SEMIPARAMETRIC MAXIMUM LIKELIHOOD ESTIMATION IN PARAMETRIC REGRESSION WITH MISSING COVARIATES

by

Zhiwei Zhang

BS, Applied Chemistry, Beijing Medical University, 1994

MS, Medicinal Chemistry, Peking Union Medical College, 1997

Submitted to the Graduate Faculty of

the Graduate School of Public Health in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

University of Pittsburgh

2003

UNIVERSITY OF PITTSBURGH

GRADUATE SCHOOL OF PUBLIC HEALTH

This dissertation was presented

by

Zhiwei Zhang

It was defended on

November 19, 2003

and approved by

Howard E. Rockette, PhD, Professor, Department of Biostatistics, Graduate School of Public Health, University of Pittsburgh

Stewart J. Anderson, PhD, Associate Professor, Department of Biostatistics, Graduate School of Public Health, University of Pittsburgh

Joyce H. Chang, PhD, Research Assistant Professor, Department of Medicine, School of Medicine, University of Pittsburgh

Sati Mazumdar, PhD, Professor, Department of Biostatistics, Graduate School of Public Health, University of Pittsburgh

Gong Tang, PhD, Assistant Professor, Department of Biostatistics, Graduate School of Public Health, University of Pittsburgh

Dissertation Director: Howard E. Rockette, PhD, Professor, Department of Biostatistics, Graduate School of Public Health, University of Pittsburgh

SEMIPARAMETRIC MAXIMUM LIKELIHOOD ESTIMATION IN PARAMETRIC REGRESSION WITH MISSING COVARIATES

Zhiwei Zhang, PhD

University of Pittsburgh, 2003

Parametric regression models are widely used in public health sciences. This dissertation is concerned with statistical inference under such models with some covariates missing at random. Under natural conditions, parameters remain identifiable from the observed (reduced) data. If the always observed covariates are discrete or can be discretized, we propose a semiparametric maximum likelihood method which requires no parametric specification of the selection mechanism or the covariate distribution. Simple conditions are given under which the semiparametric maximum likelihood estimator (MLE) exists. For ease of computation, we also consider a restricted MLE which maximizes the likelihood over covariate distributions supported by the observed values. The two MLEs are asymptotically equivalent and strongly consistent for a class of topologies on the parameter set. Upon normalization, they converge weakly to a zero-mean Gaussian process in a suitable space. The MLE of the regression parameter, in particular, achieves the semiparametric information bound, which can be consistently estimated by perturbing the profile log-likelihood. Furthermore, the profile likelihood ratio statistic is asymptotically chi-squared. An EM algorithm is proposed for computing the restricted MLE and for variance estimation. Simulation results suggest that the proposed method performs resonably well in moderate-sized samples. In contrast, the analogous parametric maximum likelihood method is subject to severe bias under model misspecification, even in large samples. The proposed method can be applied to related statistical problems.

Keywords: Aymptotic normality; Consistency; EM algorithm; Infinite-dimensional M-estimation; Missing at random; Missing covariates; Parametric regression; Profile likelihood; Semiparametric likelihood.

TABLE OF CONTENTS

LIST OF TABLES

1.0 INTRODUCTION

Parametric regression models such as generalized linear models are routinely used in clinical trials, epidemiology and other fields. Let X and Y be random vectors taking values in Borel sets $\mathcal{X}$ and $\mathcal{Y}$, respectively. Y is often referred to as the *outcome*, while X is a vector of *covariates*. Under a parametric regression model, the effect of X on Y is usually specified through $f(\cdot|x;\theta)$, the regular conditional density of Y given $X = x$ with respect to some fixed measure μ on $\mathcal{Y}$. Here f is a known function and θ is an unknown Euclidean regression parameter to be estimated. Let (X_i, Y_i), $i = 1, \ldots, n$, be independent copies of (X, Y). If the (X_i, Y_i) are completely observed, the likelihood for θ is given by

$$\prod_{i=1}^{n} f(Y_i|X_i;\theta), \tag{1.1}$$

which does not involve the marginal distribution of X. An estimator of θ is obtained by maximizing this likelihood. Under appropriate conditions this maximization reduces to solving a system of likelihood equations:

$$\sum_{i=1}^{n} \dot{\ell}_\theta(X_i, Y_i) = 0,$$

where $\dot{\ell}_\theta(x, y) := \partial \log f(y|x;\theta)/\partial\theta$. It is well known that, under regularity conditions, the maximum likelihood estimator (MLE) of θ is consistent and asymptotically normal with asymptotic variance the inverse of the Fisher information. See Cox and Hinkley (1974, chap. 9) or van der Vaart (1998, chap. 5) for a discussion of the general theory of M-estimation.

Now suppose that a portion of X is unobserved on some subjects. Write $X = (W, Z)$, where W is always observed and Z is possibly missing. Denote the supports of W and Z by $\mathcal{W}$ and $\mathcal{Z}$, respectively. Write $G(\cdot|w)$ for the conditional distribution of Z given $W = w$. Let $R = 1$ if Z is observed; 0 otherwise. As before, assume that (X_i, Y_i, R_i), $i = 1, \ldots, n$, are independent copies of (X, Y, R). However, we only observe $(R_i, R_i Z_i, W_i, Y_i)$, $i = 1, \ldots, n$. Numerous methods have been proposed for estimating θ from this data. Here we give a brief review of the few approaches which we consider to be representative. No attempt has been made to exhaust the literature. It is understood that assumptions made and notations defined in the discussion of a specific method are effective only within that context, unless otherwise stated.

Complete case analysis. Call subject or unit i a *complete case* if it is fully observed (i.e., $R_i = 1$); otherwise an *incomplete case*. An (overly) common approach is simply to apply standard procedures to the

set of complete cases, ignoring the incomplete ones. For example, an estimate of θ is obtained by maximizing

$$\prod_{i:R_i=1} f(Y_i|X_i;\theta).$$

This may or may not be valid, depending on the *selection mechanism*, the mechanism by which subjects are randomly selected for full or partial observation. Here and in the sequel, write $[\cdot]$ or $[\cdot|\cdot]$ for the (conditional) distribution of a random element (given another random element). Then the validity of the complete case analysis requires that

$$[Y|X, R = 1] = [Y|X]. \tag{1.2}$$

The left side is the outcome-covariate relationship among complete cases, whereas the right side is the inferential target. Note that (1.2) is equivalent to the conditional independence of Y and R given X; the latter can be alternatively expressed as

$$E(R|X, Y) = E(R|X) \qquad \text{almost surely.} \tag{1.3}$$

This condition is sometimes called covariate-dependent missingness in the literature. However, we shall refer to (1.3) as *outcome-independent missingness*, which seems to deliver the right message. The complete case analysis is easy to carry out and, under (1.3) and other regularity conditions, yields valid inference. An obvious drawback is loss of effiency because the information in the incomplete caes is discarded.

Pseudolikelihoods. Consider the situation where

$$E(R|X, Y) = E(R|W) \qquad \text{almost surely.} \tag{1.4}$$

This is apparently stronger than (1.3) and allows some information in the incomplete cases to be recovered. Suppose for the moment that G is known. Then the likelihood for θ is given by

$$\prod_{i=1}^{n} \{f(Y_i|X_i;\theta)\}^{R_i} \left\{ \int_{\mathcal{Z}} f(Y_i|W_i, z;\theta) G(dz|W_i) \right\}^{1-R_i}. \tag{1.5}$$

Note that, although the likelihood (1.1) with complete data does not involve the covariate distribution, the one with missing covariates does. Note also that (1.5) does not involve the selection mechanism, by assumption (1.4). Of course, G is unknown and we cannot estimate θ by maximizing (1.5). However, from assumption (1.4) we have

$$[Z|W, R = 1] = [Z|W],$$

which suggests that G can be estimated empirically from complete cases. If W is finitely discrete, then a natural estimator of $G(\cdot|w)$ is the (stratum-specific) empirical distribution of Z:

$$\frac{\sum_{i=1}^{n} 1\{W_i = w\} R_i \delta_{Z_i}}{\sum_{i=1}^{n} 1\{W_i = w\} R_i}, \tag{1.6}$$

where $1\{\cdot\}$ is the indicator function and δ_z denotes the probability distribution supported by the one-point set $\{z\}$. (These notations will be used later without further comment.) With G replaced by (1.6), (1.5) becomes a pseudolikelihood that can be maximized to give an estimate of θ. This is the approach taken by Pepe and Fleming (1991). For a general W, Carroll and Wand (1991) suggest maximizing (1.5) with $G(\cdot|w)$ replaced by a kernel estimator:

$$\frac{\sum_{i=1}^n R_i \psi\{(w - W_i)/b\}\delta_{Z_i}}{\sum_{i=1}^n R_i \psi\{(w - W_i)/b\}},$$

where ψ is a symmetric density function and $b > 0$ a bandwidth. These methods are simple to implement and can be more efficient than the complete case analysis, especially when the proportion of incomplete cases is high. It has long been realized, however, that pseudolikelihood methods are generally inefficient in the semiparametric sense; see, for example, the monograph of Bickel, Klaassen, Ritov and Wellner (1993) for an exposition of the semiparametric efficiency theory. In the present setting, the inefficiency of the pseudolikelihood methods is noted by Robins, Hsieh and Newey (1995) among others. In addition, assumption (1.4) is often unrealistic in practice.

In what follows we shall assume that

$$E(R|X, Y) = E(R|W, Y) =: \pi(W, Y) \qquad \text{almost surely,} \tag{1.7}$$

that is, Z is *missing at random* (MAR) in the sense of Rubin (1976). Compared with (1.4), this is apparently weaker and hence more acceptable.

Mean score. First assume that G is known. Then (1.5) continues to be the likelihood for θ even though (1.4) is weakened to (1.7). Differentiating the logarithm of (1.5) with respect to θ gives the score

$$\sum_{i=1}^n \left[R_i \dot{\ell}_\theta(X_i, Y_i) + (1 - R_i) E\left\{ \dot{\ell}_\theta(X, Y) | W_i, Y_i; G, \theta \right\} \right]. \tag{1.8}$$

Evaluation of (1.8) requires knowledge of $[Z|W, Y]$, which depends on θ and G. Because G is unknown in reality, one cannot just set (1.8) to 0 and solve for θ. But (1.7) implies that

$$[Z|W, Y] = [Z|W, Y, R = 1],$$

and the right side can estimated empirically. In the case that both W and Y are finitely discrete, Reilly and Pepe (1995) propose estimating $[Z|W = w, Y = y]$ by:

$$\frac{\sum_{i=1}^n 1\{W_i = w, Y_i = y\} R_i \delta_{Z_i}}{\sum_{i=1}^n 1\{W_i = w, Y_i = y\} R_i}.$$

Then (1.8) becomes a "mean score":

$$\sum_{i=1}^n \left[R_i \dot{\ell}_\theta(X_i, Y_i) + (1 - R_i) \frac{\sum_{j=1}^n 1\{W_j = W_i, Y_j = Y_i\} R_j \dot{\ell}_\theta(W_i, Z_j, Y_i)}{\sum_{j=1}^n 1\{W_j = W_i, Y_j = Y_i\} R_j} \right]. \tag{1.9}$$

An estimator of θ is obtained by setting (1.9) to 0. Like the pseudolikelihood methods, the mean score method suffers from a loss of efficiency due to replacing population quantities with empirical estimates.

Pseudoscore. Chatterjee, Chen and Breslow (2003) (CCB) develop yet another method of this type. Rewrite (1.8) as

$$\sum_{i=1}^{n}\left[R_i\dot{\ell}_\theta(X_i,Y_i)+(1-R_i)\frac{\int_{\mathcal{Z}}\dot{\ell}_\theta(W_i,z,Y_i)f(Y_i|W_i,z;\theta)G(dz|W_i)}{\int_{\mathcal{Z}}f(Y_i|W_i,z;\theta)G(dz|W_i)}\right]. \tag{1.10}$$

Their key observation is that, by Bayes's law,

$$\frac{dG(\cdot|w)}{d[Z|W=w,R=1]}(z)=\frac{E(R|W=w)}{E(R|W=w,Z=z)}, \tag{1.11}$$

provided the denominator on the right side is nonzero. If W is finitely discrete, then (1.6) is an empirical estimate of $[Z|W=w,R=1]$ (not $G(\cdot|w)$). The denominator on the right side of (1.11) is equal to

$$\int_{\mathcal{Y}}\pi(w,y)f(y|w,z;\theta)d\mu(y).$$

If a parametric model for π such as logistic regression is assumed, an estimate $\hat{\pi}$ can be obtained from the (W_i,Y_i,R_i), $i=1,\ldots,n$, by standard methods. In view of (1.11) with these estimates substituted, (1.10) becomes

$$\sum_{i=1}^{n}\Bigg(R_i\dot{\ell}_\theta(X_i,Y_i)+(1-R_i)\\ \times\frac{\sum_{j=1}^{n}\left[1\{W_j=W_i\}R_j\dot{\ell}_\theta(W_i,Z_j,Y_i)f(Y_i|W_i,Z_j;\theta)\,\big/\int\hat{\pi}(W_i,y)f(y|W_i,Z_j;\theta)d\mu(y)\,\right]}{\sum_{j=1}^{n}\left[1\{W_j=W_i\}R_jf(Y_i|W_i,Z_j;\theta)\,\big/\int\hat{\pi}(W_i,y)f(y|W_i,Z_j;\theta)d\mu(y)\,\right]}\Bigg), \tag{1.12}$$

which is called a pseudoscore. Setting (1.12) to 0 gives an estimate of θ. CCB give a heuristic discussion on the likely efficiency gain of the pseudoscore method over the mean score method.

Robins, Hsieh and Newey (1995) (RHN) derive the efficient score for estimating θ in the semiparametric model with G unspecified. (Braslow, McNeney and Wellner (2003) note a correction to their formula.) The efficient score remains the same whether the selection mechanism π is known, parametrically modeled or completely unspecified. Assuming a parametric model for π, RHN propose a class of estimators of θ which is motivated by the form of the efficient score. There is a member of this class that is semiparametrically efficient. In the case that both W and Y are finitely discrete, the efficient estimator is easy to compute. In general, however, it takes a considerable amount of computation to find the efficient estimator.

Parametric likelihood. Ibrahim, Chen and Lipsitz (1999) assume that $G(\cdot|\cdot)=G(\cdot|\cdot;\gamma)$ is known up to a Euclidean parameter γ, with conditional density $g(\cdot|\cdot;\gamma)$ with respect to some measure ν on $\mathcal{Z}$. Then the likelihood for (θ,γ) is

$$\prod_{i=1}^{n}\{f(Y_i|X_i;\theta)g(Z_i|W_i;\gamma)\}^{R_i}\left\{\int_{\mathcal{Z}}f(Y_i|W_i,z;\theta)g(z|W_i;\gamma)d\nu(z)\right\}^{1-R_i}. \tag{1.13}$$

As a consequence of the MAR assumption, the selection mechanism π is not involved. Expression (1.13) is a parametric likelihood, to which the classical maximum likelihood theory applies. Under correct model specification and some regularity conditions, θ and γ are simultaneously and efficiently estimated by maximizing (1.13). However, if the model for G is misspecified, this can lead to severely biased inference.

Semiparametric likelihood. In practice, one rarely has sufficient information to specify a parametric model for the covariate distribution, except in the trivial case that W and Z are both finitely discrete. It is then natural to consider maximizing the likelihood without specifying G. Wild (1991) explores this idea in a two-phase, outcome-dependent sampling design with a finitely discrete outcome. Assume that W is empty, so that G denotes the marginal distribution of Z. Write $\mathcal{Y} = \{y_j : j = 1, \ldots, J\}$. In phase one, a random sample $\{(Z_i, Y_i) : i = 1, \ldots, n\}$ is generated. However, one observes only the Y_i, $i = 1, \ldots, n$. Let $D_j = \{i : Y_i = y_j\}$, $j = 1, \ldots, J$. In phase two, a simple random subsample is taken from each stratum D_j for actual observation of Z. Thus the R_i are no longer independent of each other. Nevertheless the likelihood for (θ, G) takes a familiar look:

$$\prod_{i=1}^{n} [f(Y_i|Z_i;\theta)G\{Z_i\}]^{R_i} \left\{ \int_{\mathcal{Z}} f(Y_i|z;\theta)dG(z) \right\}^{1-R_i},$$

where $G\{z\}$ denotes the G-measure of the one-point set $\{z\}$. Wild (1991) then proposes estimating θ by maximizing the above semiparametric likelihood simultaneously over θ and G, the latter restricted to the (random) set of probability measures concentrated on $\{Z_i : R_i = 1\}$.

Lawless, Kalbfleisch and Wild (1999) (LKW) generalize this method in two ways. First, they consider various sampling schemes including basic stratified sampling as described in the above paragraph and Bernoulli sampling which we have considered prior to the above paragraph. To fix ideas and to keep in line with most of the foregoing discussion, we shall from now on restrict attention to Bernoulli sampling, where the (X_i, Y_i, R_i), $i = 1, \ldots, n$, are independent and identically distributed. As a second generalization, LKW relax the assumption that Y is finitely discrete and allows the selection probability to depend on covariates as well. Their key assumptions are summarized as follows. Let S_k, $k = 1, \ldots, K$, be a partition of the sample space $\mathcal{X} \times \mathcal{Y}$, let $T = \sum_{k=1}^{K} k1\{(X, Y) \in S_k\}$, and define T_i analogously, with (X, Y) replaced by (X_i, Y_i), $i = 1, \ldots, n$. Assume that T is always observed and that (X, Y) is either completely observed, in which case $R = 1$, or completely missing so that $R = 0$. Furthermore, assume that $E(R|X = x, Y = y)$ is constant on each S_k. Here we consider X as a whole and use G to denote its marginal distribution. In this setting, the semiparametric likelihood for (θ, G) is given by

$$\prod_{i=1}^{n} \left[f(Y_i|X_i;\theta)G\{X_i\} \right]^{R_i} \left\{ Q_{T_i}(\theta, G) \right\}^{1-R_i},$$

where

$$Q_k(\theta, G) := P\{(X, Y) \in S_k; \theta, G\} = \iint_{S_k} f(y|x;\theta)d\mu(y)dG(x), \qquad k = 1, \ldots, K.$$

Like Wild (1991), LKW consider estimating (θ, G) by maximizing the semiparametric likelihood with G restricted to probability distributions supported by the observed values of X. As shown by van der Vaart and Wellner (2001) and Braslow, McNeney and Wellner (2003), the LKW estimator is consistent, asymptotically normal and semiparametrically efficient.

