(Z_i)\}$. This is equivalent to drawing the imputation from the predictive distribution $p\{z|\mathcal{C}_i, G_{\mathcal{C}_i}(Z_i)\}$.

Remark 8. This logic seems a bit circular since Bayesian inference is based on deriving the posterior distribution of β given the observed data. Under suitable regularity conditions on the choice of the prior distribution, the posterior mean or mode of β is generally an efficient estimator for β. Therefore, using proper imputation, where we draw from the posterior distribution of β, we start with an efficient estimator and, after imputing m full data sets, we end up with an estimator that is not efficient. □

When the sample size is large, the posterior distribution of the parameter and the sampling distribution of the estimator are closely approximated by each other. The initial estimator $\hat{\beta}_n^I$ was assumed to be asymptotically normal; that is,

$$n^{1/2}(\hat{\beta}_n^I - \beta_0) \sim N\bigg(0, \text{var}[q\{\mathcal{C}, G_\mathcal{C}(Z)\}]\bigg),$$

where $q\{\mathcal{C}, G_{\mathcal{C}}(Z)\}$ denotes the influence function of $\hat{\beta}_n^I$. Therefore, mimicking the idea of Bayesian imputation, instead of fixing the values $\hat{\beta}_n^I$ for each of the m imputations, at the j-th imputation, we sample $\beta^{(j)}$ from

$$N\left(\hat{\beta}_n^I, \frac{\hat{\text{var}}[q\{\mathcal{C}, G_{\mathcal{C}}(Z)\}]}{n}\right),$$

where $\hat{\text{var}}[q\{\mathcal{C}, G_{\mathcal{C}}(Z)\}]$ is a consistent estimator for the asymptotic variance, and then randomly choose $Z_{ij}(\beta^{(j)})$ from the conditional distribution with conditional density

$$p_{Z|\mathcal{C},G_{\mathcal{C}}(Z)}\{z|\mathcal{C}_i, G_{\mathcal{C}_i}(Z_i), \beta^{(j)}\}.$$

Remark 9. If $\hat{\beta}_n^I$ were efficient, say the MLE, then this would approximate sampling the β's from the posterior distribution and the Z's from the predictive distribution. □

Using this approach, the j-th imputed estimator is the solution to the equation

$$\sum_{i=1}^{n} S^F\{Z_{ij}(\beta^{(j)}), \hat{\beta}_{nj}^*\} = 0,$$

and the final multiple-imputation estimator is

$$\hat{\beta}_n^* = m^{-1} \sum_{j=1}^{m} \hat{\beta}_{nj}^*.$$

Therefore, if we decide to use such an approach, the obvious questions are

1. What is the asymptotic distribution of $n^{1/2}(\hat{\beta}_n^* - \beta_0)$?
2. How does it compare with improper imputation?
3. How do we estimate the asymptotic variance?

Asymptotic Distribution of $n^{1/2}(\hat{\beta}_n^* - \beta_0)$

Using the same expansion that led us to (14.15), we again obtain that

$$n^{1/2}(\hat{\beta}_n^* - \beta_0) = n^{-1/2} \sum_{i=1}^{n} \left\{I^F(\beta_0)\right\}^{-1} \left[m^{-1} \sum_{j=1}^{m} S^F\{Z_{ij} \underbrace{(\beta^{(j)})}_{\substack{\text{note here the} \\ \text{dependence on } j}}, \beta_0\} \right] + o_p(1). \tag{14.69}$$

Also, the same logic that was used for the multiple-imputation (improper) estimator leads us to the relationship

$$n^{-1/2}\sum_{i=1}^{n}\left[m^{-1}\sum_{j=1}^{m}S^F\{Z_{ij}(\beta^{(j)}),\beta_0\}\right]$$
$$= n^{-1/2}\sum_{i=1}^{n}\left[m^{-1}\sum_{j=1}^{m}S^F\{Z_{ij}(\beta_0),\beta_0\}\right]$$
$$+\left\{I^F(\beta_0)-I(\beta_0)\right\}m^{-1}\sum_{j=1}^{m}n^{1/2}(\beta^{(j)}-\beta_0)+o_p(1). \tag{14.70}$$

Note that (14.70) can be written as

$$n^{-1/2}\sum_{i=1}^{n}\left[m^{-1}\sum_{j=1}^{m}S^F\{Z_{ij}(\beta_0),\beta_0\}\right]$$
$$+\left\{I^F(\beta_0)-I(\beta_0)\right\}m^{-1}\sum_{j=1}^{m}n^{1/2}(\beta^{(j)}-\hat{\beta}_n^I)$$
$$+\left\{I^F(\beta_0)-I(\beta_0)\right\}n^{1/2}(\hat{\beta}_n^I-\beta_0)+o_p(1). \tag{14.71}$$