Now, if we review the original problem posed at the beginning of this chapter with the added MAR assumption (1.7), and compare the aforementioned methods in terms of applicability and efficiency, we observe that each of them requires at least one of the following additional assumptions:

- Outcome-independent missingness;
- Parametric specification of the selection mechanism;
- Parametric specification of the conditional distribution of Z given W;
- Finite discreteness of both W and Y;
- Existence of a finite stratification of $\mathcal{W} \times \mathcal{Y}$ on which the selection mechanism is based.

It appears that, in CCB and RHN, parametric modeling of π could be replaced by nonparametric regression techniques. But we have not seen a formal justification of this conjecture. In terms of efficiency, pseudolikelihood-type methods and the complete case analysis generally do not achieve the semiparametric information bound. In contrast, (parametric or semiparametric) maximum likelihood methods often yield efficient estimators, at least under correct model specification in the parametric case. The RHN methodology can in principle lead to efficient inference; however, it is computationally difficult to actually find their efficient estimator, except when W and Y are both finitely discrete.

Here we propose a method based on maximizing a semiparametric likelihood; a precise definition is given in Chapter 2. For technical reasons, we shall assume that W is finitely discrete. We do, however, allow the distribution of Y to be arbitrary, under a general parametric model. The proposed method does not require parametric assumptions on the selection mechanism or on the covariate distribution, nor does it rely on outcome-independent missingness or a finite stratification of the sample space as in LKW. Except for a different sampling scheme, the method of Wild (1991) can be viewed as a special case of our method where Y is finitely discrete. The method developed here is related to, but different from, the LKW method. When W and Y are both finitely discrete, our problem trivially satisfies the assumptions of LKW, with each possible value of (W, Y) defining a stratum, and the two methods are equivalent. For our general problem, however, the proposed method has several notable advantages. First, the finite stratification assumption of LKW is motivated by a two-phase sampling design, where the experimenter defines the selection mechanism. This assumption may not be plausible if the missingness is unplanned, as is often the case in observational studies. The proposed method requires no such assumption. Second, even when their stratification assumption indeed holds, it may be difficult to ascertain the stratum membership if the definition of strata is unclear to the data analyst. Again, our method requires no such information. Third, LKW assume that only the stratum membership is observed for an incomplete case, and the semiparametric efficiency of their estimator is relative to this assumption. In our problem, we obtain more information by observing the exact value of (W, Y) on an incomplete case. Without making use of this extra piece of information, the LKW method is likely to be inefficient in this setting.

The rest of the dissertation is organized as follows. In Chapter 2, we discuss parameter identification and define the semiparametric MLEs. In Chapter 3, we show that the proposed estimators are consistent

and asymptotically equivalent. We explore two approaches to proving weak convergence. One is based on a linearization argument applied to a well-behaved system of likelihood equations, which is explained in Chapter 4. The other one, considered in Chapter 5, relies on a quadratic expansion of the profile log-likelihood. Chapter 6 gives concrete examples that illustrate the verification of regularity conditions. Numerical methods are derived and simulation results reported in Chapter 7. Chapter 8 outlines potential applications of the proposed method to related statistical problems. We finish this dissertation with a discussion in Chapter 9.

2.0 SEMIPARAMETRIC MLES

The distribution of (W, Z, Y, R) is completely determined by (F, G, θ, π), where F is the marginal distribution of W. Let $(F_0, G_0, \theta_0, \pi_0)$ denote the true values of parameters. F_0 and π_0 can be estimated from the observed data using standard methods. Here we focus on the estimation of (θ_0, G_0), with primary interest in θ_0. Our first result pertains to the identifiability of (θ_0, G_0) from the observed data. The observables (R, RZ, W, Y) certainly do not carry more information about (θ_0, G_0) than do the original regression variables (X, Y). Therefore it seems natural to require that

(A1) (F_0, G_0, θ_0) is identifiable from the distribution of (X, Y).

Since (F_0, G_0) is certainly identifiable, this means that $\theta = \theta_0$ whenever (X, Y) has the same distribution under (F_0, G_0, θ) as under (F_0, G_0, θ_0). In a generalized linear model, (A1) is equivalent to var(X) being positive definite. We allow Z to be missing at random but require that every possible configuration of (X, Y) be potentially observable. To be precise, assume that

(A2) (1.7) holds, with

(A3) $\pi_0(W, Y) > 0$ almost surely.

These conditions together ensure the identifiability of parameters from the observed data.

Theorem 2.1. *Under (A1)–(A3),* $(F_0, G_0, \theta_0, \pi_0)$ *is identifiable from the distribution of* (R, RZ, W, Y).

Proof. Let $(F_1, G_1, \theta_1, \pi_1)$ define the same distribution of (R, RZ, W, Y) as does $(F_0, G_0, \theta_0, \pi_0)$. It follows immediately that $\pi_1 = \pi_0$ almost surely under the (common) distribution of (W, Y). Let B be a Borel subset of $\mathcal{X} \times \mathcal{Y}$. A conditioning argument combined with (A2) and (A3) yields

$$E_j \frac{R1_B(X, Y)}{\pi_0(W, Y)} = P_j\{(X, Y) \in B\},$$

where E_j (resp. P_j) denotes expectation (resp. probability) evaluated under $(F_j, G_j, \theta_j, \pi_j)$, $j = 0, 1$. But

$$\frac{R1_B(X, Y)}{\pi_0(W, Y)} = \frac{R1_B(W, RZ, Y)}{\pi_0(W, Y)},$$

which follows the same distribution under either set of parameters. Therefore

$$P_1\{(X, Y) \in B\} = P_0\{(X, Y) \in B\}, \qquad \text{every } B.$$

It now follows from (A1) that $(F_1, G_1, \theta_1) = (F_0, G_0, \theta_0)$. □

The proposed method requires that W be finitely discrete. In fact, for notational convenience, we even assume that W is empty in the subsequent discussion. Only inessential (but tedious) modifications of results and proofs are needed to accomodate the inclusion of a finitely discrete W. Thus π is now defined on $\mathcal{Y}$ and G is just a probability measure on $\mathcal{Z} \subset \mathbb{R}^d$, the support of Z. Our inference is based on the semiparametric likelihood

$$L_n(\theta, G) = \prod_{i=1}^{n} [f(Y_i|Z_i;\theta)G\{Z_i\}]^{R_i} \{f(Y_i;G,\theta)\}^{1-R_i}, \tag{2.1}$$

where $G\{z\} := G(\{z\})$ and

$$f(y;G,\theta) := \int_{\mathcal{Z}} f(y|z;\theta)dG(z). \tag{2.2}$$

Let $\Theta \subset \mathbb{R}^q$ denote the parameter set for θ and $\mathcal{G}$ the set of all probability measures on $\mathcal{Z}$. It is natural to estimate (θ_0, G_0) by $(\tilde{\theta}, \tilde{G})$, a maximizer of the likelihood (2.1) over $(\theta, G) \in \Theta \times \mathcal{G}$. It is not obvious, however, that $(\tilde{\theta}, \tilde{G})$ exists; the supremum of L_n over $\Theta \times \mathcal{G}$ need not be achieved. Nevertheless, it will be shown that $(\tilde{\theta}, \tilde{G})$ indeed exists under the following simple conditions:

(B1) $\mathcal{Z}$ is compact;

(B2) Θ is compact;

(B3) The map $(z, \theta) \mapsto f(y|z;\theta)$ is continuous for every $y \in \mathcal{Y}$.

In this chapter, $\mathcal{G}$ is equipped with the weak topology and $\Theta \times \mathcal{G}$ the product topology. The following lemma is used in proving the existence of $(\tilde{\theta}, \tilde{G})$ and in later developments as well.

Lemma 2.2. *Under (B1), $\mathcal{G}$ is compact for the weak topology. Under (B1)–(B3), the map $(\theta, G) \mapsto f(y;G,\theta)$ is continuous for the product topology for every $y \in \mathcal{Y}$.*

Proof. The compactness of $\mathcal{G}$ follows from the compactness of $\mathcal{Z}$ and Prohorov's theorem. For the second assertion, fix $y \in \mathcal{Y}$ and take a sequence $(\theta_m, G_m) \to (\theta, G)$. Then

$$|f(y;G_m,\theta_m) - f(y;G,\theta)| \quad \leq \quad \left|\int_{\mathcal{Z}} \{f(y|z;\theta_m) - f(y|z;\theta)\}dG(z)\right| \; + \; \left|\int_{\mathcal{Z}} f(y|z;\theta_m)d(G_m - G)(z)\right|.$$

It follows from the assumptions that $f(y|\cdot;\cdot)$ is bounded. By the dominated convergence theorem, the first term on the right tends to 0 as $m \to \infty$. The assumptions also imply that $\{f(y|\cdot;\theta) : \theta \in \Theta\}$ is equicontinuous. By Theorem 1.12.1 of van der Vaart and Wellner (1996) (VW), the second term tends to 0 as well. □

Suppose now that a sample of size n has been drawn. Denote by $z_1, \ldots, z_k$ the distinct observed values of Z with respective multiplicities $n_1, \ldots, n_k$. Let $n_0 = \sum_{i=1}^{n}(1 - R_i)$, so that $\sum_{j=0}^{k} n_j = n$. Let (j, l) provide the index of the lth complete case taking the value z_j, $l = 1, \ldots, n_j$, $j = 1, \ldots, k$. Thus, for example, $(1, 1) = 1$ if and only if $R_1 = 1$ and $Z_1 = z_1$. Similarly, write $(0, l)$ for the index of the lth incomplete case.

Theorem 2.3. *Let (B1)–(B3) hold for a version of $f(\cdot|\cdot;\cdot)$ that is nonnegative everywhere. Then $(\tilde{\theta}, \tilde{G})$ exists.*

Proof. Let Λ denote the k-dimensional unit simplex:

$$\Lambda = \left\{ \lambda \in \mathbb{R}^k : \lambda_j \geq 0,\ j = 1, \ldots, k,\ \sum_{j=1}^{k} \lambda_j \leq 1 \right\},$$

where λ_j refers to the jth component of λ. It can be shown that

$$\sup_{(\theta,G)\in\Theta\times\mathcal{G}} L_n(\theta, G) = \sup_{(\theta,\lambda,G)\in\Theta\times\Lambda\times\mathcal{G}} \left[\prod_{j=1}^{k} \left\{ \lambda_j^{n_j} \prod_{l=1}^{n_j} f(Y_{(j,l)}|z_j;\theta) \right\} \right.$$
$$\left. \times \prod_{l=1}^{n_0} \left\{ \sum_{j=1}^{k} \lambda_j f(Y_{(0,l)}|z_j;\theta) + \left(1 - \sum_{j=1}^{k} \lambda_j\right) f(Y_{(0,l)};G,\theta) \right\} \right];$$

the left side is achieved if and only if the right side is. By Lemma 2.2, $\Theta \times \Lambda \times \mathcal{G}$ is compact and the function of (θ, λ, G) on the right side is continuous. □

In general, $(\tilde{\theta}, \tilde{G})$ is not unique. It is easy to see that $\tilde{G}\{z_j\} > 0$, $j = 1, \ldots, k$, unless L_n is identically 0. But $\tilde{G}$ need not concentrate on $\{z_j : j = 1, \ldots, k\}$, and the remaining mass can be distributed in rather arbitrary ways. Fortunately, $\tilde{G}$ can be modified, without changing any term in the likelihood, into a discrete distribution which is relatively easy to understand. To avoid trivialities, assume that

(B4) $f(\cdot|\cdot;\cdot)$ is positive everywhere.

This ensures that L_n is not identically 0.

Lemma 2.4. *Let (B4) hold and let $(\tilde{\theta}, \tilde{G})$ be an MLE of (θ, G). If the set*

$$\{a(f(Y_{(0,1)}|z;\tilde{\theta}), \ldots, f(Y_{(0,n_0)}|z;\tilde{\theta})) : 0 \leq a \leq 1, z \in \mathcal{Z}\} \tag{2.3}$$

is compact, then there exists $\tilde{G}^* \in \mathcal{G}$ which is supported by between k and $k + n_0$ points and such that

$$\begin{aligned} \tilde{G}^*\{z_j\} &= \tilde{G}\{z_j\}, & j &= 1, \ldots, k; \\ f(Y_{(0,l)};\tilde{G}^*,\tilde{\theta}) &= f(Y_{(0,l)};\tilde{G},\tilde{\theta}), & l &= 1, \ldots, n_0. \end{aligned} \tag{2.4}$$

In particular, $(\tilde{\theta}, \tilde{G}^*)$ is another MLE.

This follows from the argument of van der Vaart and Wellner (1992, theorem 2.1); the proof is omitted. The compactness of (2.3) certainly follows from (B1) and (B3).

Since all information about $\mathcal{Z}$ comes from the observed values of Z, it seems reasonable to consider also the *restricted* MLE $(\hat{\theta}, \hat{G})$, any maximizer of (2.1) with G restricted to the set of probability measures concentrated on $\{z_j : j = 1, \ldots, k\}$. In fact, this is common practice in semiparametric maximum likelihood estimation; compare Wild (1991) and LKW. It is easy to see that $(\hat{\theta}, \hat{G})$ exists, provided (B2) holds and the map $\theta \mapsto f(y|z;\theta)$ is continuous for every $(z, y) \in \mathcal{Z} \times \mathcal{Y}$. The latter certainly follows from (B3).

Where it is necessary to distinguish between $(\hat{\theta}, \hat{G})$ and $(\tilde{\theta}, \tilde{G})$, the latter shall be referred to as the *global* MLE. Both MLEs have intuitive appeal. In fact, it will be shown in Chapter 3 that, almost surely, the global maxima of (2.1) eventually fall into the restricted parameter set as $n \to \infty$, so that the two MLEs are

equivalent in an asymptotic sense. From a practical point of view, the computation of $(\hat{\theta}, \hat{G})$ has dimension at most $q + k$, while that of $(\tilde{\theta}, \tilde{G})$ is infinite-dimensional.

A simple characterization of $\hat{G}$ and $\tilde{G}$ can be obtained as follows. Let $h : \mathcal{Z} \to \mathbb{R}$ be bounded and measurable. For a (small) real number t, define $\hat{G}_t$ by

$$d\hat{G}_t/d\hat{G} = 1 + t(h - \int h d\hat{G}).$$

Clearly, with $|t|$ sufficiently small, $\hat{G}_t$ is a probability measure on $\mathcal{Z}$ and, like $\hat{G}$, is concentrated on $\{Z_i : R_i = 1\}$. By the definition of $\hat{G}$, the map $t \mapsto \log L_n(\hat{\theta}, \hat{G}_t)$ is maximized at $t = 0$. Differentiating with respect to t and setting the derivative equal to 0 at $t = 0$, we obtain

$$\int_{\mathcal{Z}} h d\hat{G} = \frac{1}{n} \sum_{i=1}^{n} \left\{ R_i h(Z_i) + (1 - R_i) \frac{\int_{\mathcal{Z}} h(z) f(Y_i|z; \hat{\theta}) d\hat{G}(z)}{f(Y_i; \hat{G}, \hat{\theta})} \right\}. \tag{2.5}$$

In particular, for $h = 1\{z\}$, $z \in \mathcal{Z}$, we have

$$\hat{G}\{z\} = \frac{1}{n} \sum_{i=1}^{n} \left\{ R_i 1\{Z_i = z\} + (1 - R_i) \frac{f(Y_i|z; \hat{\theta}) \hat{G}\{z\}}{f(Y_i; \hat{G}, \hat{\theta})} \right\}.$$

Simple algebraic manipulation then yields

$$\hat{G}\{z\} = \frac{\sum_{i=1}^{n} R_i 1\{Z_i = z\}}{n - \sum_{i=1}^{n} (1 - R_i) f(Y_i|z; \hat{\theta}) / f(Y_i; \hat{G}, \hat{\theta})}, \tag{2.6}$$

provided the denominator on the right side is nonzero (which is the case if the numerator is nonzero). These identities remain valid with $(\hat{\theta}, \hat{G})$ replaced by $(\tilde{\theta}, \tilde{G})$.

3.0 CONSISTENCY AND ASYMPTOTIC EQUIVALENCE

Because of the restricted nature of $\hat{G}$, one might expect $(\hat{\theta}, \hat{G})$ to be easier to analyze. Indeed, we will start our analysis with the consistency of $(\hat{\theta}, \hat{G})$, first for the weak topology on $\mathcal{G}$ and then for more general topologies. These follow from the representation (2.6) and arguments similar to those of Chen (2002, theorem 2) and van der Vaart and Wellner (2001, theorem 2). By discretizing $\tilde{G}$ as in Lemma 2.4, we then demonstrate the asymptotic equivalence of $(\hat{\theta}, \hat{G})$ and $(\tilde{\theta}, \tilde{G})$ and hence the consistency of the latter.

Recall that (G_0, θ_0, π_0) is the true value of (G, θ, π). Let E_0 denote (conditional) expectation taken under (θ_0, G_0, π_0). Let P_0 and $\mathbb{P}_n$ denote the true and empirical distributions of (Z, Y, R), respectively. We shall use operator notation for integrals, writing

$$\mathbb{P}_n h(Z, Y, R) = \mathbb{P}_n h = \int h d\mathbb{P}_n = \frac{1}{n} \sum_{i=1}^{n} h(Z_i, Y_i, R_i)$$

for instance. Let us assume that

(C0) (A1)–(A3) and (B1)–(B4) hold;

(C1) $\{f(\cdot|z;\theta) : z \in \mathcal{Z}, \theta \in \Theta\}$ is VC-subgraph;

(C2) $P_0 \sup_{z,\theta} f(Y|z,\theta) < \infty$;

(C3) $P_0 \inf_{z,\theta} \log f(Y|z,\theta) > -\infty$;

(C4) $P_0 \sup_{z,\theta,G} (1-R) f(Y|z,\theta)/f(Y;G,\theta) < \infty$;

(C5) $P_0\{R \log E_0(R|Z)\} > -\infty$.

In what follows, $(\hat{\theta}, \hat{G})$ will be subscripted with n and treated as a random sequence.

Theorem 3.1. *Let (C0)–(C5) be satisfied. Then, almost surely, $\hat{\theta}_n \to \theta_0$ and $\hat{G}_n \to G_0$ weakly as $n \to \infty$.*

Proof. Let

$$D_0(z, \theta, G) = 1 - P_0\{(1-R) f(Y|z,\theta)/f(Y;G,\theta)\}, \qquad z \in \mathcal{Z}, \theta \in \Theta, G \in \mathcal{G},$$

and define D_n similarly, with P_0 replaced by $\mathbb{P}_n$. Note that $D_0(Z, \theta_0, G_0)$ is a version of $E_0(R|Z)$. For every real-valued measurable function h,

$$\hat{G}_n h = \mathbb{P}_n\{R h(Z)/D_n(Z, \hat{\theta}_n, \hat{G}_n)\}, \tag{3.1}$$

as another way to write (2.6). This motivates the definition of G_n by

$$G_n h = \frac{\mathbb{P}_n\{Rh(Z)/D_0(Z,\theta_0,G_0)\}}{\mathbb{P}_n\{R/D_0(Z,\theta_0,G_0)\}}.$$

With probability tending to 1, G_n is a well-defined probability measure. Furthermore, by the law of large numbers, $G_n h \to G_0 h$ almost surely, for every bounded measurable function h. Take a countable collection of functions, such as in Theorem 1.12.2 of van der Vaart and Wellner (1996) (VW), to see that

$$G_n \to G_0 \qquad \text{almost surely} \tag{3.2}$$

for the weak topology (the only topology on $\mathcal{G}$ considered in this proof).

Define

$$B_0(\epsilon,\theta,G) = \{z \in \mathcal{Z} : D_0(z,\theta,G) < \epsilon\}, \qquad \epsilon > 0, \theta \in \Theta, G \in \mathcal{G}. \tag{3.3}$$

Lemma 2.2 ensures that $\mathcal{G}$ is compact, hence separable. Let Θ_0 and $\mathcal{G}_0$ be countable and dense in Θ and $\mathcal{G}$, respectively, and let $\mathbb{Q}^+$ denote the set of positive rational numbers. By the law of large numbers, we have that, almost surely,

$$\mathbb{P}_n 1_{B_0(\epsilon,\theta,G)}(Z) \to G_0(B_0(\epsilon,\theta,G)), \qquad \epsilon \in \mathbb{Q}^+, \theta \in \Theta_0, G \in \mathcal{G}_0. \tag{3.4}$$

For later use, we also notice that

$$\mathbb{P}_n R \log\left[D_0(Z,\theta_0,G_0)\mathbb{P}_n\left\{R/D_0(Z,\theta_0,G_0)\right\}\right] \to P_0\{R \log E_0(R|Z)\} \tag{3.5}$$

almost surely, by (C5) and the law of large numbers.

Fix an ω in the probability space which satisfies (3.2), (3.4), (3.5) and the (countably many) Glivenko-Cantelli (GC) properties established in Lemma 3.2 ahead. (Almost every ω qualifies.) It will be shown that, at this ω, $(\hat\theta_n, \hat G_n) \to (\theta_0, G_0)$. By the compactness of $\Theta \times \mathcal{G}$, it suffices to show that every limit point (θ', G') of $(\hat\theta_n, \hat G_n)$ is equal to (θ_0, G_0). Abusing notation, we still use (n) to denote the convergent subsequence.

It follows from the definition of $(\hat\theta_n, \hat G_n)$ that

$$\begin{aligned} 0 &\geq \frac{1}{n}\log\frac{L_n(\theta_0,G_n)}{L_n(\hat\theta_n,\hat G_n)} \\ &= \mathbb{P}_n\left[R\log\frac{f(Y|Z;\theta_0)}{f(Y|Z;\hat\theta_n)} + R\log\frac{D_n(Z,\hat\theta_n,\hat G_n)}{D_0(Z,\theta_0,G_0)\mathbb{P}_n\left\{R/D_0(Z,\theta_0,G_0)\right\}} + (1-R)\log\frac{f(Y;G_n,\theta_0)}{f(Y;\hat G_n,\hat\theta_n)}\right]. \end{aligned} \tag{3.6}$$

By Lemma 3.2 and the dominated convergence theorem,

$$\mathbb{P}_n R\log\frac{f(Y|Z;\theta_0)}{f(Y|Z;\hat\theta_n)} \to P_0 R\log\frac{f(Y|Z;\theta_0)}{f(Y|Z;\theta')}.$$

Similarly, we have

$$\mathbb{P}_n(1-R)\log\frac{f(Y;G_n,\theta_0)}{f(Y;\hat G_n,\hat\theta_n)} \to P_0(1-R)\log\frac{f(Y;G_0,\theta_0)}{f(Y;G',\theta')}.$$

Suppose for the moment that

$$G_0 \ll G', \tag{3.7}$$

$$\liminf \mathbb{P}_n R \log \frac{D_n(Z, \hat{\theta}_n, \hat{G}_n)}{D_0(Z, \theta_0, G_0)\mathbb{P}_n \{R/D_0(Z, \theta_0, G_0)\}} \geq P_0 R \log \frac{dG_0}{dG'}(Z). \tag{3.8}$$

Then the right side of (3.6) will have limit inferior at least

$$\begin{aligned}
&P_0 \left\{ R \log \frac{f(Y|Z;\theta_0)}{f(Y|Z;\theta')} + R \log \frac{dG_0}{dG'}(Z) + (1-R) \log \frac{f(Y;G_0,\theta_0)}{f(Y;G',\theta')} \right\} \\
=&P_0 \left\{ R \log \frac{f(Y|Z;\theta_0)\frac{dG_0}{dG'}(Z)\,/f(Y;G_0,\theta_0)}{f(Y|Z;\theta')\,/f(Y;G',\theta')} + \log \frac{f(Y;G_0,\theta_0)}{f(Y;G',\theta')} \right\} \\
=&E_0 \left[\pi_0(Y) E_0 \left\{ \log \frac{f(Y|Z;\theta_0)\frac{dG_0}{dG'}(Z)\,/f(Y;G_0,\theta_0)}{f(Y|Z;\theta')\,/f(Y;G',\theta')} \middle| Y \right\} \right] + E_0 \log \frac{f(Y;G_0,\theta_0)}{f(Y;G',\theta')}.
\end{aligned} \tag{3.9}$$

By Jensen's inequality, both terms in (3.9) are ≥ 0 (hence $= 0$). The first term being 0 implies that

$$f(y|z;\theta_0)\frac{dG_0}{dG'}(z)\,/f(y;G_0,\theta_0) = f(y|z;\theta')\,/f(y;G',\theta') \tag{3.10}$$

for $G' \times P_0^Y$-almost every (z, y), where P_0^Y denotes the true distribution of Y. By (B4), $\mu \ll P_0^Y$; hence (3.10) actually holds for $G' \times \mu$-almost every (z, y). The second term in (3.9) being 0 implies that

$$f(y;G_0,\theta_0) = f(y;G',\theta'), \qquad \mu\text{-almost every } y.$$

Multiplying the two equations above, we obtain

$$f(y|z;\theta') = f(y|z;\theta_0)\frac{dG_0}{dG'}(z), \qquad G' \times \mu\text{-almost every } (z, y). \tag{3.11}$$

Integrating both sides with respect to $d\mu(y)$ gives

$$\frac{dG_0}{dG'}(z) = 1, \qquad G'\text{-almost every } z,$$

whence $G' = G_0$. In view of this and (3.7), (3.11) becomes

$$f(y|z;\theta') = f(y|z;\theta_0), \qquad G_0 \times \mu\text{-almost every } (z, y).$$

It now follows from (A1) that $\theta' = \theta_0$.

Thus the proof will be complete as soon as (3.7) and (3.8) are verified. To this end, observe that

$$\int_{(-\infty,t]} D_n(z, \hat{\theta}_n, \hat{G}_n) d\hat{G}_n(z) = \mathbb{P}_n R1\{Z \leq t\} \to P_0 R1\{Z \leq t\} = \int_{(-\infty,t]} D_0(z, \theta_0, G_0) dG_0(z) \tag{3.12}$$

for every $t \in \mathbb{R}^d$ (in fact, uniformly). On the other hand, for every continuity point t of G',

$$\begin{aligned}
\text{LHS of (3.12)} &= \hat{G}_n(t) - \mathbb{P}_n \frac{(1-R)\int_{(-\infty,t]} f(Y|z;\hat{\theta}_n) d\hat{G}_n(z)}{f(Y;\hat{G}_n,\hat{\theta}_n)} \\
&= \hat{G}_n(t) - P_0 \frac{(1-R)\int_{(-\infty,t]} f(Y|z;\hat{\theta}_n) d\hat{G}_n(z)}{f(Y;\hat{G}_n,\hat{\theta}_n)} + o(1) \\
&\to G'(t) - P_0 \frac{(1-R)\int_{(-\infty,t]} f(Y|z;\theta') dG'(z)}{f(Y;G',\theta')} \\
&= \int_{(-\infty,t]} D_0(z, \theta', G') dG'(z),
\end{aligned}$$

where the first step follows from Fubini's theorem, the second from Lemma 3.2, the third from Lemma 3.4 and the dominated convergence theorem, and the last from Fubini's theorem again. Combining the two displays above, we have

$$\int_{(-\infty,t]} D_0(z,\theta',G')dG'(z) = \int_{(-\infty,t]} D_0(z,\theta_0,G_0)dG_0(z) \tag{3.13}$$

for every continuity point t of G'. As functions of t, the left side defines a finite (by (C4)) signed measure and the right side a finite measure; hence both are right continuous in t. For a discontinuity point t of G', take a sequence $t_m \downarrow t$ to see that (3.13) remains valid. Therefore both sides define the same measure and, by the π-λ theorem, their integrands are the respective Radon-Nikodym derivatives. In particular, the integrand on the left is nonnegative G'-almost everywhere. By (A3) and (B4), the integrand on the right is even positive everywhere. Thus (3.7) follows, with

$$\frac{dG_0}{dG'}(z) = \frac{D_0(z,\theta',G')}{D_0(z,\theta_0,G_0)}. \tag{3.14}$$

It remains to verify (3.8). In view of (3.5) and (3.14), it suffices to show that

$$\liminf \mathbb{P}_n R \log D_n(Z,\hat{\theta}_n,\hat{G}_n) \geq P_0 R \log D_0(Z,\theta',G').$$

By Lemma 3.2,

$$D_n(z,\hat{\theta}_n,\hat{G}_n) - D_0(z,\hat{\theta}_n,\hat{G}_n) \to 0 \tag{3.15}$$

uniformly in z. Let $\epsilon \in \mathbb{Q}^+$. Then

$$\mathbb{P}_n R \log\{D_n(Z,\hat{\theta}_n,\hat{G}_n) \vee \epsilon\} - \mathbb{P}_n R \log\{D_0(Z,\hat{\theta}_n,\hat{G}_n) \vee \epsilon\} \to 0,$$

where $a \vee b := \max\{a,b\}$. Also by Lemma 3.2,

$$(\mathbb{P}_n - P_0) R \log\{D_0(Z,\hat{\theta}_n,\hat{G}_n) \vee \epsilon\} \to 0.$$

The continuity of D_0 (shown in the proof of Lemma 3.2) and the dominated convergence theorem together imply

$$P_0 R \log\{D_0(Z,\hat{\theta}_n,\hat{G}_n) \vee \epsilon\} \to P_0 R \log\{D_0(Z,\theta',G') \vee \epsilon\}.$$

Combining the three displays above gives

$$\mathbb{P}_n R \log\{D_n(Z,\hat{\theta}_n,\hat{G}_n) \vee \epsilon\} \to P_0 R \log\{D_0(Z,\theta',G') \vee \epsilon\}. \tag{3.16}$$

Similarly to (3.3), define

$$B_n(\epsilon,\theta,G) = \{z \in \mathcal{Z} : D_n(z,\theta,G) < \epsilon\}.$$

Now we can write

$$\mathbb{P}_n R \log D_n(Z, \hat{\theta}_n, \hat{G}_n) = \mathbb{P}_n R \log\{D_n(Z, \hat{\theta}_n, \hat{G}_n) \vee \epsilon\} + \mathbb{P}_n \left[R1_{B_n(\epsilon, \hat{\theta}_n, \hat{G}_n)}(Z) \left\{ \log D_n(Z, \hat{\theta}_n, \hat{G}_n) - \log \epsilon \right\} \right].$$

In view of (3.16), we have, for $\epsilon < 1$,

$$\liminf \mathbb{P}_n R \log D_n(Z, \hat{\theta}_n, \hat{G}_n) \geq P_0 R \log D_0(Z, \theta', G') + \liminf \mathbb{P}_n R 1_{B_n(\epsilon, \hat{\theta}_n, \hat{G}_n)}(Z) \log D_n(Z, \hat{\theta}_n, \hat{G}_n), \quad (3.17)$$

and it suffices to show that the last term becomes nonnegative as $\epsilon \to 0$. To this end, notice that, by Jensen's inequality,

$$\mathbb{P}_n R 1_{B_n(\epsilon, \hat{\theta}_n, \hat{G}_n)}(Z) \log D_n(Z, \hat{\theta}_n, \hat{G}_n) \geq -\mathbb{P}_n\{R 1_{B_n(\epsilon, \hat{\theta}_n, \hat{G}_n)}(Z)\} \times \log \frac{\mathbb{P}_n\{R 1_{B_n(\epsilon, \hat{\theta}_n, \hat{G}_n)}(Z)/D_n(Z, \hat{\theta}_n, \hat{G}_n)\}}{\mathbb{P}_n\{R 1_{B_n(\epsilon, \hat{\theta}_n, \hat{G}_n)}(Z)\}}. \quad (3.18)$$

But

$$\mathbb{P}_n\{R 1_{B_n(\epsilon, \hat{\theta}_n, \hat{G}_n)}(Z)/D_n(Z, \hat{\theta}_n, \hat{G}_n)\} = \hat{G}_n(B_n(\epsilon, \hat{\theta}_n, \hat{G}_n)) \leq 1,$$

so that the left side of (3.18) is at least

$$\mathbb{P}_n\{R 1_{B_n(\epsilon, \hat{\theta}_n, \hat{G}_n)}(Z)\} \log \mathbb{P}_n\{R 1_{B_n(\epsilon, \hat{\theta}_n, \hat{G}_n)}(Z)\}.$$

The continuity of D_0 and the compactness of $\mathcal{Z}$ together imply that $\{D_0(z, \cdot, \cdot) : z \in \mathcal{Z}\}$ is equcontinuous. This, together with (3.15), implies that

$$D_n(z, \hat{\theta}_n, \hat{G}_n) \to D_0(z, \theta', G') \quad (3.19)$$

uniformly in z. It follows that, for large n,

$$B_n(\epsilon, \hat{\theta}_n, \hat{G}_n) \subset B_0(2\epsilon, \theta', G'). \quad (3.20)$$

Also by the equicontinuity mentioned above, there exists $(\theta'', G'') \in \Theta_0 \times \mathcal{G}_0$ (defined before (3.4)) such that

$$\sup_{z \in \mathcal{Z}} |D_0(z, \theta', G') - D_0(z, \theta'', G'')| < \epsilon,$$

whence

$$B_0(2\epsilon, \theta', G') \subset B_0(3\epsilon, \theta'', G'') \subset B_0(4\epsilon, \theta', G'). \quad (3.21)$$

Combining (3.20) and (3.21), we see that

$$0 \leq \limsup \mathbb{P}_n\{R 1_{B_n(\epsilon, \hat{\theta}_n, \hat{G}_n)}(Z)\} \leq \lim \mathbb{P}_n\{R 1_{B_0(3\epsilon, \theta'', G'')}(Z)\} = P_0\{R 1_{B_0(3\epsilon, \theta'', G'')}(Z)\} \leq P_0\{R 1_{B_0(4\epsilon, \theta', G')}(Z)\} \leq G_0(B_0(4\epsilon, \theta', G')).$$

As $\epsilon \to 0$, the last quantity tends to $G_0\{z \in \mathcal{Z} : D_0(z, \theta', G') \leq 0\}$, which is 0 by (3.14). This, together with the fact that $\lim_{x \to 0+} x \log x = 0$, shows that the last term in (3.17) can be removed. □

Lemma 3.2. *In the situation of Theorem 3.1, the following classes of functions of (Z, Y, R) are P_0-GC:*

$$\{R1_{(-\infty,t]}(Z) : t \in \mathbb{R}^d\},$$
$$\{R \log f(Y|Z,\theta) : \theta \in \Theta\},$$
$$\{R \log f(Y;G,\theta) : \theta \in \Theta, G \in \mathcal{G}\},$$
$$\{R \log\{D_0(Z,\theta,G) \vee \epsilon\} : \theta \in \Theta, G \in \mathcal{G}\},$$
$$\{(1-R)f(Y|z;\theta)/f(Y;G,\theta) : z \in \mathcal{Z}, \theta \in \Theta, G \in \mathcal{G}_d\},$$
$$\left\{\frac{(1-R)}{f(Y;G,\theta)} \int_{(-\infty,t]} f(Y|z,\theta)dG(z) : t \in \mathbb{R}^d, \theta \in \Theta, G \in \mathcal{G}_d\right\},$$

where $\epsilon \in \mathbb{Q}^+$ and $\mathcal{G}_d$ denotes the collection of finitely discrete probability measures on $\mathcal{Z}$.