Because $n^{1/2}(\hat{\beta}_n^I-\beta_0)=n^{-1/2}\sum_{i=1}^{n}q\{\mathcal{C}_i,G_{\mathcal{C}_i}(Z_i)\}+o_p(1)$, we can write (14.71) as

$$n^{-1/2}\sum_{i=1}^{n}\left(\left[m^{-1}\sum_{j=1}^{m}S^F\{Z_{ij}(\beta_0),\beta_0\}\right]+\left\{I^F(\beta_0)-I(\beta_0)\right\}q\{\mathcal{C}_i,G_{\mathcal{C}_i}(Z_i)\}\right) \tag{14.72}$$

$$+\,m^{-1}\sum_{j=1}^{m}\left\{I^F(\beta_0)-I(\beta_0)\right\}n^{1/2}(\beta^{(j)}-\hat{\beta}_n^I)+o_p(1). \tag{14.73}$$

Note that (14.72) is a term that was derived when we considered type B multiple-imputation estimators in the previous section (improper imputation), whereas (14.73) is an additional term due to sampling the $\beta^{(j)}$'s from

$$N\left(\hat{\beta}_n^I,\frac{\hat{\text{var}}[q\{\mathcal{C}_i,G_{\mathcal{C}_i}(Z_i)\}]}{n}\right).$$

Therefore

$$n^{1/2}(\hat{\beta}_n^*-\beta_0)=[\left\{I^F(\beta_0)\right\}^{-1}\{(14.72)+(14.73)\}].$$

The expression (14.72) multiplied by $\{I^F(\beta_0)\}^{-1}$ is, up to $o_p(1)$, identical to representation (14.32) of the type B multiple-imputation estimator and hence is asymptotically normally distributed with mean zero and variance equal to (14.59).

By construction, $n^{1/2}(\beta^{(j)} - \hat{\beta}_n^I), j = 1, \ldots, m$ are m independent draws from a normal distribution with mean zero and variance $\hat{\text{var}}[q\{\mathcal{C}_i, G_{\mathcal{C}_i}(Z_i)\}]$. Because $\hat{\text{var}}\{q(\cdot)\}$ is a consistent estimator for $\text{var}\{q(\cdot)\}$, this implies that $n^{1/2}(\beta^{(j)} - \hat{\beta}_n^I), j = 1, \ldots, m$ is asymptotically equivalent to $V_1, \ldots, V_m$, which are iid normal random variables with mean zero and variance $\text{var}[q\{\mathcal{C}, G_{\mathcal{C}}(Z)\}]$ and independent of all the data $\{\mathcal{C}_i, G_{\mathcal{C}_i}(Z_i)\}, i = 1, \ldots, n$. Therefore,

$$
\begin{aligned}
& n^{1/2}(\hat{\beta}_n^* - \beta_0) = \\
& n^{-1/2} \sum_{i=1}^{n} \{I^F(\beta_0)\}^{-1} \left(\left[m^{-1} \sum_{j=1}^{m} S^F\{Z_{ij}(\beta_0), \beta_0\} \right] \right. \\
& \qquad \left. + \left\{ I^F(\beta_0) - I(\beta_0) \right\} q\{\mathcal{C}_i, G_{\mathcal{C}_i}(Z_i)\} \right) \qquad (14.74) \\
& + m^{-1} \sum_{j=1}^{m} \{I^F(\beta_0)\}^{-1} \left\{ I^F(\beta_0) - I(\beta_0) \right\} V_j \qquad (14.75) \\
& + o_p(1).
\end{aligned}
$$

Because (14.74) converges to a normal distribution with mean zero and variance matrix (14.59), and (14.75) is distributed as normal with mean zero and variance matrix

$$
\begin{aligned}
m^{-1} \left\{ I^F(\beta_0) \right\}^{-1} & \left\{ I^F(\beta_0) - I(\beta_0) \right\} \ \text{var}[q\{\mathcal{C}, G_{\mathcal{C}}(Z)\}] \\
& \times \left\{ I^F(\beta_0) - I(\beta_0) \right\} \left\{ I^F(\beta_0) \right\}^{-1} \qquad (14.76)
\end{aligned}
$$

and is independent of (14.74), this implies that $n^{1/2}(\hat{\beta}_n^* - \beta_0)$ is asymptotically normal with mean zero and asymptotic variance equal to (14.59) + (14.76), which equals

$$
\begin{aligned}
& \left\{ I^F(\beta_0) \right\}^{-1} + \left(\frac{m+1}{m} \right) \left\{ I^F(\beta_0) \right\}^{-1} \left\{ I^F(\beta_0) - I(\beta_0) \right\} \left\{ I^F(\beta_0) \right\}^{-1} \\
& + \left(\frac{m+1}{m} \right) \left\{ I^F(\beta_0) \right\}^{-1} \left\{ I^F(\beta_0) - I(\beta_0) \right\} \ \text{var}[q\{\mathcal{C}, G_{\mathcal{C}}(Z)\}] \\
& \times \left\{ I^F(\beta_0) - I(\beta_0) \right\} \left\{ I^F(\beta_0) \right\}^{-1}. \qquad (14.77)
\end{aligned}
$$

Comparing (14.77) with (14.59), we see that the estimator using "proper" imputation has greater variance (is less efficient) than the corresponding "improper" imputation estimator, which fixes $\hat{\beta}_n^I$ at each imputation. This makes intuitive sense since we are introducing additional variability by sampling the β's from some distribution at each imputation. The increase in the variance is given by (14.76). The variances of the two methods converge as m goes to infinity.