Proof. The first GC property follows from the classical Glivenko-Cantelli theorem and the GC preservation theorem (van der Vaart and Wellner, 2000, theorem 3). The second follows from the compactness, continuity and integrability assumptions and Lemma 3.3 below. Lemma 2.2 says that $\mathcal{G}$ (hence $\Theta \times \mathcal{G}$) is compact and that the map $(\theta, G) \mapsto f(y;G,\theta)$ is continuous for every y. This, together with Lemma 3.3, proves the GC property of the third display. By (C4) and the dominated convergence theorem, D_0 is continuous (in all three arguments together). Thus Lemma 3.3 applies to the fourth display as well. The last two require a different approach. The class $\{f(\cdot|z,\theta) : z \in \mathcal{Z}, \theta \in \Theta\}$ is VC, has an integrable envelope and can be shown to be pointwise seperable, hence GC. Its convex hull (also GC) contains $\{f(\cdot;G,\theta) : \theta \in \Theta, G \in \mathcal{G}_d\}$; so the latter too is GC. The GC property of the fifth display now follows from the GC preservation theorem and the integrability condition (C4). For the last display, it suffices to note that it is contained in the convex hull of the fifth GC class. □

Lemma 3.3. *Let T be a compact semi-metric space and $\mathcal{F} = \{f_t : t \in T\}$ a class of measurable functions with an integrable envelope. Suppose that for P-almost every x, the map $t \mapsto f_t(x)$ is continuous. Then $\mathcal{F}$ is P-GC.*

Proof. For each $t \in T$,

$$\left\| \sup_{d(t',t)<\delta} f_{t'} - \inf_{d(t',t)<\delta} f_{t'} \right\|_{P,1} \to 0, \qquad \delta \to 0,$$

by the dominated convergence theorem. Fix $\epsilon > 0$. Then the above norm can be made $< \epsilon$ for some $\delta_t > 0$ for every t. Now write $T = \bigcup_{t \in T} B(t, \delta_t)$, where $B(t,\delta)$, $\delta > 0$, denotes the open ball in T with center t and radius δ. Since T is compact, this can be rewritten as a finite union: $T = \bigcup_{j=1}^{J} B(t_j, \delta_{t_j})$ for some $J \in \mathbb{N}$ and $t_1, \ldots, t_J \in T$. It follows that $\mathcal{F}$ is covered by finitely many $L_1(P)$-brackets of size $< \epsilon$:

$$\left[\inf_{t \in B(t_j,\delta_{t_j})} f_t, \sup_{t \in B(t_j,\delta_{t_j})} f_t\right], \qquad j = 1, \ldots, J.$$

□

Lemma 3.4. *Let (B1)–(B4) be satisfied and let $\mathcal{G}$ be equipped with the weak topology. If $(\theta_m, G_m) \to (\theta, G)$ and t is a continuity point of G, then*

$$\int_{(-\infty,t]} f(y|z,\theta_m)dG_m(z) \to \int_{(-\infty,t]} f(y|z,\theta)dG(z), \qquad y \in \mathcal{Y}. \tag{3.22}$$

Proof. Fix $y \in \mathcal{Y}$. Define the probability measures $\bar{G}_m$ and $\bar{G}$ by

$$\frac{d\bar{G}_m}{dG_m}(z) = \frac{f(y|z;\theta_m)}{f(y;G_m,\theta_m)} \qquad \text{and} \qquad \frac{d\bar{G}}{dG}(z) = \frac{f(y|z;\theta)}{f(y;G,\theta)}.$$

Let $h : \mathcal{Z} \to \mathbb{R}$ be bounded and continuous and write

$$\bar{G}_m h = \frac{\int h(z)f(y|z;\theta_m)dG_m(z)}{f(y;G_m,\theta_m)}.$$

By Lemma 2.2, the denominator tends to $f(y;G,\theta)$ as $m \to \infty$. Noting that $\{h(\cdot)f(y|\cdot;\theta) : \theta \in \Theta\}$ is equicontinuous, the same technique applies to the numerator and yields

$$\bar{G}_m h \to \frac{\int h(z)f(y|z;\theta)dG(z)}{f(y;G,\theta)} = \bar{G}h.$$

This is true for every h, whence $\bar{G}_m \to \bar{G}$. Each continuity point t of G is also a continuity point of $\bar{G}$. Therefore,

$$\text{LHS of (3.22)} = \bar{G}_m(-\infty,t]f(y;G_m,\theta_m) \to \bar{G}(-\infty,t]f(y;G,\theta) = \text{RHS of (3.22)}.$$

□

Now let us consider other topologies on $\mathcal{G}$. Let $\mathcal{H}$ be a class of real-valued measurable functions on $\mathcal{Z}$ that are uniformly bounded. Each $G \in \mathcal{G}$ defines an element of $\ell^\infty(\mathcal{H})$ by $h \in \mathcal{H} \mapsto \int h dG$. Thus $\mathcal{G}$ can be viewed as a subset of $\ell^\infty(\mathcal{H})$ with

$$\|G\|_{\mathcal{H}} := \sup_{h\in\mathcal{H}} \left| \int h dG \right|, \qquad G \in \mathcal{G}.$$

One example of $\mathcal{H}$ is $C_1^1(\mathcal{Z})$, the collection of real-valued functions on $\mathcal{Z}$ that are uniformly bounded by 1 and are Lipschitz with Lipschitz constant at most 1. It is well known that, with $\mathcal{H} = C_1^1(\mathcal{Z})$, $\|\cdot\|_{\mathcal{H}}$ generates the weak topology on $\mathcal{G}$; cf. Theorem 1.12.4 of VW. As a second example, if $\mathcal{H} = \{1(-\infty,t] : t \in \mathbb{R}^d\}$, then $\|G\|_{\mathcal{H}} = 1$ for every $G \in \mathcal{G}$ and convergence for $\|\cdot\|_{\mathcal{H}}$ corresponds to uniform convergence of distribution functions. In what follows we even write $\|\cdot\|_{\mathcal{H}}$ for an $\mathcal{H}$ that is not uniformly bounded, provided the norm is well defined for the (signed) measures of interest.

The D_n and D_0 notations defined in the proof of Theorem 3.1 will continue to be used. Assume that

(C5′) $\inf_{z\in\mathcal{Z}} D_0(z,\theta_0,G_0) > 0$.

This is a stronger version of (C5), and enables us to strengthen the consistency of $\hat{G}_n$.

Theorem 3.5. *Let (C0)–(C4) and (C5′) hold and let $\mathcal{H}$ be an $L_1(G_0)$-bounded GC class. Then, almost surely, $\hat{\theta}_n \to \theta_0$ and $\|\hat{G}_n - G_0\|_{\mathcal{H}} \to 0$.*

Proof. Given the conclusions of Theorem 3.1, we only need to consider $\hat{G}_n$. For each $h \in \mathcal{H}$, write

$$\begin{aligned}(\hat{G}_n - G_0)h &= \mathbb{P}_n \frac{Rh(Z)}{D_n(Z, \hat{\theta}_n, \hat{G}_n)} - P_0 \frac{Rh(Z)}{E_0(R|Z)} \\ &= \mathbb{P}_n \frac{Rh(Z)\{D_0(Z, \theta_0, G_0) - D_n(Z, \hat{\theta}_n, \hat{G}_n)\}}{D_n(Z, \hat{\theta}_n, \hat{G}_n) D_0(Z, \theta_0, G_0)} + (\mathbb{P}_n - P_0) \frac{Rh(Z)}{D_0(Z, \theta_0, G_0)}.\end{aligned}$$

The $L_1(G_0)$-boundedness of $\mathcal{H}$ implies the existence of an integrable envelope H; cf. VW (p. 125). It follows that

$$\|\hat{G}_n - G_0\|_{\mathcal{H}} \le \mathbb{P}_n H \sup_{z \in \mathcal{Z}} \left| \frac{D_0(z, \theta_0, G_0) - D_n(z, \hat{\theta}_n, \hat{G}_n)}{D_n(z, \hat{\theta}_n, \hat{G}_n) D_0(z, \theta_0, G_0)} \right| + \left\| (\mathbb{P}_n - P_0) \frac{Rh(Z)}{D_0(Z, \theta_0, G_0)} \right\|_{\mathcal{H}}.$$

By the law of large numbers, $\mathbb{P}_n H = O(1)$ almost surely. (3.19) and (C5′) together imply that the supremum on the right tends to 0 almost surely. By the GC preservation theorem, the class

$$\{Rh(Z)/D_0(Z, \theta_0, G_0) : h \in \mathcal{H}\}$$

is GC. Hence the last term on the right tends to 0 almost surely. □

We are now ready to establish the asymptotic equivalence of $(\hat{\theta}_n, \hat{G}_n)$ and $(\tilde{\theta}_n, \tilde{G}_n)$. For each n, let $\tilde{\mathcal{E}}_n$ denote the collection of global MLEs, that is,

$$\tilde{\mathcal{E}}_n = \{(\theta, G) \in \Theta \times \mathcal{G} : L_n(\theta, G) = \max_{\theta', G'} L_n(\theta', G')\}.$$

Define $\hat{\mathcal{E}}_n$ analogously.

Theorem 3.6. *Let (C0)–(C4) and (C5′) be satisfied. Then, almost surely, $\hat{\mathcal{E}}_n = \tilde{\mathcal{E}}_n$ for large n. In particular, $\tilde{\theta}_n \to \theta_0$ and $\|\tilde{G}_n - G_0\|_{\mathcal{H}} \to 0$ almost surely for every sequence $(\tilde{\theta}_n, \tilde{G}_n)$ of global MLEs and every $L_1(G_0)$-bounded GC class $\mathcal{H}$.*

Proof. Fix an ω in the probability space as in Theorem 3.1. We show that, at this ω, $\hat{\mathcal{E}}_n = \tilde{\mathcal{E}}_n$ for large n. In fact, it suffices to show that $\tilde{\mathcal{E}}_n \subset \hat{\mathcal{E}}_n$ for large n. By contradiction, assume there is a sequence $(\tilde{\theta}_n, \tilde{G}_n)$ of global MLEs such that $(\tilde{\theta}_n, \tilde{G}_n) \notin \hat{\mathcal{E}}_n$ for infinitely many n. For each n, let $\tilde{G}_n^*$ be a discrete modification of $\tilde{G}_n$ given by Lemma 2.4. $\tilde{G}_n^*$ does not admit a representation of the form (3.1). Nevertheless the arguments in the proof of Theorem 3.1 remain valid with $(\hat{\theta}_n, \hat{G}_n)$ replaced by $(\tilde{\theta}_n, \tilde{G}_n^*)$. A key step in verifying this is the first equality in (3.12). For the latter, it suffices to note that every z such that $D_n(z, \tilde{\theta}_n, \tilde{G}_n^*) \neq 0$ satisfies (2.6) with $(\hat{\theta}, \hat{G})$ replaced by $(\tilde{\theta}_n, \tilde{G}_n^*)$. As in Theorem 3.1, we conclude that $\tilde{\theta}_n \to \theta_0$ and $\tilde{G}_n^* \to G_0$ weakly. As argued in that same proof, $D_n(z, \tilde{\theta}_n, \tilde{G}_n^*) \to D_0(z, \theta_0, G_0)$ uniformly in z. It then follows from C5′ that $D_n(z, \tilde{\theta}_n, \tilde{G}_n^*) \neq 0$ for any $z \in \mathcal{Z}$ for large n. In that case (2.6) implies that $\tilde{G}_n^*$ has no discrete components other than the observed values of Z. As a discrete measure, $\tilde{G}_n^*$ must then be concentrated on the obcerved values of Z. By (2.4), the same can be said about $\tilde{G}_n$. The global MLE $(\tilde{\theta}_n, \tilde{G}_n)$ then qualifies as a restricted MLE. This is true for all large n, which contradicts the hypothesis about $(\tilde{\theta}_n, \tilde{G}_n)$. □

The assumptions we made on the selection mechanism are simple and reasonably weak; (A2), (A3) and (C5′) are weaker than conditions (2a) and (2b) of Robins, Hsieh and Newey (1995). Their plausibility depends on the application at hand. The verification of the assumptions on the regression model, especially (C1) and (C3), is facilitated if the conditional distribution of Y given Z belongs to an exponential family. Suppose we can write $f(y|z;\theta) = \exp\{\sum_{j=1}^{J} C_j(z,\theta)T_j(y)\}$. Then the class $\{\log f(\cdot|z,\theta) : z \in \mathcal{Z}, \theta \in \Theta\}$ is finite-dimensional and VC by Lemma 2.6.15 of VW. This VC property is preserved under a monotone transformation (Lemma 2.6.18 (viii) of VW). Thus (C1) is verified. If, for each j, C_j is continuous and $T_j(Y)$ has finite mean, then (C3) follows from (B1) and (B2). Condition (C4) appears more difficult to verify. If $\mathcal{Y}$ is finite, such as in logistic regression, then (C4) follows from (B1)–(B4). It can also be shown that, under (B1) and (B2), (C4) holds for the normal linear model and any Poisson regression model where the natural parameter is expressible as a continuous function of the systematic component (see Chapter 6). In general, however, the validity of (C4) may depend on the selection mechanism and even the true values of regression parameters. It would be desirable to relax or remove this condition with a more sophisticated argument.

4.0 ASYMPTOTIC NORMALITY VIA LINEARIZATION

In this section we show that $(\hat{\theta}_n, \hat{G}_n)$ is asymptotically normal and that $\hat{\theta}_n$ achieves the semiparametric information bound. (In view of Theorem 3.6, the same can then be said about $(\tilde{\theta}_n, \tilde{G}_n)$.) These results are deduced from Theorem 3.3.1 and Lemma 3.3.5 (also known as the Z-theorem) of van der Vaart and Wellner (1996) (VW). The latter essentially follows from a linearization argument applied to a well-behaved system of likelihood equations and can be seen as an extension to infinite-dimensional parameters of similar results for finite-dimensional parameters. We will derive a representation of the efficient score for estimating θ_0, construct a system of likelihood equations, and then apply the Z-theorem. Unfortunately, it seems difficult to characterize the hypotheses of the Z-theorem in terms of simple conditions on the regression model while maintaining a sufficient degree of generality. As a consequence, our main results will involve abstract conditions. We do, however, suggest some techniques for verifying these conditions and give illustrative examples in Chapter 6.

4.1 INFORMATION CALCULATION

We shall start with a calculation of the information bound for estimating θ, which is a criterion for determining the efficiency of an estimator. A systematic exposition of the semiparametric efficiency theory is given in, for example, the monograph of Bickel, Klaassen, Ritov and Wellner (1993). For the problem considered here, Robins, Hsieh and Newey (1995) (RHN) have derived a representation of the efficient score for θ as a functional of the unique solution to an integral equation, and Braslow, McNeney and Wellner (2003) have noted a minor correction to their formula. Here we derive an alternative representation of the efficient score for θ which is convenient to use in our arguments.

Let us assume that

(D1) θ_0 is an interior point of Θ;

(D2) There is a neighborhood of θ_0 where $f(y|z;\cdot)$ is twice differentiable for every (y, z) with derivatives $\dot{f}(y|z;\cdot)$ and $\ddot{f}(y|z;\cdot)$.

With (Z, Y) completely observed, the score for θ at θ_0 is $\dot{\ell}^{ZY}_{\theta_0}$, where

$$\dot{\ell}^{ZY}_{\theta}(z, y) := \frac{\partial}{\partial\theta} \log f(y|z;\theta) = \frac{\dot{f}}{f}(y|z;\theta),$$

for θ in the aforementioned neighborhood of θ_0. (In this context, the $\dot{\ell}_\theta$ notation defined in Chapter 1 may create confusion and therefore is abandoned.) On the other hand, if Z is unobserved but $G = G_0$ is known, then the score for θ at θ_0 is $\dot{\ell}^{Y}_{\theta_0,G_0}$, where

$$\dot{\ell}^{Y}_{\theta,G}(y) := \frac{\partial}{\partial\theta} \log f(y; G, \theta) = \frac{\int \dot{\ell}^{ZY}_{\theta}(z, y) f(y|z;\theta) dG(z)}{f(y; G, \theta)}.$$

$\dot{\ell}^{Y}_{\theta_0,G_0}$ is well defined under the following assumption:

(D3) $\dot{\ell}^{ZY}_{\theta_0}$ consists of linearly independent vectors in $L_2(P_0)$.

In that case, $\dot{\ell}^{Y}_{\theta_0,G_0}$ is the componentwise projection of $\dot{\ell}^{ZY}_{\theta_0}$ into $L_2(P^Y_0)$, where P^Y_0 denotes the true distribution of Y. In general, R is random and the observed variables can be written as (R, RZ, Y). If $G = G_0$ is known, then the score for θ at θ_0 can be found, by differentiating the log-density, to be $\dot{\ell}_{\theta_0,G_0}$, where

$$\dot{\ell}_{\theta,G}(z, y, r) := r\dot{\ell}^{ZY}_{\theta}(z, y) + (1 - r)\dot{\ell}^{Y}_{\theta,G}(y),$$

provided the right side exists.

More generally, let $P^Y_{\theta,G}$ denote the distribution of Y under (θ, G); similarly for $P^{ZY}_{\theta,G}$, etc. Define the conditional expectation operators $\Pi^Y_{\theta,G} : L_2(P_{\theta,G}) \to L_2(P^Y_{\theta,G})$ by

$$\begin{aligned} \Pi^Y_{\theta,G} h(y) &= E_{\theta,G}\{h(Z, Y, R)|Y = y\} \\ &= \frac{\int_{\mathcal{Z}} \{\pi_0(y)h(z, y, 1) + (1 - \pi_0(y))h(z, y, 0)\} f(y|z;\theta) dG(z)}{f(y; G, \theta)}, \end{aligned} \tag{4.1}$$

and $\Pi^Z_\theta : L_2(P_{\theta,G}) \to L_2(G)$ by

$$\begin{aligned} \Pi^Z_\theta h(z) &= E_\theta\{h(Z, Y, R)|Z = z\} \\ &= \int_{\mathcal{Y}} \{\pi_0(y)h(z, y, 1) + (1 - \pi_0(y))h(z, y, 0)\} f(y|z;\theta) d\mu(y). \end{aligned} \tag{4.2}$$

Let $A_{\theta,G} : L_2(P_{\theta,G}) \to L_2(P_{\theta,G})$ be defined by

$$A_{\theta,G} h(z, y, r) = rh(z, y, r) + (1 - r)\Pi^Y_{\theta,G} h(y).$$

It is easy to see that $\Pi^Y_{\theta,G}$, Π^Z_θ and $A_{\theta,G}$ as Hilbert space operators are linear and continuous. Here and in the sequel, we use the subscript 0 to denote either θ_0 or (θ_0, G_0), depending on the context. Thus, for example,

$$\dot{\ell}^Y_0 = \Pi^Y_0 \dot{\ell}^{ZY}_0 \qquad \text{and} \qquad \dot{\ell}_0 = A_0 \dot{\ell}^{ZY}_0.$$

The efficient score function for θ at θ_0 is given by $\dot{\ell}_0$ minus its projection into the tangent space Γ for (G, π) at (G_0, π_0). Recall that Γ is defined as the closed linear span of the set of scores at (G_0, π_0) in all

regular one-dimensional submodels passing through (G_0, π_0) with θ fixed at θ_0. For reasons that will become clear later, it suffices to consider regular one-dimensional submodels passing through G_0 with (θ, π) fixed at (θ_0, π_0). Such a submodel can be constructed as in Chapter 2. Let $h : \mathcal{Z} \to \mathbb{R}$ be bounded and measurable. For a (small) real number t, define G_t by

$$dG_t/dG_0 = 1 + t(h - G_0 h).$$

Note that when $t = 0$, G_t as defined above is just G_0, so that no ambiguity arises. Clearly, with $|t|$ sufficiently small, G_t is a probability measure on $\mathcal{Z}$. By differentiating the log-likelihood with respect to t and evaluating the derivative at $t = 0$, the score at $t = 0$ is found to be $B_0 h - G_0 h$, where $B_{\theta,G}$ is the restriction of $A_{\theta,G}$ to $L_2(G)$. For a general $h \in L_2(G_0)$, there is a sequence (h_m) of bounded measurable functions such that $h_m \to h$ in $L_2(G_0)$. It follows that $G_0 h_m \to G_0 h$ and, by the continuity of B_0, also that $B_0 h_m \to B_0 h$ in $L_2(P_0)$. In view of the closedness of Γ, we now have

$$\Gamma \supset \{B_0 h - G_0 h : h \in L_2(G_0)\} = B_0 L_2^0(G_0) = \mathcal{R}(B_0) \cap L_2^0(P_0), \tag{4.3}$$

where $\mathcal{R}(\cdot)$ denotes the range of an operator, $L_2^0(G_0) := \{h \in L_2(G_0) : G_0 h = 0\}$, and similarly for $L_2^0(P_0)$.

Let A_0^* and B_0^* denote the respective Hilbert adjoint operators of A_0 and B_0. Then $A_0^* : L_2(P_0) \to L_2(P_0)$ is given by

$$\begin{aligned} A_0^* h(z, y, r) &= r h(z, y, r) + E_0\{(1 - R) h(Z, Y, R) | Y = y\} \\ &= r h(z, y, r) + \frac{\int (1 - \pi(y)) h(t, y, 0) f(y|t; \theta_0) dG_0(t)}{f(y; G_0, \theta_0)}, \end{aligned} \tag{4.4}$$

and $B_0^* = \Pi_0^Z A_0^*$. It is easily verified that $B_0^* = \Pi_0^Z A_0 = \Pi_0^Z A_0 A_0 = B_0^* A_0$ on $L_2(P_0^{ZY})$ and, in particular, $B_0^* = B_0^* B_0$ on $L_2(P_0^Z)$. A closer look at these operators actually yields more.

Lemma 4.1. *Let (A2) and (A3) hold. Then*

(a) A_0 restricted to $L_2(P_0^{ZY})$ is one-to-one;

(b) $B_0^* B_0$ is coutinously invertible;

(c) $\mathcal{R}(B_0)$ is closed in $L_2(P_0)$.

Proof. Let $A_0 h = 0$ with $h \in L_2(P_0^{ZY})$. In random variable notation, this means $A_0 h(Z, Y, R) = 0$ almost surely. It follows that

$$0 = R A_0 h(Z, Y, R) = R h(Z, Y) \qquad \text{almost surely,}$$

hence

$$0 = E_0\{R h(Z, Y) | Z, Y\} = \pi_0(Y) h(Z, Y) \qquad \text{almost surely.}$$

By assumption, this further implies that $h(Z, Y) = 0$ almost surely, i.e., $h = 0$ in $L_2(P_0^{ZY})$. Thus (a) is established; in particular, B_0 is one-to-one. It follows that the self-adjoint operator $B_0^* B_0$ is positive-definite. Therefore its spectrum is contained in $(0, \infty)$, which does not include 0. This proves (b). In

particular, $\mathcal{R}(B_0^*) = L_2(G_0)$, which is closed. By Theorem 4.14 of Rudin (1973, p. 96), this implies that $\mathcal{R}(B_0)$ is closed in $L_2(P_0)$. □

Lemma 4.1 says that the projection of $\dot{\ell}_0$ into $\mathcal{R}(B_0)$ exists; by a standard result in functional analysis, it is given by $B_0(B_0^*B_0)^{-1}B_0^*\dot{\ell}_0$. This can also be written as $B_0(B_0^*B_0)^{-1}B_0^*A_0\dot{\ell}_0^{ZY} = B_0(B_0^*B_0)^{-1}B_0^*\dot{\ell}_0^{ZY}$ by a previous remark. It is verified that B_0, B_0^* and hence $B_0^*B_0$ and $(B_0^*B_0)^{-1}$ are all mean-preserving, i.e., $P_0B_0h = P_0h$. Therefore $B_0(B_0^*B_0)^{-1}B_0^*\dot{\ell}_0$ is in $L_2^0(P_0)$ and is also the projection of $\dot{\ell}_0$ into the right side of (4.3) (an intersection of two closed subspaces). Denote

$$\dot{\ell}_e = \dot{\ell}_0 - B_0(B_0^*B_0)^{-1}B_0^*\dot{\ell}_0.$$

It is not yet clear that $\dot{\ell}_e$ is the efficient score for θ. The right side of (4.3) is in general smaller than Γ, and $I_e := P_0(\dot{\ell}_e\dot{\ell}_e^T)$ is larger than the efficient Fisher information in the sense of nonnegative definiteness. However, RHN have demonstrated the existence of a regular estimator of θ with asymptotic variance I_e^{-1}. Then $\dot{\ell}_e$ must be the efficient score and I_e the efficient information.

This projection is taken as the starting point in the information calculation of RHN, which focuses on the characterization of $\dot{\ell}_e$ in terms of an integral equation. The discussion here clarifies some technical issues and yields a different representation of $\dot{\ell}_e$ as well as intermediate results which will be useful in later developments.

4.2 LIKELIHOOD EQUATIONS

Assume that

(D0) (C0)–(C6) hold.

The strong consistency proved in Chapter 3 and the definition of $(\hat{\theta}_n, \hat{G}_n)$ together imply that

$$0 = \frac{\partial}{\partial\theta}\frac{1}{n}\log L_n(\theta, \hat{G}_n)\bigg|_{\theta=\hat{\theta}_n} = \mathbb{P}_n\dot{\ell}_{\hat{\theta}_n,\hat{G}_n} \tag{4.5}$$

for large n, almost surely. (The left side is well defined because $\hat{G}_n$ is finitely discrete.) Let $\mathcal{H}$ be a uniformly bounded GC class of real-valued measurable functions on $\mathcal{Z}$. Specific choices of $\mathcal{H}$ will be made later on in this chapter. Recall from Chapter 2 that $(\hat{\theta}_n, \hat{G}_n)$ satisfies

$$\hat{G}_nh = \mathbb{P}_nB_{\hat{\theta}_n,\hat{G}_n}h, \qquad h \in \mathcal{H}, \tag{4.6}$$

where the operator notation is defined in the last subsection.

We are now ready to define a system of likelihood equations. Let $\ell^\infty(\mathcal{H})$ denote the collection of bounded real-valued functions on $\mathcal{H}$; this is a Banach space under the uniform norm:

$$\|T\|_\mathcal{H} = \sup_{h\in\mathcal{H}}|Th|, \qquad T \in \ell^\infty(\mathcal{H}).$$

The product space $R^q \times \ell^\infty(\mathcal{H})$ is a Banach space too, under the product norm:

$$\|(a,T)\| = |a| \vee \|T\|_{\mathcal{H}}, \qquad a \in \mathbb{R}^q,\ T \in \ell^\infty(\mathcal{H}),$$

where $|\cdot|$ denotes the Euclidean norm. Let the random maps $\Psi_n : \Theta \times \mathcal{G} \to \mathbb{R}^q \times \ell^\infty(\mathcal{H})$ be defined by $\Psi_n(\theta, G) = (\Psi_{n1}(\theta, G), \Psi_{n2}(\theta, G))$, where

$$\Psi_{n1}(\theta, G) = \mathbb{P}_n \dot{\ell}_{\theta,G} \qquad \text{and} \qquad \Psi_{n2}(\theta, G)h = \mathbb{P}_n B_{\theta,G} h - Gh.$$

The first equation above is a valid definition only if $\dot{\ell}_{\theta,G}$ exists almost surely under $\mathbb{P}_n$. When the latter fails, the definition of $\Psi_{n1}(\theta, G)$ is arbitrary. Note also that $B_{\theta,G}$ preserves boundedness, so that $\Psi_{n2}(\theta, G)$ is indeed in $\ell^\infty(\mathcal{H})$. As remarked in Chapter 3, each $G \in \mathcal{G}$ can be identified with an element of $\ell^\infty(\mathcal{H})$. Before embedding the domain of Ψ_n into $R^q \times \ell^\infty(\mathcal{H})$, however, there is a technical issue to address: we would like to know that $\Psi_n(\theta, G_1) = \Psi_n(\theta, G_2)$ whenever $\|G_1 - G_2\|_{\mathcal{H}} = 0$, so that Ψ_n is well defined on the new domain. This can be ascertained by requiring that the linear span of $\mathcal{H}$ contain a vector lattice of bounded continuous functions on $\mathcal{Z}$ which contains constant functions and separates points in $\mathcal{Z}$. In that case $\|G_1 - G_2\|_{\mathcal{H}} = 0$ implies that $G_1 = G_2$ (VW, lemma 1.3.12). Now the Ψ_n can be regarded as maps from $\mathbb{R}^q \times \ell^\infty(\mathcal{H})$ into itself whose domain is the product of Θ with the set of probability measures in $\ell^\infty(\mathcal{H})$ under the given identification. It follows from (4.5) and (4.6) that $\Psi_n(\hat{\theta}_n, \hat{G}_n) = 0$ for large n, almost surely. Let us assume that

(D4) For some $\tau > 0$, $\dot{\ell}_{\theta,G}$ is P_0-almost surely well defined whenever $\|(\theta - \theta_0, G - G_0)\| < \tau$ and

$$P_0^* \sup \left\{ \left|\dot{\ell}_{\theta,G}\right|^2 : \|(\theta - \theta_0, G - G_0)\| < \tau \right\} < \infty,$$

where the superscript * denotes outer integral.

This justifies the definition of the population version of Ψ_n by $\Psi = (\Psi_1, \Psi_2)$, where

$$\Psi_1(\theta, G) = P_0 \dot{\ell}_{\theta,G} \qquad \text{and} \qquad \Psi_2(\theta, G)h = P_0 B_{\theta,G} h - Gh. \tag{4.7}$$

(Again, the exact definition of Ψ outside a neighborhood of (θ_0, G_0) is irrelevant for our purpose.) It can be shown by simple algebra that $\Psi(\theta_0, G_0) = 0$.

Under the Z-theorem, Ψ is required to be Fréchet-differentiable at (θ_0, G_0), with derivative $\dot{\Psi}$ defined on the linear span of $\Theta \times \mathcal{G} - (\theta_0, G_0)$. Heuristically, this can be derived as follows. First, for $(\theta, G) \approx (\theta_0, G_0)$,

$$\begin{aligned}
\Psi_1(\theta, G) - \Psi_1(\theta_0, G_0) &= P_0(\dot{\ell}_{\theta,G} - \dot{\ell}_0) \\
&= P_0(\dot{\ell}_{\theta,G} - \dot{\ell}_{\theta_0,G}) + P_0(\dot{\ell}_{\theta_0,G} - \dot{\ell}_0) \\
&\approx P_0 \ddot{\ell}_0(Z, Y, R)(\theta - \theta_0) \\
&\quad + \iint (1 - \pi_0(y))(\dot{\ell}_0^{ZY}(z, y) - \dot{\ell}_0^{Y}(y)) f(y|z; \theta_0) d\mu(y) d(G - G_0)(z),
\end{aligned} \tag{4.8}$$

where

$$\ddot{\ell}_{\theta,G}(z,y,r) := \frac{\partial}{\partial\theta^T}\dot{\ell}_{\theta,G}(z,y,r) = r\ddot{\ell}_{\theta}^{ZY}(z,y) + (1-r)\ddot{\ell}_{\theta,G}^{Y}(y)$$
$$\ddot{\ell}_{\theta}^{ZY}(z,y) := \frac{\partial}{\partial\theta^T}\dot{\ell}_{\theta}^{ZY}(z,y) = \frac{\partial^2}{\partial\theta\partial\theta^T}\log f(y|z;\theta),$$
$$\ddot{\ell}_{\theta,G}^{Y}(y) := \frac{\partial}{\partial\theta^T}\dot{\ell}_{\theta,G}^{Y}(y) = \frac{\partial^2}{\partial\theta\partial\theta^T}\log f(y;G,\theta).$$

Under regularity conditions, the first term on the right side of (4.8) is equal to $-I_0(\theta-\theta_0)$, where $I_0 := P_0(\dot{\ell}_0\dot{\ell}_0^T)$ is the Fisher information for θ when G is known to be G_0. In operator notation, the second term can be rewritten as $-(G-G_0)B_0^*\dot{\ell}_0^{ZY}$. The derivative of the second component of Ψ can be obtained in a similar fashion. Uniformly over $h \in \mathcal{H}$,

$$\begin{aligned}\Psi_2(\theta,G)h - \Psi_2(\theta_0,G_0)h &= (P_0 - P_{\theta,G})B_{\theta,G}h \\ &= (P_0 - P_{\theta,G_0})B_{\theta,G}h + (P_{\theta,G_0} - P_{\theta,G})B_{\theta,G}h \\ &\approx -(P_0B_0h\dot{\ell}_0^T)(\theta-\theta_0) - (G-G_0)B_0^*B_0h.\end{aligned}$$

The foregoing discussion suggests that $\dot{\Psi}$ is given by the map

$$\begin{pmatrix}\theta-\theta_0\\ G-G_0\end{pmatrix} \mapsto \begin{pmatrix}\dot{\Psi}_{11} & \dot{\Psi}_{12}\\ \dot{\Psi}_{21} & \dot{\Psi}_{22}\end{pmatrix}\begin{pmatrix}\theta-\theta_0\\ G-G_0\end{pmatrix}, \tag{4.9}$$

where

$$\dot{\Psi}_{11}(\theta-\theta_0) = -I_0(\theta-\theta_0), \tag{4.10}$$
$$\dot{\Psi}_{12}(G-G_0) = -(G-G_0)B_0^*\dot{\ell}_0, \tag{4.11}$$
$$\dot{\Psi}_{21}(\theta-\theta_0)h = -(P_0B_0h\dot{\ell}_0^T)(\theta-\theta_0), \tag{4.12}$$
$$\dot{\Psi}_{22}(G-G_0)h = -(G-G_0)B_0^*B_0h. \tag{4.13}$$

This derivation can be made rigorous by imposing regularity conditions. An intermediate set of sufficient conditions for (4.9) is given below.

(D5) As $(\theta,G)\to(\theta_0,G_0)$,

$$P_0\ddot{\ell}_0^{ZY} + \mathrm{var}(\dot{\ell}_0^{ZY}) = 0; \tag{4.14}$$
$$P_0\left|\dot{\ell}_\theta^{ZY} - \dot{\ell}_0^{ZY} - \ddot{\ell}_0^{ZY}(\theta-\theta_0)\right| = o(|\theta-\theta_0|); \tag{4.15}$$
$$P_0\left|\dot{\ell}_{\theta,G}^{Y} - \dot{\ell}_{\theta_0,G}^{Y} - \ddot{\ell}_0^{Y}(\theta-\theta_0)\right| = o(|\theta-\theta_0|); \tag{4.16}$$
$$\iint(1-\pi_0(y))(\dot{\ell}_{\theta_0,G}^{Y} - \dot{\ell}_0^{Y})(y)f(y|z;\theta_0)d\mu(y)d(G-G_0)(z) = o(\|G-G_0\|_{\mathcal{H}}); \tag{4.17}$$
$$\iint\left|f(y|z;\theta) - f(y|z;\theta_0) - \dot{f}(y|z;\theta_0)(\theta-\theta_0)\right|d\mu(y)dG_0(z) = o(|\theta-\theta_0|); \tag{4.18}$$
$$\sup_{h\in\mathcal{H}} P_0(B_{\theta,G_0}h - B_0h)^2 = o(1); \tag{4.19}$$
$$\sup_{h\in\mathcal{H}}\left|(G-G_0)(B_{\theta,G}^*B_{\theta,G} - B_0^*B_0)h\right| = o(\|G-G_0\|_{\mathcal{H}}). \tag{4.20}$$

Specifically, (4.14) says that

$$P_0 \frac{\ddot{f}(Y|Z;\theta_0)}{f(Y|Z;\theta_0)} = 0. \tag{4.21}$$

Observe that

$$\ddot{\ell}_0(z,y,r) + \dot{\ell}_0(z,y,r)\dot{\ell}_0(z,y,r)^T = r\frac{\ddot{f}(y|z;\theta_0)}{f(y|z;\theta_0)} + (1-r)\frac{\int \ddot{f}(y|t;\theta_0)dG_0(t)}{\int f(y|t;\theta_0)dG_0(t)}.$$

This has mean 0, by (4.21) and a conditioning argument. Hence $P_0\ddot{\ell}_0 = -I_0$. This, (4.15), and (4.16) together justify the approximation of $P_0(\dot{\ell}_{\theta,G} - \dot{\ell}_{\theta_0,G})$ by (4.10). That (4.11) approximates $P_0(\dot{\ell}_{\theta_0,G} - \dot{\ell}_0)$ follows from (4.17) and the following identities:

$$P_0(\dot{\ell}_{\theta_0,G} - \dot{\ell}_0) = \iint (1-\pi_0(y))(\dot{\ell}^Y_{\theta_0,G} - \dot{\ell}^Y_0)(y) f(y|z;\theta_0)d\mu(y)dG_0(z), \tag{4.22}$$

$$0 = \iint (1-\pi_0(y))(\dot{\ell}^{ZY}_0(z,y) - \dot{\ell}^Y_{\theta_0,G}(y)) f(y|z;\theta_0)d\mu(y)dG(z), \tag{4.23}$$

$$0 = \iint (1-\pi_0(y))(\dot{\ell}^{ZY}_0(z,y) - \dot{\ell}^Y_0(y)) f(y|z;\theta_0)d\mu(y)dG_0(z). \tag{4.24}$$

Subtract (4.24) from the sum of (4.22), (4.17) (reversed) and (4.23) to obtain

$$P_0(\dot{\ell}_{\theta_0,G} - \dot{\ell}_0) = (4.11) + o(\|G - G_0\|_{\mathcal{H}}).$$

In order to explain (4.12) we write

$$\begin{aligned}(P_{\theta,G_0} - P_0)B_{\theta,G}h &= (P_0 B_{\theta,G}h\dot{\ell}^{ZY}_0)^T(\theta-\theta_0) + o(|\theta-\theta_0|)\\ &= (P_0 B_0 h\dot{\ell}^{ZY}_0)^T(\theta-\theta_0) + o(|\theta-\theta_0|)\\ &= (P_0 B_0 h\dot{\ell}^{T}_0)(\theta-\theta_0) + o(|\theta-\theta_0|),\end{aligned}$$

where the first step follows from (4.18), the second from (D3), (4.19) and the Cauchy-Schwartz inequality, and the last from a remark in Section 4.1. All remainders in the above display are negligible uniformly in h; thus (4.12) is a valid approximation of $(P_0 - P_{\theta,G_0})B_{\theta,G}h$. Lastly, note that

$$\begin{aligned}(P_{\theta,G} - P_{\theta,G_0})B_{\theta,G}h &= Gh - G_0\Pi^Z_\theta B_{\theta,G}h\\ &= (G-G_0)B^*_{\theta,G}B_{\theta,G}h\\ &= (G-G_0)B^*_0 B_0 h + o(\|G-G_0\|_{\mathcal{H}}),\end{aligned}$$

where the last step follows from (4.20) and is uniform in h.

Condition (D5) can in turn be deduced from simpler conditions on the regression model (usually smoothness and integrability properties). The arguments involved typically depend on the choice of $\mathcal{H}$. Some common techniques are illustrated by the examples in Chapter 6.