Let us now study the properties of Rubin's formula for the asymptotic variance when applied to type A (proper imputation) multiple-imputation estimators.

Rubin's Estimator for the Asymptotic Variance

Using arguments that led us to (14.74) and (14.75), we obtain that

$$n^{1/2}(\hat{\beta}^*_{nj} - \hat{\beta}^*_n) = n^{-1/2} \sum_{i=1}^{n} \{I^F(\beta_0)\}^{-1} [S^F\{Z_{ij}(\beta_0), \beta_0\} - \bar{S}^F_i(\beta_0)] + \{I^F(\beta_0)\}^{-1} \{I^F(\beta_0) - I(\beta_0)\}(V_j - \bar{V}) + o_p(1),$$

where $\bar{V} = m^{-1} \sum_{j=1}^{m} V_j$. Therefore,

$$E\left\{ n \sum_{j=1}^{m} \left(\hat{\beta}^*_{nj} - \hat{\beta}^*_n\right)\left(\hat{\beta}^*_{nj} - \hat{\beta}^*_n\right)^T \right\} \xrightarrow{n\to\infty}$$
$$\left\{I^F(\beta_0)\right\}^{-1} \times$$
$$\left\{ E\left(\sum_{j=1}^{m} [S^F\{Z_{ij}(\beta_0), \beta_0\} - \bar{S}^F_i(\beta_0)][S^F\{Z_{ij}(\beta_0), \beta_0\} - \bar{S}^F_i(\beta_0)]^T \right) \right.$$
$$\left. + \left\{I^F(\beta_0) - I(\beta_0)\right\} E\left\{ \sum_{j=1}^{m} (V_j - \bar{V})(V_j - \bar{V})^T \right\} \left\{I^F(\beta_0) - I(\beta_0)\right\} \right\}$$
$$\times \left\{I^F(\beta_0)\right\}^{-1}. \tag{14.78}$$

Remark 10. The term involving $q\{\mathcal{C}_i, G_{\mathcal{C}_i}(Z_i)\}$ is common for all j and hence drops out when considering $\{\hat{\beta}^*_{nj} - \hat{\beta}^*_n\}$. Also, the additivity of the expectations in (14.78) is a consequence of the fact that V_j are generated independently from all the data. □

We showed in (14.64) that

$$E\left(\sum_{j=1}^{m} [S^F\{Z_{ij}(\beta_0), \beta_0\} - \bar{S}^F_i(\beta_0)][S^F\{Z_{ij}(\beta_0), \beta_0\} - \bar{S}^F_i]^T \right) = (m-1)\left\{I^F(\beta_0) - I(\beta_0)\right\}. \tag{14.79}$$

Also

$$E\left\{ \sum_{j=1}^{m} (V_j - \bar{V})(V_j - \bar{V})^T \right\} = (m-1)\mathrm{var}[q\{\mathcal{C}, G_{\mathcal{C}}(Z)\}]. \tag{14.80}$$

Therefore, substituting (14.79) and (14.80) into (14.78) yields

$$E\left\{n\sum_{j=1}^{m}\left(\hat{\beta}_{nj}^{*}-\hat{\beta}_{n}^{*}\right)\left(\hat{\beta}_{nj}^{*}-\hat{\beta}_{n}^{*}\right)^{T}\right\}\rightarrow$$
$$(m-1)\left\{I^{F}(\beta_0)\right\}^{-1}\Bigg(\left\{I^{F}(\beta_0)-I(\beta_0)\right\}+\left\{I^{F}(\beta_0)-I(\beta_0)\right\}$$
$$\times\ \mathrm{var}\left[q\{\mathcal{C},G_{\mathcal{C}}(Z)\}\right]\left\{I^{F}(\beta_0)-I(\beta_0)\right\}\Bigg)\left\{I^{F}(\beta_0)\right\}^{-1}. \tag{14.81}$$

Consequently, Rubin's variance estimator, given by (14.60), when used with "proper" multiple imputation, will converge in expectation to (14.77), which indeed is the asymptotic variance of the "proper" multiple-imputation estimator.

Summary

1. Type B or "improper" imputation (i.e., holding the initial estimator fixed across imputations) results in a more efficient estimator than "proper" imputation (where the β's are sampled from $N\left(\hat{\beta}_n^I, \frac{\hat{\mathrm{var}}[q\{\mathcal{C},G_{\mathcal{C}}(Z)\}]}{n}\right)$ for each imputation); however, this difference in efficiency disappears as $m\rightarrow\infty$.
2. Rubin's variance estimator underestimates the asymptotic variance when used with "improper" imputation.
3. Rubin's variance estimator correctly estimates the asymptotic variance when used with "proper" imputation (i.e., the variance estimator converges in expectation to the asymptotic variance). As m, the number of imputations, goes to infinity, Rubin's estimator is also consistent as well as asymptotically unbiased.