4.3 MAIN RESULTS

We are finally ready to formulate a set of sufficient conditions for the joint asymptotic normality of $(\hat{\theta}_n, \hat{G}_n)$. The following result is essentially a translation of Theorem 3.3.1 and Lemma 3.3.5 of VW into the present

context. Let us assume that

(D6) $\mathcal{H}$ and $\left\{\dot{\ell}_{\theta,G}, \Pi^Y_{\theta,G} h : \|(\theta-\theta_0, G-G_0)\| < \tau, h \in \mathcal{H}\right\}$ are Donsker;

(D7) $\left|\dot{\ell}_{\theta,G} - \dot{\ell}_0\right| \vee \left\|\Pi^Y_{\theta,G} - \Pi^Y_0\right\|_{\mathcal{H}} \to 0$ almost surely as $\|(\theta-\theta_0, G-G_0)\| \to 0$;

(D8) $\dot{\Psi}$ given by (4.9) is continuously invertible.

Write $\mathbb{G}_n = \sqrt{n}(\mathbb{P}_n - P_0)$.

Theorem 4.2. *Let (D0)–(D8) hold for a suitable class $\mathcal{H}$ of uniformly bounded functions on $\mathcal{Z}$. Then*

$$\sqrt{n}\dot{\Psi}\begin{pmatrix}\hat{\theta}_n - \theta_0 \\ \hat{G}_n - G_0\end{pmatrix} = -\mathbb{G}_n \begin{pmatrix}\dot{\ell}_0 \\ B_0 h - G_0 h : h \in \mathcal{H}\end{pmatrix} + o^*_p(1). \tag{4.25}$$

In particular, $\sqrt{n}(\hat{\theta}_n - \theta_0, \hat{G}_n - G_0)$ converges weakly to $\dot{\Psi}^{-1}Q$ in $\mathbb{R}^q \times \ell^\infty(\mathcal{H})$, where $Q = (Q_1, Q_2)$ is a tight Gaussian process in $\mathbb{R}^q \times \ell^\infty(\mathcal{H})$ with mean 0 and covariances given by

$$\begin{aligned}\operatorname{var} Q_1 &= I_0, \\ \operatorname{cov}(Q_1, Q_2 h) &= P_0(\dot{\ell}_0 B_0 h), \\ \operatorname{cov}(Q_2 h_1, Q_2 h_2) &= G_0\{B_0 h_1 (B_0 h_2 - G_0 h_2)\}.\end{aligned}$$

Conditions (D1)–(D3) are standard assumptions in the asymptotic theory of maximum likelihood estimation without missing data (see van der Vaart, 1998, chap. 5). The other conditions may appear more complicated. In particular, it seems difficult to give a general characterization of the Donsker property (D6) in terms of simple and easy-to-verify conditions. In fact, different regression models may require different techniques. Chapter 6 contains examples that illustrate the techniques for verifying these conditions. Here we focus on the characterization of (D8), which will lead to an explicit formula for $\dot{\Psi}^{-1}$ and a better understanding of the limit distribution of $\sqrt{n}(\hat{\theta}_n - \theta_0, \hat{G}_n - G_0)$. The arguments of VW (example 3.3.10) and van der Vaart (1994a) play a central role in the following discussion.

It is apparent from the block form (4.9) of $\dot{\Psi}$ that (D8) would follow from the continuous invertibility of both $\dot{\Psi}_{11}$ and $\dot{V} := \dot{\Psi}_{22} - \dot{\Psi}_{21}\dot{\Psi}_{11}^{-1}\dot{\Psi}_{12}$. Recall that $I_0 = P_0\{A_0 \dot{\ell}_0^{ZY}(A_0\dot{\ell}_0^{ZY})^T\}$. Lemma 4.1 and assumption (D3) together then imply that I_0 is positive definite, hence invertible. The second operator has the form

$$\dot{V}(G-G_0)h = (G-G_0)\{(P_0 B_0 h \dot{\ell}_0^T) I_0^{-1} B_0^* \dot{\ell}_0 - B_0^* B_0 h\}, \qquad h \in \mathcal{H}, G \in \mathcal{G}.$$

$\dot{V}$ is continuously invertible if and only if there exists $c > 0$ such that

$$\|\dot{V}(G_1 - G_2)\|_{\mathcal{H}} \geq c\|G_1 - G_2\|_{\mathcal{H}}, \qquad G_1, G_2 \in \mathcal{G}.$$

The latter certainly would follow from the existence of $c > 0$ such that

$$\{(P_0 B_0 h \dot{\ell}_0^T) I_0^{-1} B_0^* \dot{\ell}_0 - B_0^* B_0 h : h \in \mathcal{H}\} \supset c\mathcal{H}. \tag{4.26}$$

Now it seems natural to take $\mathcal{H}$ to be the unit ball of a Banach space $(\mathbb{B}, \|\cdot\|)$ contained in $\ell^\infty(\mathcal{Z})$ with $\|\cdot\| \geq \|\cdot\|_\infty$. Then (4.26) will hold if the operator C defined by

$$Ch = (P_0 B_0 h \dot{\ell}_0^T) I_0^{-1} B_0^* \dot{\ell}_0 - B_0^* B_0 h$$

maps $\mathbb{B}$ into itself and is continuously invertible. Note, however, that objects like $B_0^*\dot{\ell}_0$ or $B_0^*B_0h$ are originally defined as vectors in $L_2(G_0)$ and therefore represent equivalence classes of functions. Instead of redefining these operators, we shall simply take the "natural" versions given by the far right sides of (4.1), (4.2) and (4.4). Thus it is understood that

$$B_0^*\dot{\ell}_0(z) = B_0^*\dot{\ell}_0^{ZY}(z) = \int_{\mathcal{Y}} \left\{ \pi_0(y)\dot{\ell}_0^{ZY}(z,y) + (1-\pi_0(y)) \frac{\int_{\mathcal{Z}} \dot{\ell}_0^{ZY}(t,y) f(y|t;\theta_0) dG_0(t)}{f(y;G_0,\theta_0)} \right\} f(y|z;\theta_0) d\mu(y).$$

Similarly, for $h \in \ell^\infty(\mathcal{Z})$,

$$B_0^*B_0h(z) = B_0^*h(z) = \int_{\mathcal{Y}} \left\{ \pi_0(y)h(z) + (1-\pi_0(y)) \frac{\int_{\mathcal{Z}} h(t) f(y|t;\theta_0) dG_0(t)}{f(y;G_0,\theta_0)} \right\} f(y|z;\theta_0) d\mu(y). \tag{4.27}$$

This slight abuse of notation is nevertheless very convenient. We will indicate which spaces are being considered whenever the context does not implicitly do so.

Suppose now that

(D8′a) All components of $B_0^*\dot{\ell}_0$ are in $\mathbb{B}$;

(D8′b) $B_0^*B_0 : \mathbb{B} \to \mathbb{B}$ is continuously invertible.

Then $C : \mathbb{B} \to \mathbb{B}$ is linear and continuous. (This makes use of the assumption that $\mathbb{B}$ has a norm stronger than the uniform norm.) Furthermore, the first component of C has finite rank and therefore is compact. It then follows from a standard result in functional analysis that C is Fredolm. In particular, it is continuously invertible if and only if it is one-to-one. The latter indeed holds under a stronger version of (D3):

(D3′) No nontrivial linear combination of the components of $\dot{\ell}_0^{ZY}(z,y)$ depends on (z,y) only through z;

in other words, the components of $\dot{\ell}_0^{ZY}$, considered as vectors in the quotient space $L_2(P_0^{ZY})\,/L_2(G_0)$, are linearly independent. Condition (D3′) is intimately related to the positivity of the efficient information I_e.

Lemma 4.3. *(a) If I_e is positive definite, then (D3′) holds. (b) Conversely, (A2), (A3) and (D3′) together imply the positive definiteness of I_e. (c) Under (A2), (A3), (D3′) and (D8′b), $C : \mathbb{B} \to \mathbb{B}$ is one-to-one.*

Proof. (a) Let I_e be positive definite and let $a \in \mathbb{R}^q$ be such that $a^T\dot{\ell}_0^{ZY} \in L_2(G_0)$. Then $a^T\dot{\ell}_0 = A_0(a^T\dot{\ell}_0^{ZY}) = B_0(a^T\dot{\ell}_0^{ZY})$, whence $a^T\dot{\ell}_e = 0$ almost surely, whence $a^TI_ea = 0$. The positivity of I_e now implies $a = 0$.

(b) Assume (A2), (A3) and (D3′) and let $a^TI_ea = 0$. Then, almost surely, $0 = a^T\dot{\ell}_e = A_0\{a^T\dot{\ell}_0^{ZY} - (B_0^*B_0)^{-1}B_0^*(a^T\dot{\ell}_0^{ZY})\}$. By Lemma 4.1, A_0 is one-to-one on $L_2(P_0^{ZY})$, so that $a^T\dot{\ell}_0^{ZY} \in L_2(G_0)$. By (D3′), this implies $a = 0$.

(c) Let $h \in \mathbb{B}$ be such that $Ch = 0$ in $\mathbb{B}$ (i.e., pointwise). Simple algebraic manipulation then yields

$$(P_0B_0h\dot{\ell}_0^T)\left[I_0^{-1}P_0\left\{B_0(B_0^*B_0)^{-1}B_0^*\dot{\ell}_0\dot{\ell}_0^T\right\} - I\right] = 0,$$

where I is the identity matrix. This can be written in a simpler form:

$$I_eI_0^{-1}P_0(B_0h\dot{\ell}_0) = 0.$$

It follows from the positivity (invertibility) of I_e that $P_0(B_0 h\dot{\ell}_0) = 0$. Substituting this into the definition of C gives that $B_0^* B_0 h = 0$ in $\mathbb{B}$. The result now follows from assumption (D8′b). □

It follows from this discussion that, under the newly introduced assumptions, $\dot{V}$ and hence $\dot{\Psi}$ are continuously invertible, with

$$\dot{V}^{-1} T h = T C^{-1} h, \qquad h \in \mathcal{H},\ T \in \mathcal{R}(\dot{V}),$$
$$\dot{\Psi}^{-1} = \begin{pmatrix} \dot{\Psi}_{11}^{-1}(\dot{\Psi}_{11} + \dot{\Psi}_{12}\dot{V}^{-1}\dot{\Psi}_{21})\dot{\Psi}_{11}^{-1} & -\dot{\Psi}_{11}^{-1}\dot{\Psi}_{12}\dot{V}^{-1} \\ -\dot{V}^{-1}\dot{\Psi}_{21}\dot{\Psi}_{11}^{-1} & \dot{V}^{-1} \end{pmatrix}. \tag{4.28}$$

This allows us to derive the asymptotic distribution of $\sqrt{n}(\hat{\theta}_n - \theta_0, \hat{G}_n - G_0)$ explicitly.

Corollary 4.4. *Let (D0)–(D2), (D3′), (D4)–(D7) and (D8′) be satisfied for a suitable Banach space $\mathbb{B}$ with unit ball $\mathcal{H}$. Then*

$$\sqrt{n}\begin{pmatrix} \hat{\theta}_n - \theta_0 \\ \hat{G}_n - G_0 \end{pmatrix} = \mathbb{G}_n \begin{pmatrix} I_e^{-1}\dot{\ell}_e \\ B_0 C^{-1} h - (P_0 B_0 C^{-1} h \dot{\ell}_0^T) I_0^{-1} \dot{\ell}_0 : h \in \mathcal{H} \end{pmatrix} + o_p^*(1). \tag{4.29}$$

In particular, $\sqrt{n}(\hat{\theta}_n - \theta_0)$ is asymptotically normal with mean 0 and variance I_e^{-1}.

Proof. It follows from (4.25) and the continuous mapping theorem that

$$\sqrt{n}\begin{pmatrix} \hat{\theta}_n - \theta_0 \\ \hat{G}_n - G_0 \end{pmatrix} = -\dot{\Psi}^{-1}\mathbb{G}_n \begin{pmatrix} \dot{\ell}_0 \\ B_0 h - G_0 h : h \in \mathcal{H} \end{pmatrix} + o_p^*(1).$$

A term-by-term examination shows that $\dot{\Psi}^{-1}$ and $\mathbb{G}_n$ can be interchanged in the above display. A direct application of (4.28) then gives (4.29), a key observation being that

$$C^{-1}(B_0^* \dot{\ell}_0) = I_0 I_e^{-1} (B_0^* B_0)^{-1} B_0^* \dot{\ell}_0,$$

which is easily verified. □

A convenient choice for $\mathbb{B}$ is a class of smooth functions on $\mathcal{Z}$. To be precise, let $\alpha > 0$ and let $\underline{\alpha}$ denote the largest integer strictly smaller than α. Let $C^\alpha(\mathcal{Z})$ be the class of continuous real-valued functions on $\mathcal{Z}$ that are differentiable up to order $\underline{\alpha}$ with bounded partial derivatives and whose $\underline{\alpha}$-th order partial derivatives are Lipschitz of order $\alpha - \underline{\alpha}$. A norm on $C^\alpha(\mathcal{Z})$ can be defined as follows. For a vector $k = (k_1, \ldots, k_d)$ of nonnegative integers, define the differential operator

$$D^k = \frac{\partial^{k_\cdot}}{\partial z_1^{k_1} \ldots \partial z_d^{k_d}},$$

where $k_\cdot := \sum_{j=1}^d k_j$. Then for $h \in C^\alpha(\mathcal{Z})$, let

$$\|h\|_\alpha = \max_{k_\cdot \le \underline{\alpha}} \sup_z \left|D^k h(z)\right| \vee \max_{k_\cdot = \underline{\alpha}} \sup_{z,z'} \frac{\left|D^k h(z) - D^k h(z')\right|}{|z - z'|^{\alpha - \underline{\alpha}}},$$

where the suprema are taken over all z, z' in the interior of $\mathcal{Z}$ with $z \neq z'$. For $t > 0$, let $C_t^\alpha(\mathcal{Z}) = \{h \in C^\alpha(\mathcal{Z}) : \|h\|_\alpha \leq t\}$. Thus if $\mathbb{B} = C^\alpha(\mathcal{Z})$, then $\mathcal{H} = C_1^\alpha(\mathcal{Z})$. This is consistent with our previous definition of $C_1^1(\mathcal{Z})$.

As will be discussed later, the choice of α depends on d, the dimension of $\mathcal{Z}$, among other things. For example, it will be required that $\alpha > d/2$. For $\alpha \in (0, 1]$, (D8′a) holds if there exist measurable functions M and M' such that

$$\sup_{z \neq z'} \frac{|f(y|z;\theta_0) - f(y|z';\theta_0)|}{|z - z'|^\alpha} \leq M(y), \tag{4.30}$$

$$\int (1 - \pi_0(y)) \left|\dot{\ell}_0^Y(y)\right| M(y) d\mu(y) < \infty, \tag{4.31}$$

$$\sup_{z \neq z'} \frac{\left|\dot{f}(y|z;\theta_0) - \dot{f}(y|z';\theta_0)\right|}{|z - z'|^\alpha} \leq M'(y), \tag{4.32}$$

$$\int \pi_0(y) M'(y) d\mu(y) < \infty. \tag{4.33}$$

If $\alpha = 1$, $\mathcal{Z}$ is convex and $f(y|z;\theta_0)$ and $\dot{f}(y|z;\theta_0)$ are continuously differentiable with respect to z, then it seems natural to take

$$M(y) = \sup_z \left|\frac{\partial}{\partial z} f(y|z;\theta_0)\right|, \tag{4.34}$$

$$M'(y) = \sup_z \left|\frac{\partial}{\partial z^T} \dot{f}(y|z;\theta_0)\right|. \tag{4.35}$$

The theory of Fredolm operators again proves useful for verifying (D8′b).

Lemma 4.5. *Let (A3), (B1) and (B4) be satisfied.*

(a) Suppose that (4.30) *holds for some $\alpha \in (0, 1]$ and μ-integrable M. Then (D8′b) holds for $\mathbb{B} = C^\beta(\mathcal{Z})$ for every $\beta < \alpha$.*

(b) Suppose that (4.30) *holds for $\alpha = 1$ and some μ-integrable M. Furthermore, assume that $\mathcal{Z}$ is convex and that for every y, $f(y|z;\theta_0)$ is differentiable with respect to z, with*

$$\int \left|\frac{\partial}{\partial z} f(y|z;\theta_0) - \frac{\partial}{\partial z} f(y|z';\theta_0)\right| d\mu(y) \leq K|z - z'|^\beta, \tag{4.36}$$

$$\int \left|\frac{\partial}{\partial z} f(y|z;\theta_0)\right| d\mu(y) \leq K, \tag{4.37}$$

for all z,z' in $\mathcal{Z}$ and fixed constants K and $\beta > 0$. Then (D8′b) holds for $\mathbb{B} = C^1(\mathcal{Z})$.

Proof. Consider (a) first. In view of (4.27), the continuous invertibility of $B_0^* B_0 : C^\beta(\mathcal{Z}) \to C^\beta(\mathcal{Z})$ can be deduced from the following claims:

(i) The map $h \in C^\beta(\mathcal{Z}) \mapsto (\Pi_0^Z \pi_0) h =: \pi_0^Z h$ has range contained in $C^\beta(\mathcal{Z})$ and is continuously invertible;

(ii) The map $h \in C^\beta(\mathcal{Z}) \mapsto \Pi_0^Z\{(1 - \pi_0)\Pi_0^Y h\}$ has range contained in $C^\beta(\mathcal{Z})$ and is compact;

(iii) $B_0^* B_0 : C^\beta(\mathcal{Z}) \to C^\beta(\mathcal{Z})$ is one-to-one.

It follows from (4.30) and the μ-integrability of M that π_0^Z is Lipschitz of order α, hence of order β. By compactness of $\mathcal{Z}$ (B1), π_0^Z is bounded, hence in $C^\beta(\mathcal{Z})$. For every $h \in C^\beta(\mathcal{Z})$, we have

$$\begin{aligned}\left|\pi_0^Z(z_1)h(z_1) - \pi_0^Z(z_2)h(z_2)\right| &\le \left\|\pi_0^Z\right\|_\infty \left|h(z_1) - h(z_2)\right| + \left\|h\right\|_\infty \left|\pi_0^Z(z_1) - \pi_0^Z(z_2)\right| \\ &\le 2\left\|\pi_0^Z\right\|_\beta \left\|h\right\|_\beta \left|z_1 - z_2\right|^\beta, \qquad z_1, z_2 \in \mathcal{Z},\end{aligned}$$

so that $\pi_0^Z h \in C^\beta(\mathcal{Z})$. It is easy to see that $h \mapsto \pi_0^Z h$ is linear, continuous and one-to-one. (A3) says that π_0 is positive P_0^Y-almost everywhere on $\mathcal{Y}$; by (B4), this actually holds μ-almost everywhere. It follows that π_0^Z is positive everywhere on $\mathcal{Z}$. Combined with the compactness of $\mathcal{Z}$ and the continuity of π_0^Z, this implies that π_0^Z is bounded away from 0. It can then be shown that $h \mapsto \pi_0^Z h$ is onto $C^\beta(\mathcal{Z})$. Now (i) follows from the inverse mapping theorem. Next, $h \in C^\beta(\mathcal{Z}) \mapsto (1 - \pi_0)\Pi_0^Y h \in \ell^\infty(\mathcal{Y})$ is linear and continuous because $\|\cdot\|_\infty \le \|\cdot\|_\beta$. By Lemma 3.1 of van der Vaart (1994a), Π_0^Z maps $\ell^\infty(\mathcal{Y})$ into $C^\beta(\mathcal{Z})$ and is compact. Thus (ii) follows. To see (iii), let $h \in C^\beta(\mathcal{Z})$ be such that $B_0^* B_0 h = 0$ in $C^\beta(\mathcal{Z})$ (i.e., pointwise). By Lemma 4.1, $h = 0$ G_0-almost everywhere. It follows that $\Pi_0^Y h = 0$ P_0^Y-almost everywhere and, by (B4), μ-almost everywhere. This shows that the second component of $B_0^* B_0 h$ is 0 pointwise, whence $\pi_0^Z h = 0$ pointwise. By the strict positivity of π_0^Z, $h = 0$ pointwise (i.e., in $C^\beta(\mathcal{Z})$).

(b) can be proved in a similar manner, except that the analogue of (ii) will be deduced from Lemma 5.1 of van der Vaart (1994a) instead. □

These arguments can in principle be extended to larger α (> 1), by imposing smoothness conditions on derivatives of higher orders (depending on α). However, it appears difficult to treat all α simultaneously with one set of conditions in one proof.

When $\mathcal{Z} = [l, u]$ is a compact interval in the real line, another choice for $\mathbb{B}$ is the space of functions of bounded variation. The total variation of a real-valued function h on $\mathcal{Z}$ is given by

$$\|h\|_{BV} = \sup \sum_{j=1}^{k} \left|h(t_j) - h(t_{j-1})\right|,$$

where the supremum is taken over all finite partitions $l = t_0 < t_1 < \cdots < t_k = u$ of $\mathcal{Z}$. Let $BBV(\mathcal{Z})$ consist of functions h with

$$\|h\|_{BBV} := \|h\|_\infty \vee \|h\|_{BV} < \infty$$

and let $BBV_1(\mathcal{Z})$ denote its unit ball. It is clear that

$$\|\cdot\|_{BBV} \le \max\{u - l, 1\} \|\cdot\|_1, \tag{4.38}$$

where $\|\cdot\|_1$ refers to the $C^1(\mathcal{Z})$ norm. It follows that (D8′a) is satisfied for $\mathbb{B} = BBV(\mathcal{Z})$ if (4.30)–(4.33) hold for $\alpha = 1$, with potential candidates for M and M' given by (4.34) and (4.35) respectively. (D8′b) can be verified the same way as in Lemma 4.5.

Lemma 4.6. *In the situation of Lemma 4.5(b) with $d = 1$, (D8′b) is satisfied for $\mathbb{B} = BBV(\mathcal{Z})$.*

Proof. We shall restrict attention to particularities of the BBV space, without duplicating the details already explained in the proof of Lemma 4.5. π_0^Z is in $C^1(\mathcal{Z})$ and hence in $BBV(\mathcal{Z})$. For $h \in BBV(\mathcal{Z})$, we have

$$\left|\pi_0^Z(z_1)h(z_1) - \pi_0^Z(z_2)h(z_2)\right| \leq \left\|\pi_0^Z\right\|_\infty \left|h(z_1) - h(z_2)\right| + \left\|h\right\|_\infty \left|\pi_0^Z(z_1) - \pi_0^Z(z_2)\right|, \qquad z_1, z_2 \in \mathcal{Z},$$

hence

$$\left\|\pi_0^Z h\right\|_{BV} \leq \left\|\pi_0^Z\right\|_\infty \left\|h\right\|_{BV} + \left\|h\right\|_\infty \left\|\pi_0^Z\right\|_{BV} \leq 2\left\|\pi_0^Z\right\|_{BBV} \left\|h\right\|_{BBV}.$$

It follows that the map $h \in BBV(\mathcal{Z}) \mapsto \pi_0^Z h$ has range contained in $BBV(\mathcal{Z})$ and is linear and continuous. Furthermore, it is one-to-one and onto because π_0^Z is bounded away from 0. By the inverse mapping theorem, this map is continuously invertible. Lemma 5.1 of van der Vaart (1994a) implies that the map $h \in BBV(\mathcal{Z}) \subset \ell^\infty(\mathcal{Z}) \mapsto \Pi_0^Z\{(1-\pi_0)\Pi_0^Y h\}$ has range contained in $C^1(\mathcal{Z})$ and is compact for the 1-norm. By (4.38), it maps into $BBV(\mathcal{Z})$ and is compact for the BBV norm as well. Lastly, $B_0^* B_0$ is one-to-one. □

5.0 ASYMPTOTIC NORMALITY VIA QUADRATIC EXPANSION

In this chapter we explore a different approach to proving the asymptotic normality of $\hat{\theta}_n$ (not $\hat{G}_n$). The main technical tool here is a quadratic expansion of the profile log-likelihood for θ near θ_0, established by Murphy and van der Vaart (2000) (MV) for a general semiparametric model. Aside from the asymptotic normality of $\hat{\theta}_n$, the results of this section yield a consistent estimate of the asymptotic variance of $\hat{\theta}_n$ and justify the asymptotic chi-squared distribution of the profile likelihood ratio statistic, which can be used to test hypotheses about and construct confidence sets for θ.

The key structural requirement of the theory of MV is a well-behaved least favorable submodel which helps reduce the dimension of the problem. In what follows we propose a candidate submodel and show that it satisfies the conditions imposed by MV. Interestingly, the key arguments involved here are similar to those used in Chapter 4. So let us start by assuming that

(E0) Conditions (D0), (D1), (D2′), (D3′), (D4′), (D5), (D6), (D7′) and (D8′) hold for a suitable Banach space $\mathbb{B}$ with unit ball $\mathcal{H}$,

where (D2′), (D4′) and (D7′) are variants of their unprimed versions:

(D2′) (D2) holds with $\dot{f}(y|\cdot;\cdot)$ and $\ddot{f}(y|\cdot;\cdot)$ continuous for every y;

(D4′) $P_0 \sup\left\{\left|\dot{\ell}_\theta^{ZY}(z,Y)\right|^2 + \left|\ddot{\ell}_\theta^{ZY}(z,Y)\right| : z \in \mathcal{Z},\ |\theta-\theta_0|<\tau\right\} < \infty$ for some $\tau > 0$;

(D7′) For P_0-almost every y, $\sup\left\{\left|(\Pi_{\theta,G}^Y - \Pi_0^Y)h(y)\right| : h = \dot{\ell}_0^{ZY},\ \left(\dot{\ell}_0^{ZY}\right)^{\otimes 2},\ \dot{\ell}_0^{ZY}h_0^T,\ \ddot{\ell}_0^{ZY} \text{ or } \in \mathcal{H}\right\} \to 0$ as $\|(\theta-\theta_0, G-G_0)\| \to 0$, where $a^{\otimes 2} := aa^T$ and h_0 will be defined immediately.

For $G \in \mathcal{G}$ and θ, t in a neighborhood of θ_0, define $G_t(\theta, G)$ by

$$dG_t(\theta,G)/dG = 1 - (t-\theta)^T h_G,$$

where $h_G := h_0 - Gh_0$, and $h_0 := (B_0^* B_0)^{-1} B_0^* \dot{\ell}_0$ is the least favorable direction for θ at (θ_0, G_0). D8′ says that all components of h_0 are in $\mathbb{B}$, hence bounded. It follows that $G_t(\theta, G) \in \mathcal{G}$ for $|t-\theta|$ sufficiently small. A parametric submodel can then be defined by $t \mapsto (t, G_t(\theta, G))$ (for θ, t in a small neighborhood of θ_0). This submodel clearly passes through (θ, G) at $t = \theta$:

$$G_\theta(\theta, G) = G, \qquad \text{every } (\theta, G).$$

Thus condition (8) of MV is satisfied. Under this submodel, the log-density of (R, RZ, Y) with respect to some dominating measure is given by

$$\begin{aligned}\ell_{t,\theta,G}(z,y,r) &= \log\left[\left\{f(y|z;t)\frac{dG_t(\theta,G)}{dG}(z)\right\}^r\left\{\int f(y;u,t)dG_t(\theta,G)(u)\right\}^{1-r}\right] \\ &= r\log f(y|z;t) + r\log\{1-(t-\theta)^T h_G(z)\} \\ &\quad + (1-r)\log\int f(y|u;t)\{1-(t-\theta)^T h_G(u)\}dG(u).\end{aligned} \tag{5.1}$$

This does not correspond exactly to the semiparametric likelihood (expression (2.1)) we use, as no point mass appears in the above display. Adding the term $r\log G\{z\}$ to the right side would make an exact correspondence with (2.1). However, the resulting function would be difficult, if not impossible, to work with, precisely because of the point mass. A close look at the proof of MV's Theorem 1 reveals that, in connection with the profile likelihood, one may take (in their notation) $l(t,\theta,\eta) = \log \mathrm{l}(t,\eta_t(\theta,\eta)) + j(\theta,\eta)$ for any function j indexed by (θ,η) only. In particular, $\ell_{t,\theta,G}$ defined above is a legitimate choice, provided it satisfies the regularity conditions given in their theorem.

Differentiating (5.1) with respect to t gives

$$\begin{aligned}\dot\ell_{t,\theta,G}(z,y,r) &= r\dot\ell_t^{ZY}(z,y) - rh_{t,\theta,G}(z) + (1-r)\Pi^Y_{t,G_t(\theta,G)}\dot\ell_t^{ZY}(y) - (1-r)\Pi^Y_{t,G_t(\theta,G)}h_{t,\theta,G}(y) \\ &= A_{t,G_t(\theta,G)}(\dot\ell_t^{ZY} - h_{t,\theta,G})(z,y,r) \\ &= \dot\ell_{t,G_t(\theta,G)}(z,y,r) - B_{t,G_t(\theta,G)}h_{t,\theta,G}(z,y,r)\end{aligned} \tag{5.2}$$

and

$$\begin{aligned}\ddot\ell_{t,\theta,G}(z,y,r) &= r\ddot\ell_t^{ZY}(z,y) - rh^{\otimes 2}_{t,\theta,G}(z) + (1-r)\Pi^Y_{t,G_t(\theta,G)}\{(\ddot f/f)(\cdot|\cdot;t)\}(y) \\ &\quad - (1-r)\Pi^Y_{t,G_t(\theta,G)}(\dot\ell_t^{ZY}h_{t,\theta,G}^T)(y) - (1-r)(\Pi^Y_{t,G_t(\theta,G)}\dot\ell_t^{ZY})^{\otimes 2}(y) \\ &\quad + (1-r)(\Pi^Y_{t,G_t(\theta,G)}\dot\ell_t^{ZY})(\Pi^Y_{t,G_t(\theta,G)}h_{t,\theta,G})^T(y) - (1-r)\Pi^Y_{t,G_t(\theta,G)}(h_{t,\theta,G}\dot\ell_t^{ZY\,T})(y) \\ &\quad + (1-r)(\Pi^Y_{t,G_t(\theta,G)}h_{t,\theta,G})(\Pi^Y_{t,G_t(\theta,G)}\dot\ell_t^{ZY})^T(y) - (1-r)(\Pi^Y_{t,G_t(\theta,G)}h_{t,\theta,G})^{\otimes 2}(y) \\ &= A_{t,G_t(\theta,G)}\ddot\ell_t^{ZY}(z,y,r) - B_{t,G_t(\theta,G)}(h^{\otimes 2}_{t,\theta,G})(z,y,r) \\ &\quad + (1-r)\,\mathrm{var}\{\dot\ell_t^{ZY}(Z,Y) - h_{t,\theta,G}(Z)|Y=y;t,G_t(\theta,G)\},\end{aligned} \tag{5.3}$$

where $h_{t,\theta,G} := h_G/\{1-(t-\theta)^T h_G\}$. It follows that

$$\dot\ell_{\theta_0,\theta_0,G_0} = \dot\ell_e, \tag{5.4}$$

so that the submodel $t \mapsto (t, G_t(\theta_0,G_0))$ is least favorable for estimating θ at (θ_0,G_0) and condition (9) of MV is met.

The profile likelihood for θ is given by

$$PL_n(\theta) = \sup\big\{L_n(\theta,G) : G\in\mathcal{G}, G\{Z_i : R_i = 1\} = 1\big\}.$$

Under our assumptions, the supremum above is in fact a maximum. Denote by $\hat{G}_n(\theta)$ any maximizer in the above display, so that $PL_n(\theta) = L_n(\theta, \hat{G}_n(\theta))$. Then

$$\hat{\theta}_n = \arg\max_{\theta \in \Theta} PL_n(\theta),$$
$$\hat{G}_n = \hat{G}_n(\hat{\theta}_n).$$

Condition (10) of MV requires that $\|\hat{G}_n(\bar{\theta}_n) - G_0\|_{\mathcal{H}} \to 0$ in probability whenever $\bar{\theta}_n \to \theta_0$ in probability. By arguing along subsequences, this will follow from the same statement with "in probability" replaced by "almost surely". So let

$$\bar{\theta}_n \to \theta_0 \qquad \text{almost surely.} \tag{5.5}$$

That $\|\hat{G}_n(\bar{\theta}_n) - G_0\|_{\mathcal{H}} \to 0$ almost surely follows from the proof of Theorem 3.1 with the following modifications. First, choose an ω that satisfies (5.5) in addition to the conditions specified in that proof. Second, replace the middle term in (3.6) by

$$\frac{1}{n} \log \frac{L_n(\bar{\theta}_n, G_n)}{L_n(\bar{\theta}_n, \hat{G}_n(\bar{\theta}_n))}$$

and adjust other expressions accordingly. The same argument is then applicable, because (2.6) remains valid with $(\hat{\theta}, \hat{G})$ replaced by $(\bar{\theta}_n, \hat{G}_n(\bar{\theta}_n))$.

Now let us consider MV's (11), a nontrivial condition. Let $\bar{\theta}_n \to \theta_0$ in probability. We need to show that

$$P_0 \dot{\ell}_{\theta_0, \bar{\theta}_n, \hat{G}_n(\bar{\theta}_n)} = o_p(|\bar{\theta}_n - \theta_0| + n^{-1/2}).$$

In light of the discussion on page 457 of MV, this is equivalent to

$$P_0 \dot{\ell}_{\theta_0, \theta_0, \hat{G}_n(\bar{\theta}_n)} = o_p(|\bar{\theta}_n - \theta_0| + n^{-1/2}). \tag{5.6}$$

Write

$$\begin{aligned} P_0 \dot{\ell}_{\theta_0, \theta_0, \hat{G}_n(\bar{\theta}_n)} &= P_0(\dot{\ell}_{\theta_0, \hat{G}_n(\bar{\theta}_n)} - B_{\theta_0, \hat{G}_n(\bar{\theta}_n)} h_{\hat{G}_n(\bar{\theta}_n)}) \\ &= P_0(\dot{\ell}_{\theta_0, \hat{G}_n(\bar{\theta}_n)} - B_{\theta_0, \hat{G}_n(\bar{\theta}_n)} h_0 + \hat{G}_n(\bar{\theta}_n) h_0) \\ &= \Psi_1(\theta_0, \hat{G}_n(\bar{\theta}_n)) - \Psi_2(\theta_0, \hat{G}_n(\bar{\theta}_n)) h_0 \\ &= \xi \Psi(\theta_0, \hat{G}_n(\bar{\theta}_n)), \end{aligned} \tag{5.7}$$

where the first step follows from (5.2), the second from the definitions of h_G and $B_{\theta,G}$, the third from (4.7), and $\xi : \mathbb{R}^q \times \ell^\infty(\mathcal{H}) \to \mathbb{R}^q$ is defined by $\xi(a, T) = a - Th_0$. By (D5), Ψ is Fréchet differentiable at (θ_0, G_0), so that

$$\Psi(\theta_0, \hat{G}_n(\bar{\theta}_n)) = \Psi(\theta_0, \hat{G}_n(\bar{\theta}_n)) - \Psi(\theta_0, G_0) = \dot{\Psi}(0, \hat{G}_n(\bar{\theta}_n) - G_0) + o_p(\|\hat{G}_n(\bar{\theta}_n) - G_0\|_{\mathcal{H}}).$$

With ξ linear and continuous, (5.7) now becomes

$$P_0 \dot{\ell}_{\theta_0, \theta_0, \hat{G}_n(\bar{\theta}_n)} = \xi \dot{\Psi}(0, \hat{G}_n(\bar{\theta}_n) - G_0) + o_p(\|\hat{G}_n(\bar{\theta}_n) - G_0\|_{\mathcal{H}}) = o_p(\|\hat{G}_n(\bar{\theta}_n) - G_0\|_{\mathcal{H}}), \tag{5.8}$$

where the second step is due to the fact that $\xi\dot{\Psi}(0, G - G_0) = 0$ for all G. In view of (5.8), (5.6) will follow as soon as

$$\hat{G}_n(\bar{\theta}_n) - G_0 = O_p(|\bar{\theta}_n - \theta_0| + n^{-1/2}). \tag{5.9}$$

To this end, note that (2.5) continues to hold with $(\hat{\theta}, \hat{G})$ replaced by $(\bar{\theta}_n, \hat{G}_n(\bar{\theta}_n))$. In particular, for $h \in \mathcal{H}$,

$$\begin{aligned}\sqrt{n}(\hat{G}_n(\bar{\theta}_n) - G_0)h &= \sqrt{n}\mathbb{P}_n B_{\bar{\theta}_n, \hat{G}_n(\bar{\theta}_n)}h - \sqrt{n}P_0 B_0 h \\ &= \mathbb{G}_n B_{\bar{\theta}_n, \hat{G}_n(\bar{\theta}_n)}h + \sqrt{n}P_0(B_{\bar{\theta}_n, \hat{G}_n(\bar{\theta}_n)} - B_0)h \\ &= \mathbb{G}_n B_{\bar{\theta}_n, \hat{G}_n(\bar{\theta}_n)}h + \sqrt{n}\{\Psi_2(\bar{\theta}_n, \hat{G}_n(\bar{\theta}_n)) - \Psi_2(\theta_0, G_0)\}h + \sqrt{n}(\hat{G}_n(\bar{\theta}_n) - G_0)h.\end{aligned} \tag{5.10}$$

Applying once again the differentiability of Ψ at (θ_0, G_0), we obtain

$$\begin{aligned}\sqrt{n}\{\Psi_2(\bar{\theta}_n, \hat{G}_n(\bar{\theta}_n)) - \Psi_2(\theta_0, G_0)\}h = -(P_0 B_0 h\dot{\ell}_0^T)\sqrt{n}(\bar{\theta}_n - \theta_0) - \sqrt{n}(\hat{G}_n(\bar{\theta}_n) - G_0)B_0^* B_0 h \\ + o_p(\sqrt{n}|\bar{\theta}_n - \theta_0|) + o_p(\sqrt{n}\|\hat{G}_n(\bar{\theta}_n) - G_0\|_{\mathcal{H}})\end{aligned} \tag{5.11}$$

uniformly in h. It follows from (D8′b) that

$$\{B_0^* B_0 h : h \in \mathcal{H}\} \supset c\mathcal{H} \tag{5.12}$$

for some $c > 0$. Combine (5.10)–(5.12), take suprema over h, and conclude that

$$\sqrt{n}\|\hat{G}_n(\bar{\theta}_n) - G_0\|_{\mathcal{H}} \le c^{-1}\|\mathbb{G}_n B_{\bar{\theta}_n, \hat{G}_n(\bar{\theta}_n)}\|_{\mathcal{H}} + O_p(\sqrt{n}|\bar{\theta}_n - \theta_0|) + o_p(\sqrt{n}\|\hat{G}_n(\bar{\theta}_n) - G_0\|_{\mathcal{H}}).$$

From this (5.9) follows, provided the first term on the right is $O_p(1)$. The latter can be ascertained using the next lemma, an extension of Lemma 19.24 of van der Vaart (1998).