14.7 Surrogate Marker Problem Revisited

We now return to Example 1, which was introduced at the beginning of the chapter. In this problem, we were interested in estimating the regression parameters θ in a logistic regression model of a binary variable Y as a function of covariates X given by (14.2). However, because X was expensive to measure, a cheaper surrogate variable W for X was also collected and the design was to collect X only on a validation subsample chosen at random with prespecified probability that depended on Y and W. A surrogate variable W for X is assumed to satisfy the property that

$$p(Y=1|W,X)=P(Y=1|X). \tag{14.82}$$

In addition, we also assume that

$$\begin{pmatrix} X \\ W \end{pmatrix} \sim N\left[\begin{pmatrix} \mu_X \\ \mu_W \end{pmatrix}, \begin{bmatrix} \sigma_{XX} & \sigma_{XW} \\ \sigma_{XW}^T & \sigma_{WW} \end{bmatrix} \right].$$

Letting R denote the indicator of a complete case (i.e., if $R = 1$, then we observe (Y, W, X), whereas when $R = 0$, we only observe (Y, W)), the observed data can be represented as

$$(R_i, Y_i, W_i, R_i X_i), i = 1, \ldots, n.$$

We are interested in obtaining an estimator for the parameter

$$\beta = (\theta^T, \mu_X^T, \mu_W^T, \sigma_{XX}, \sigma_{XW}, \sigma_{WW})^T$$

using the observed data. We will now illustrate how we would use multiple imputation. First, we need to derive an initial estimator $\hat{\beta}_n^I$. One possibility is to use an inverse weighted complete-case estimator. This is particularly attractive because we know the probability of being included in the validation set by design; that is, $P[R = 1|Y, W] = \pi(Y, W)$.

If we had full data, then we could estimate β using standard likelihood estimating equations; namely,

$$\begin{aligned} \sum_{i=1}^{n} X_i \left\{ Y_i - \text{expit}(\theta^T X_i) \right\} &= 0, \\ \sum_{i=1}^{n} (X_i - \mu_X) &= 0, \\ \sum_{i=1}^{n} (W_i - \mu_W) &= 0, \\ \sum_{i=1}^{n} \left\{ (X_i - \mu_X)(X_i - \mu_X)^T - \sigma_{XX} \right\} &= 0, \\ \sum_{i=1}^{n} \left\{ (W_i - \mu_W)(W_i - \mu_W)^T - \sigma_{WW} \right\} &= 0, \\ \sum_{i=1}^{n} \left\{ (X_i - \mu_X)(W_i - \mu_W)^T - \sigma_{XW} \right\} &= 0, \end{aligned} \tag{14.83}$$

where

$$\text{expit}(u) = \frac{\exp(u)}{1 + \exp(u)}.$$

Using the observed data, an inverse weighted complete-case estimator for β can be obtained by solving the equations

$$\sum_{i=1}^{n} \frac{R_i}{\pi(Y_i, W_i)} X_i \left\{ Y_i - \text{expit}(\theta^T X_i) \right\} = 0$$
$$\sum_{i=1}^{n} \frac{R_i}{\pi(Y_i, W_i)} (X_i - \mu_X) = 0$$
$$\vdots \qquad \text{etc.} \tag{14.84}$$

Remark 11. If we are interested only in the parameter θ of the logistic regression model, then we only need to solve (14.84) and not rely on any assumption of normality of X, W. However, to use multiple imputation, we need to derive initial estimators for all the parameters. □

Solving the estimating equations above gives us an initial estimator $\hat{\beta}_n^I$. In addition, using the methods described for IPWCC estimators, say in Chapter 7, we can also derive the influence function $q(R, Y, W, RX)$ for $\hat{\beta}_n^I$ and a consistent estimate of its asymptotic variance $\hat{\text{var}}\{q(R, Y, W, RX)\}$.

In order to carry out the imputation, we must be able to sample from the conditional distribution of the full data given the observed data. For our problem, we must sample from the conditional distribution

$$p_{X|R,Y,W,RX}(x|R_i, Y_i, W_i, R_i X_i, \hat{\beta}_n^I).$$

Clearly, when $R_i = 1$ (the complete case), we just use the observed value X_i. However, when $R_i = 0$, then we must sample from

$$p_{X|R,Y,W,RX}(x|R_i = 0, Y_i, W_i, \hat{\beta}_n^I).$$

Because of the MAR assumption, this is the same as

$$p_{X|Y,W}(x|Y_i, W_i, \hat{\beta}_n^I).$$

How Do We Sample?

We now describe the use of rejection sampling to obtain random draws from the conditional distribution of $p_{X|Y,W}(x|Y_i, W_i, \hat{\beta}_n^I)$. Using Bayes's rule and the surrogacy assumption (14.82), the conditional density is derived as

$$p_{X|Y,W}(x|y, w) = \frac{p_{Y|X}(y|x) p_{X|W}(x|w)}{\int p_{Y|X}(y|x) p_{X|W}(x|w) dx}.$$

Therefore, $p_{X|Y,W}(x|y, w)$ equals

$$\frac{\left[\exp(\hat{\theta}_n^{I^T} xy) / \{1 + \exp(\hat{\theta}_n^{IT} x)\} \right] p_{X|W}(x, w)}{\underset{\text{constant}}{\text{normalizing}} = \int [\text{numerator}]\, dx}, \; y = 0, 1.$$

Because $(X^T, W^T)^T$ are multivariate normal, this implies that the conditional distribution of $X|W$ is also normally distributed with mean

$$E(X|W) = \hat{\mu}^I_{X_n} + \hat{\sigma}^I_{XW_n}[\hat{\sigma}^I_{WW_n}]^{-1}(W - \hat{\mu}^I_{W_n}) \tag{14.85}$$

and variance

$$\text{var }(X|W) = \hat{\sigma}^I_{XX_n} - \hat{\sigma}^I_{XW_n}[\hat{\sigma}^I_{WW_n}]^{-1}\hat{\sigma}^{I^T}_{XW_n}. \tag{14.86}$$

Therefore, at the j-th imputation, if $R_i = 0$, we can generate $X_{ij}(\hat{\beta}^I_n)$ by first randomly sampling from a normal distribution with mean

$$\hat{\mu}^I_{X_n} + \hat{\sigma}^I_{XW_n}[\hat{\sigma}^I_{WW_n}]^{-1}(W_i - \hat{\mu}^I_{W_n})$$

and variance (14.86).

After generating such an X in this fashion, we either "keep" this X if another randomly generated uniform random variable is less than

$$\frac{\exp(\hat{\theta}^{IT}_n XY_i)}{1 + \exp(\hat{\theta}^{IT}_n X)}$$

or we keep repeating this process until we "keep" an X that we use for the j-th imputation $X_{ij}(\hat{\beta}^I_n)$. This rejection sampling scheme guarantees a random draw from

$$p_{X|Y,W}(x|Y_i, W_i).$$

Therefore, at the j-th imputation, together with Y_i and W_i, which we always observe, we use X_i if $R_i = 1$ and $X_{ij}(\beta^I_n)$ if $R_i = 0$ to create the j-th pseudo-full data. This j-th imputed data set is then used to obtain estimators $\hat{\beta}^*_{nj}$ as described by (14.83). Standard software packages will do this.

The final estimate is

$$\hat{\beta}^*_n = m^{-1}\sum_{j=1}^{m}\hat{\beta}^*_{nj}.$$

A consistent estimator of the asymptotic variance can be obtained by substituting consistent estimators for $I^F(\beta_0)$ and I, say by using (14.66) and (14.67), respectively, and a consistent estimator $\hat{\text{var}}\{q(R, Y, W, RX)\}$ for $\text{var}\{q(R, Y, W, RX)\}$ in equation (14.58).

References

Allison, P.D. (2002). *Missing Data*. Sage, Thousand Oaks, CA.

Andersen, P.K., Borgan, O., Gill, R.D., and Kieding, N. (1992). *Statistical Models Based on Counting Processes*. Springer-Verlag, Berlin.

Anderson, P.K. and Gill, R. (1982). Cox's regression model for counting processes: a large sample study. *Annals of Statistics* **10**, 1100–1120.

Bang, H. and Robins, J.M. (2005). Doubly robust estimation in missing data and causal inference models. *Biometrics* **61**, 962–973.

Bang, H. and Tsiatis, A.A. (2000). Estimating medical costs with censored data. *Biometrika* **87**, 329–343.

Bang, H. and Tsiatis, A.A. (2002). Median regression with censored cost data. *Biometrics* **58**, 643–649.

Begun, J.M., Hall, W.J., Huang, W., and Wellner, J.A. (1983). Information and asymptotic efficiency in parametric-nonparametric models. *Annals of Statistics* **11**, 432–452.

Bickel, P.J., Klaassen, C.A.J., Ritov, Y., and Wellner, J.A. (1993). *Efficient and Adaptive Estimation for Semiparametric Models*. The Johns Hopkins University Press, Baltimore.

Breslow, N. (1974). Covariance analysis of censored survival data. *Biometrics* **30**, 89–99.

Casella, G. and Berger, R.L. (2002). *Statistical Inference, Second Edition*. Duxbury Press, Belmont, CA.

Chamberlain, G. (1987). Asymptotic efficiency in estimation with conditional moment restrictions. *Journal of Econometrics* **34**, 305–334.

Cox, D.R. (1972). Regression models and life-tables (with discussion). *Journal of the Royal Statistical Society, Series B* **34**, 187–220.

Cox, D.R. (1975). Partial likelihood. *Biometrika* **62**, 269–276.

Cox, D.R. and Snell, E.J. (1989). *Analysis of Binary Data*. Chapman and Hall, London.

Dempster, A.P., Laird, N.M., and Rubin, D.B. (1977). Maximum likelihood from incomplete data via the EM algorithm (with discussion). *Journal of the Royal Statistical Society, Series B* **39**, 1–37.

Fleming, T.R. and Harrington, D.P. (1991). *Counting Processes and Survival Analysis*. Wiley, New York.

Gill, R.D., van der Laan, M.J., and Robins, J.M. (1997). Coarsening at random: Characterizations, conjectures and counterexamples. *Proceedings of the First Seattle Symposium in Biostatistics: Survival Analysis*, Springer, New York, pp. 255–294.

Hájek, J. (1970). A characterization of limiting distributions of regular estimates. *Zeitschrift Wahrscheinlichkeitstheorie und Verwandte Gebiete* **14**, 323–330.

Hájek, J. and Sidak, Z. (1967). *Theory of Rank Tests*. Academic Press, New York.

Hampel, F.R. (1974). The influence curve and its role in robust estimation. *Journal of the American Statistical Association* **69**, 383–393.

Heitjan, D.F. (1993). Ignorability and coarse data: Some biomedical examples. *Biometrics* **49**, 1099–1109 .

Heitjan, D.F. and Rubin, D.B. (1991). Ignorability and coarse data. *Annals of Statistics* **19**, 2244–2253.

Holland, P.W. (1986). Statistics and causal inference (with discussion). *Journal of the American Statistical Association* **81**, 945–970.

Horvitz, D.G. and Thompson, D.J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association* **47**, 663–685.

Hu, P. and Tsiatis, A.A. (1996). Estimating the survival distribution when ascertainment of vital status is subject to delay. *Biometrika* **83**, 371–380.

Kaplan, E.L. and Meier, P. (1958). Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association* **53**, 457–481.

Kress, R. (1989). *Linear Integral Equations*. Springer-Verlag, Berlin.

LeCam, L. (1953). On some asymptotic properties of maximum likelihood estimates and related Bayes estimates. *University of California Publications in Statistics* **1**, 227–330.

Leon, S., Tsiatis, A.A., and Davidian, M. (2003). Semiparametric estimation of treatment effect in a pretest-posttest study. *Biometrics* **59**, 1046–1055.

Liang, K-Y. and Zeger, S.L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika* **73**, 13–22.

Lipsitz, S.R., Ibrahim, J.G., and Zhao, L.P. (1999). A weighted estimating equation for missing covariate data with properties similar to maximum likelihood. *Journal of the American Statistical Association* **94**, 1147–1160.

Littell, R.C., Milliken, G.A., Stroup, W.W., and Wolfinger, R.D. (1996). *SAS System for Mixed Models*. SAS Institute, Inc., Cary, NC.

Little, R.J.A. and Rubin, D.B. (1987). *Statistical Analysis with Missing Data*. Wiley, New York.

Loève, M. (1963). *Probability Theory (third edition)*. Springer-Verlag, Berlin.

Luenberger, D.G. (1969). *Optimization by Vector Space Methods*. Wiley, New York.

Lunceford, J.K. and Davidian, M. (2004). Stratification and weighting via the propensity score in estimation of causal treatment effects: A comparative study. *Statistics in Medicine* **23**, 2937–2960.

Manski, C.F. (1984). Adaptive estimation of non-linear regression models (with discussion). *Econometric Reviews* **3**, 145–210.

McCullagh, P. and Nelder, J.A. (1989). *Generalized Linear Models (2nd edition)*. Chapman and Hall, London.

Neugebauer, R. and van der Laan, M.J. (2005). Why prefer double robust estimators in causal inference? *Journal of Statistical Planning and Inference* **129**, 405–426.

Newey, W.K. (1988). Adaptive estimation of regression models via moment restrictions. *Journal of Econometrics* **38**, 301–339.

Newey, W.K. (1990). Semiparametric efficiency bounds. *Journal of Applied Econometrics* **5**, 99–135.

Newey, W.K. and McFadden, D. (1994). Large sample estimation and hypothesis testing. *Handbook of Econometrics* **4**, 2111–2245.

Neyman, J. (1923). Sur les applications de la thar des probabilities aux experiences Agaricales: Essay de principle. English translation of excerpts by Dabrowska, D. and Speed, T. (1990). *Statistical Science* **5**, 465–480.

Oakes, D. (1999). Direct calculation of the information matrix via the EM algorithm. *Journal of the Royal Statistical Society, Series B* **61**, 479–482.

Quale, C.M., van der Laan, M.J., and Robins, J.M. (2003). Locally efficient estimation with bivariate right censored data. University of California, Berkeley, Department of Statistics Technical Report.

Rao C.R. (1973). *Linear Statistical Inference and Its Applications*. Wiley, New York.

Robins, J.M. (1986). A new approach to causal inference in mortality studies with sustained exposure periods – Application to control of the healthy worker survivor effect. *Mathematical Modelling* **7**, 1393–1512.

Robins J.M. (1999). Robust estimation in sequentially ignorable missing data and causal inference models. *Proceedings of the American Statistical Association Section on Bayesian Statistical Science*, American Statistical Association, Alexandria, VA, pp. 6–10.

Robins, J.M. and Gill, R.D. (1997). Non-response models for the analysis of non-monotone ignorable missing data. *Statistics in Medicine* **16**, 39–56.

Robins, J.M. and Rotnitzky, A. (1992). Recovery of information and adjustment for dependent censoring using surrogate markers. In *AIDS Epidemiology, Methodological Issues*, Jewell, N., Dietz, K., and Farewell, W., eds. Birkhäuser, Boston, pp. 297–331.

Robins, J.M. and Rotnitzky, A. (2001). Comment on "Inference for semiparametric models: Some questions and an answer." *Statistica Sinica* **11**, 920–936.

Robins, J.M., Rotnitzky, A., and Bonetti, M. (2001). Comment on "Addressing an idiosyncrasy in estimating survival curves using double sampling in the presence of self-selected right censoring." *Biometrics* **57**, 343–347.

Robins J.M., Rotnitzky A., and Scharfstein D.O. (2000). Sensitivity analysis for selection bias and unmeasured confounding in missing data and causal inference models. In *Statistical Models in Epidemiology: The Environment and Clinical Trials*. Halloran, M.E. and Berry, D., eds. IMA Volume **116**, Springer-Verlag, New York, pp. 1–95.

Robins, J.M., Rotnitzky, A., and Zhao, L.P. (1994). Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association* **89**, 846–866.

Robins, J.M., Rotnitzky, A., and van der Laan, M. (2000). Comment on "On profile likelihood." *Journal of the American Statistical Association* **95**, 477–482.

Robins, J.M. and Wang, N. (2000). Inference for imputation estimators. *Biometrika* **87**, 113–124.

Rosenbaum, P.R. (1984). Conditional permutation tests and the propensity score in observational studies. *Journal of the American Statistical Association* **79**, 565–574.

Rosenbaum, P.R. (1987). Model-based direct adjustment. *Journal of the American Statistical Association* **82**, 387–394.

Rosenbaum, P.R. and Rubin, D.B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika* **70**, 41–55.

Rosenbaum, P.R. and Rubin, D.B. (1984). Reducing bias in observational studies using subclassification on the propensity score. *Journal of the American Statistical Association* **79**, 516–524.

Rosenbaum, P.R. and Rubin, D.B. (1985). Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. *American Statistician* **39**, 3–38.

Rotnitzky, A., Scharfstein, D.O., Su, T.L., and Robins, J.M. (2001). Methods for conducting sensitivity analysis of trials with potentially nonignorable competing causes of censoring. *Biometrics* **57**, 103–113.

Rubin, D.B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology* **66**, 688–701.

Rubin, D.B. (1978a). Bayesian inference for causal effects: The role of randomization. *Annals of Statistics* **6**, 34–58.

Rubin, D.B. (1978b). Multiple imputations in sample surveys: A phenomenological Bayesian approach to nonresponse (with discussion). *American Statistical Association Proceedings of the Section on Survey Research Methods*, American Statistical Association, Alexandria, VA, pp. 20–34.

Rubin, D.B. (1987). *Multiple Imputation for Nonresponse in Surveys*. Wiley, New York.

Rubin, D.B. (1990). Comment: Neyman (1923) and causal inference in experiments and observational studies. *Statistical Science* **5**, 472–480.

Rubin, D.B. (1996). Multiple imputation after 18+ years (with discussion). *Journal of the American Statistical Association* **91**, 473–520.

Rubin, D.B. (1997). Estimating causal effects from large data sets using propensity scores. *Annals of Internal Medicine* **127**, 757–763.

Schafer, J.L. (1997). *Analysis of Incomplete Multivariate Data*. Chapman and Hall, London.

Scharfstein, D.O., Rotnitzky, A., and Robins, J.M. (1999). Adjusting for nonignorable drop-out using semiparametric nonresponse models (with discussion) *Journal of the American Statistical Association* **94**, 1096–1146.

Stefanski, L.A. and Boos, D.D. (2002). The calculus of M-estimation. *American Statistician* **56**, 29–38.

Strawderman, R.L. (2000). Estimating the mean of an increasing stochastic process at a censored stopping time. *Journal of the American Statistical Association* **95**, 1192–1208.

Tsiatis, A.A. (1998). Competing risks. In *Encyclopedia of Biostatistics*. Wiley, New York, pp. 824–834.

van der Laan, M.J. and Hubbard, A.E. (1998). Locally efficient estimation of the survival distribution with right-censored data and covariates when collection of data is delayed. *Biometrika* **85**, 771–783.

van der Laan, M.J. and Hubbard, A.E. (1999). Locally efficient estimation of the quality-adjusted lifetime distribution with right-censored data and covariates. *Biometrics* **55**, 530–536.

van der Laan, M.J., Hubbard, A.E., and Robins, J.M. (2002). Locally efficient estimation of a multivariate survival function in longitudinal studies. *Journal of the American Statistical Association* **97**, 494–507.

van der Laan, M.J. and Robins, J.M. (2003). *Unified Methods for Censored Longitudinal Data and Causality*. Springer-Verlag, New York.

van der Vaart, A.W. and Wellner, J.A. (1996). *Weak Convergence and Empirical Processes with Applications to Statistics*. Springer-Verlag, New York.

Verbeke, G. and Molenberghs, G. (2000). *Linear Mixed Models for Longitudinal Data*. Springer-Verlag, New York.

Wang, N. and Robins, J.M. (1998). Large-sample theory for parametric multiple imputation procedures. *Biometrika* **85**, 935–948.

Wu, C.F.J. (1983). On the convergence properties of the EM algorithm. *Annals of Statistics* **11**, 95–103.

Yang, L. and Tsiatis, A.A. (2001). Efficiency study of estimators for a treatment effect in a pretest-posttest trial. *The American Statistician* **55**, 314–321.

Zhao, H. and Tsiatis, A.A. (1997). A consistent estimator for the distribution of quality adjusted survival time. *Biometrika* **84**, 339–348.

Zhao, H. and Tsiatis, A.A. (1999). Efficient estimation of the distribution of quality-adjusted survival time. *Biometrics* **55**, 1101–1107.

Zhao, H. and Tsiatis, A.A. (2000). Estimating mean quality adjusted lifetime with censored data. *Sankhya, Series B* **62**, 175–188.

Index

Springer Series in Statistics

(continued from p. ii)

Huet/Bouvier/Poursat/Jolivet: Statistical Tools for Nonlinear Regression: A Practical Guide with S-PLUS and R Examples, 2nd edition.
Ibrahim/Chen/Sinha: Bayesian Survival Analysis.
Jolliffe: Principal Component Analysis, 2nd edition.
Knottnerus: Sample Survey Theory: Some Pythagorean Perspectives.
Kolen/Brennan: Test Equating: Methods and Practices.
Kotz/Johnson (Eds.): Breakthroughs in Statistics Volume I.
Kotz/Johnson (Eds.): Breakthroughs in Statistics Volume II.
Kotz/Johnson (Eds.): Breakthroughs in Statistics Volume III.
Küchler/Sørensen: Exponential Families of Stochastic Processes.
Kutoyants: Statistical Influence for Ergodic Diffusion Processes.
Lahiri: Resampling Methods for Dependent Data.
Le Cam: Asymptotic Methods in Statistical Decision Theory.
Le Cam/Yang: Asymptotics in Statistics: Some Basic Concepts, 2nd edition.
Liu: Monte Carlo Strategies in Scientific Computing.
Longford: Models for Uncertainty in Educational Testing.
Manski: Partial Identification of Probability Distributions.
Mielke/Berry: Permutation Methods: A Distance Function Approach.
Molenberghs/Verbeke: Models for Discrete Longitudinal Data.
Mukerjee/Wu: A Modern Theory of Factorial Designs.
Nelsen: An Introduction to Copulas. 2nd edition
Pan/Fang: Growth Curve Models and Statistical Diagnostics.
Parzen/Tanabe/Kitagawa: Selected Papers of Hirotugu Akaike.
Politis/Romano/Wolf: Subsampling.
Ramsay/Silverman: Applied Functional Data Analysis: Methods and Case Studies.
Ramsay/Silverman: Functional Data Analysis, 2nd edition.
Rao/Toutenburg: Linear Models: Least Squares and Alternatives.
Reinsel: Elements of Multivariate Time Series Analysis. 2nd edition.
Rosenbaum: Observational Studies, 2nd edition.
Rosenblatt: Gaussian and Non-Gaussian Linear Time Series and Random Fields.
Särndal/Swensson/Wretman: Model Assisted Survey Sampling.
Santner/Williams/Notz: The Design and Analysis of Computer Experiments.
Schervish: Theory of Statistics.
Seneta: Non-negative Matrices and Markov Chains, Revised Printing.
Shao/Tu: The Jackknife and Bootstrap.
Simonoff: Smoothing Methods in Statistics.
Singpurwalla and Wilson: Statistical Methods in Software Engineering: Reliability and Risk.
Small: The Statistical Theory of Shape.
Sprott: Statistical Inference in Science.
Stein: Interpolation of Spatial Data: Some Theory for Kriging.
Taniguchi/Kakizawa: Asymptotic Theory of Statistical Inference for Time Series.
Tanner: Tools for Statistical Inference: Methods for the Exploration of Posterior Distributions and Likelihood Functions, 3rd edition.
Tillé: Sampling Algorithms.
Tsiatis: Semiparametric Theory and Missing Data.
van der Laan: Unified Methods for Censored Longitudinal Data and Causality.
van der Vaart/Wellner: Weak Convergence and Empirical Processes: With Applications to Statistics.
Verbeke/Molenberghs: Linear Mixed Models for Longitudinal Data.
Weerahandi: Exact Statistical Methods for Data Analysis.
West/Harrison: Bayesian Forecasting and Dynamic Models, 2nd edition.