Lemma 5.1. *Let $\mathcal{H}$ be a set, $\mathcal{F} \subset L_2(P)$ a Donsker class, $B : \mathcal{H} \to \mathcal{F}$, and (B_m) a sequence of random maps such that $\sup_{h\in\mathcal{H}} \|(B_m - B)h\|_{P,2} = o_p(1)$. Then $\mathbb{G}_m(B_m - B) = o_p(1)$ in $\ell^\infty(\mathcal{H})$.*

Proof. Without loss of generality, assume $B = 0$. Let $\ell^\infty(\mathcal{H}, \mathcal{F})$ denote the space of L_2-bounded maps from $\mathcal{H}$ into $\mathcal{F}$, and let $g : \ell^\infty(\mathcal{F}) \times \ell^\infty(\mathcal{H}, \mathcal{F}) \to \ell^\infty(\mathcal{H})$ be defined by $g(a, b) = a \circ b - a \circ 0$, where $\circ$ denotes composition. Then g is continuous at every (a, b) such that a is uniformly L_2-continuous. Indeed, if $(a_l, b_l) \to (a, b)$, then

$$\begin{aligned}\|a_l \circ 0 - a \circ 0\|_{\mathcal{H}} &= |a_l(0) - a(0)| \le \|a_l - a\|_{\mathcal{F}} \to 0, \\ \|a_l \circ b_l - a \circ b\|_{\mathcal{H}} &\le \|a_l \circ b_l - a \circ b_l\|_{\mathcal{H}} + \|a \circ b_l - a \circ b\|_{\mathcal{H}} \\ &\le \|a_l - a\|_{\mathcal{F}} + \|a \circ b_l - a \circ b\|_{\mathcal{H}} \to 0,\end{aligned}$$

by uniform continuity of a and uniform convergence of b_l to b.

By assumption, $B_m \to 0$ in probability in $\ell^\infty(\mathcal{H}, \mathcal{F})$. $\mathcal{F}$ being Donsker means that $\mathbb{G}_m$ tends weakly to $\mathbb{G}_p$, a P-Brownian bridge in $\ell^\infty(\mathcal{F})$. It follows that $(\mathbb{G}_m, B_m)$ tends weakly to $(\mathbb{G}_p, 0)$ in $\ell^\infty(\mathcal{F}) \times \ell^\infty(\mathcal{H}, \mathcal{F})$.

Almost all sample paths of $\mathbb{G}_p$ are uniformly continuous for the L_2-metric. By the continuous mapping theorem,

$$\mathbb{G}_m B_m = \mathbb{G}_m \circ B_m - \mathbb{G}_m \circ 0 = g(\mathbb{G}_m, B_m) \to g(\mathbb{G}_p, 0) = 0$$

weakly and hence in probability. □

(D7′) and the dominated convergence theorem together imply that

$$\sup_{h\in\mathcal{H}} \|(B_{\theta,G} - B_0)h\|_{P_0,2} \to 0, \qquad (\theta, G) \to (\theta_0, G_0).$$

This continuity, combined with the weak consistency of $(\bar{\theta}_n, \hat{G}_n(\bar{\theta}_n))$, further impies that

$$\sup_{h\in\mathcal{H}} \|(B_{\bar{\theta}_n,\hat{G}_n(\bar{\theta}_n)} - B_0)h\|_{P_0,2} = o_p(1).$$

Using this and (D6) in Lemma 5.1 then gives

$$\mathbb{G}_n B_{\bar{\theta}_n,\hat{G}_n(\bar{\theta}_n)} = \mathbb{G}_n B_0 + o_p(1) = O_p(1)$$

in $\ell^\infty(\mathcal{H})$. This establishes (5.9) and hence (11) of MV.

Theorem 1 of MV requires that almost surely,

$$\begin{aligned} \dot{\ell}_{t,\theta,G} &\to \dot{\ell}_{\theta_0,\theta_0,G_0}, \\ \ddot{\ell}_{t,\theta,G} &\to \ddot{\ell}_{\theta_0,\theta_0,G_0}, \end{aligned} \tag{5.13}$$

as $(t, \theta, G) \to (\theta_0, \theta_0, G_0)$. To see this, note first that

$$\|G_t(\theta, G) - G_0\|_{\mathcal{H}} \to 0, \tag{5.14}$$

$$\|h_{t,\theta,G} - h_0\|_\infty \to 0, \tag{5.15}$$

as $(t, \theta, G) \to (\theta_0, \theta_0, G_0)$, which can be checked by elementary means. Next write

$$\begin{aligned} \left|\left(\Pi^Y_{t,G_t(\theta,G)}\dot{\ell}^{ZY}_t - \Pi^Y_0\dot{\ell}^{ZY}_0\right)(y)\right| &\le \left|\Pi^Y_{t,G_t(\theta,G)}\left(\dot{\ell}^{ZY}_t - \dot{\ell}^{ZY}_0\right)(y)\right| + \left|\left(\Pi^Y_{t,G_t(\theta,G)} - \Pi^Y_0\right)\dot{\ell}^{ZY}_0(y)\right| \\ &\le \sup_{z\in\mathcal{Z}}\left|\left(\dot{\ell}^{ZY}_t - \dot{\ell}^{ZY}_0\right)(z,y)\right| + \left|\left(\Pi^Y_{t,G_t(\theta,G)} - \Pi^Y_0\right)\dot{\ell}^{ZY}_0(y)\right|. \end{aligned}$$

(B1), (B3) and (D2′) together imply that the class $\{\theta \mapsto \dot{\ell}^{ZY}_\theta(z, y) : z \in \mathcal{Z}\}$ is equicontinuous. Hence the first term on the right side tends to 0 as $(t, \theta, G) \to (\theta_0, \theta_0, G_0)$. (D7′) and (5.14) together imply that the second term tends to 0 as well. This takes care of one term in the expressions ((5.2) and (5.3)) for $\dot{\ell}_{t,\theta,G}$ and $\ddot{\ell}_{t,\theta,G}$; the others can be treated similarly.

Under (D6), the Donsker condition of MV reduces to

(E1) $\{B_{t,G_t(\theta,G)}h_{t,\theta,G} : \|(t - \theta_0, \theta - \theta_0, G - G_0)\| < \tau\}$ is Donsker for some $\tau > 0$.

In view of (5.14) and (D6), (E1) will follow if for (t, θ, G) in a neighborhood of $(\theta_0, \theta_0, G_0)$, $h_{t,\theta,G}$ is bounded in $\mathbb{B}$. The latter certainly holds under a stronger version of (5.15):

$$h_{t,\theta,G} \to h_0 \text{ componentwise in } \mathbb{B}, \qquad (t, \theta, G) \to (\theta_0, \theta_0, G_0).$$

This, however, does not appear to be automatic, and E1 remains a condition. MV also assume that

(E2) $\{\ddot{\ell}_{t,\theta,G} : |(t-\theta_0, \theta-\theta_0, G-G_0)\| < \tau\}$ is GC for some $\tau > 0$.

Lastly, (D4′) garantees the existence of (square-)integrable envelope functions required by Theorem 1 of MV.

Thus all conditions of MV's Theorem 1 and its corollaries have been established. In return, we have the following result. Write $pl_n(\theta) = \log PL_n(\theta)$.

Theorem 5.2. *Assume (E0)–(E2). Then, for every sequence $\bar{\theta}_n = \theta_0 + o_p(1)$, we have*

$$pl_n(\bar{\theta}_n) = pl_n(\theta_0) + n(\bar{\theta}_n - \theta_0)^T \mathbb{P}_n \dot{\ell}_e - n(\bar{\theta}_n - \theta_0)^T I_e (\bar{\theta}_n - \theta_0)/2 + o_p(\sqrt{n}\|\bar{\theta}_n - \theta_0\| + 1)^2.$$

In particular,

$$\sqrt{n}(\hat{\theta}_n - \theta_0) = \mathbb{G}_n I_e^{-1} \dot{\ell}_e + o_p(1), \tag{5.16}$$

$$2\{pl_n(\hat{\theta}_n) - pl_n(\theta_0)\} = n(\hat{\theta}_n - \theta_0)^T I_e (\hat{\theta}_n - \theta_0) + o_p(1), \tag{5.17}$$

$$-2\{pl_n(\hat{\theta}_n + u_n v_n) - pl_n(\hat{\theta}_n)\}/(nu_n^2) = v^T I_e v + o_p(1), \tag{5.18}$$

for all sequences $v_n = v(\in \mathbb{R}^q) + o_p(1)$ and $u_n = o_p(1)$ with $(\sqrt{n}u_n)^{-1} = O_p(1)$.

(5.16) says that $\sqrt{n}(\hat{\theta}_n - \theta_0) \to N(0, I_e^{-1})$. This is part of the conclusion of Corollary 4.4 and is not new to us. What is new here is that, by (5.18), I_e can be consistently estimated by perturbing the profile log-likelihood for θ around $\hat{\theta}_n$. This makes possible Wald tests and related confidence statements about θ. Furthermore, (5.17) implies that the profile likelihood ratio statistic, like its parametric analogue, is asymptotically chi-squared with q degrees of freedom. This justifies testing hypotheses about θ using a profile likelihood ratio test and constructing confidence sets by inverting this test, just like in a parametric model.

6.0 EXAMPLES

So far we have made a number of assumptions on the regression model. While some of them (such as (D3′)) are fairly simple and indeed hold quite generally, others appear more complicated and may take considerable effort to verify. The verification of the Donsker conditions (D6) and (E1), in particular, depends heavily on special structures of the regression model. Here we use two examples to illustrate some techniques for verifying these conditions. It is not difficult to see that when $\mathcal{Y}$ is finite, many problems (including (D6) and (E1)) become much easier. The results obtained, however, will not be very different from the existing results of Wild (1991). With this in mind we focus on an infinite $\mathcal{Y}$ such as an interval in the real line or the set of nonnegative integers.

6.1 LINEAR REGRESSION

Suppose $\mathcal{Z} = [l, u] \subset \mathbb{R}$, $\mathcal{Y} = \mathbb{R}$ and Θ is a compact subset of $\mathbb{R}^2 \times (0, \infty)$. Write $\theta = (\beta_0, \beta_1, \sigma^2)$ and assume

$$f(y|z;\theta) = \big(2\pi\sigma^2\big)^{-1/2} \exp\big\{ -(y - \beta_0 - \beta_1 z)^2 / \big(2\sigma^2\big)\big\}.$$

(μ is, of course, the Lebesgue measure.) Assume that G_0 is nondegenerate and that the true parameter, $\theta_0 = (\beta_{00}, \beta_{10}, \sigma_0^2)$, lies in the interior of Θ. Also assume (A2) and (A3).

This model has an exponential family structure. In light of this and the discussion in Chapter 3, conditions (C0)–(C3) are all satisfied. For (C4), write

$$\begin{aligned}
\sup_{z,\theta,G} (1-r)\frac{f(y|z;\theta)}{f(y;G,\theta)} &\le \sup_{\theta} \frac{\sup_z f(y|z;\theta)}{\inf_z f(y|z;\theta)} \\
&= \sup_{\theta} \big(2\pi\sigma^2\big)^{-1/2} \exp\big[\beta_1(z_1 - z_2)\{y - \beta_0 - \beta_1(z_1 + z_2)/2)\}/\sigma^2\big] \\
&\le a \exp(b|y| + c),
\end{aligned}$$

where $z_1, z_2 \in \mathcal{Z}$ depend on y but a, b, c do not. The conditional moment generating function of Y given Z exists everywhere on the real line and is continuous in Z. With $\mathcal{Z}$ compact, the marginal moment generating function of Y (or $|Y|$) must exist everywhere. Hence the bound above is P_0-integrable and (C4) follows. It is easy to see that

$$\sup_z f(y|z;\theta_0) \le f(y|l;\theta_0) + f(y|u;\theta_0) + (2\pi\sigma_0^2)^{-1} 1\{\beta_{00} + \beta_{10} l \le y \le \beta_{00} + \beta_{10} u\}.$$

By the dominated convergence theorem, π_0^Z is continuous. Combined with (A3), (B1) and (B4), this implies (C5′). It then follows from Theorem 3.5 that, for every $L_1(G_0)$-bounded GC class $\mathcal{H}$, $|\hat{\theta}_n-\theta_0|\vee\|\hat{G}_n-G_0\|_{\mathcal{H}}\to 0$ almost surely.

We now show that, for $\mathcal{H}=C_1^1(\mathcal{Z})$, $\sqrt{n}(\hat{\theta}_n-\theta_0,\hat{G}_n-G_0)$ converges weakly to a tight Gaussian process in $\mathbb{R}^3\times\ell^\infty(\mathcal{H})$ as described in Corollary 4.4. This follows by verifying the conditions of Corollary 4.4. Since (D1), (D2) and (D3′) are straightforward, let us start with (D4). In general,

$$|\dot{\ell}_{\theta,G}(z,y,r)|\le\sup_{z',\theta'}|\dot{\ell}_{\theta'}^{ZY}(z',y)| \tag{6.1}$$

for all (θ,G) and all (z,y,r). Thus (D4) would follow, with much to spare, from the square-integrability of the right side of (6.1). In this example,

$$\dot{\ell}_\theta^{ZY}(z,y)=\begin{pmatrix}(y-\beta_0-\beta_1 z)/\sigma^2\\ z(y-\beta_0-\beta_1 z)/\sigma^2\\ -1/(2\sigma^2)+(y-\beta_0-\beta_1 z)^2/(2\sigma^4)\end{pmatrix}. \tag{6.2}$$

Hence the right side of (6.1) is majorized by a quadratic function of $|y|$. Given Z, Y has conditional moments of all orders, and they are all continuous functions of Z. With $\mathcal{Z}$ compact, any polynomial of Y (or $|Y|$) has finite mean. This verifies (D4).

In view of the continuity of $\dot{\ell}_\theta^{ZY}$ in θ, (D7) will follow if for every y,

$$\frac{\int\dot{f}(y|z;\theta)dG(z)}{f(y;G,\theta)}-\frac{\int\dot{f}(y|z;\theta_0)dG_0(z)}{f(y;G_0,\theta_0)}\to 0, \tag{6.3}$$

$$\sup_{h\in\mathcal{H}}\left|\frac{\int h(z)f(y|z;\theta)dG(z)}{f(y;G,\theta)}-\frac{\int h(z)f(y|z;\theta_0)dG_0(z)}{f(y;G_0,\theta_0)}\right|\to 0, \tag{6.4}$$

as $|\theta-\theta_0|\vee\|G-G_0\|_{\mathcal{H}}\to 0$. With $\mathcal{H}=C_1^1(\mathcal{Z})$, convergence in $\ell^\infty(\mathcal{H})$ is simply weak convergence. Lemma 2.2 then yields $f(y;G,\theta)\to f(y;G_0,\theta_0)$. In fact, since $\dot{f}(y|\cdot;\cdot)$ is continuous, the same argument can be used to show that $\int\dot{f}(y|z;\theta)dG(z)\to\int\dot{f}(y|z;\theta_0)dG_0(z)$. Thus (6.3) follows. In the proof of Lemma 3.4, it is shown that the probability measure with density $f(y|\cdot;\theta)/f(y;G,\theta)$ with respect to G converges weakly to the one with density $f(y|\cdot;\theta_0)/f(y;G_0,\theta_0)$ with respect to G_0. Now (6.4) follows from Theorem 1.12.1 of van der Vaart and Wellner (1996) (VW).

To check (D8′a), observe that

$$\begin{aligned}\big|df(y|z;\theta_0)/dz\big|&=\big|f(y|z;\theta_0)(y-\beta_{00}-\beta_{10}z)\beta_{10}/\sigma_0^2\big|\\&\le(a|y|+b)\big[f(y|l;\theta_0)+f(y|u;\theta_0)+1\{\beta_{00}+\beta_{10}l\le y\le\beta_{00}+\beta_{10}u\}\big]\end{aligned} \tag{6.5}$$

for some constants a and b. $M(y)$ defined as (6.5) satisfies (4.30) and (4.31) for $\alpha=1$ because $|\dot{\ell}_0^Y(y)|$ is bounded by a quadratic function of $|y|$. By differentiating $\dot{f}(y|z;\theta_0)$ with respect to z, a similar procedure gives an M' that satisfies (4.32) and (4.33). Thus (D8′a) is verified. For (D8′b), note that M defined by (6.5) is μ-integrable, which implies (4.37). Furthermore, the second derivative of $f(y|z;\theta_0)$ with respect to z can be bounded in a familiar way (by a polynomial of $|y|$ times the second component of (6.5)), so that (4.36) is true for $\beta=1$. Now deduce (D8′b) from Lemma 4.5.

We now turn to (D5). Condition (4.14) typically follows from a dominated convergence argument which requires that, in a neighborhood of θ_0, $|\ddot{f}(\cdot|\cdot;\theta)|$ is dominated by a $\mu \times G_0$-integrable function. This is true quite generally. In the present example, the majorization argument in the above paragraph can be used to show that $\sup_{z,\theta} |\ddot{f}(\cdot|z;\theta)|$ is μ-integrable, a much stronger result. (4.15) follows from the mean-value theorem and the dominated convergence theorem, upon noting that $\ddot{\ell}_\theta^{ZY}$ is continuous in θ and is bounded by a polynomial of $|y|$. (4.16) follows in a similar way. First, $\ddot{\ell}_{\theta,G}^{Y}$ is continuous in (θ, G) because $f(y|z;\theta)$, $\dot{f}(y|z;\theta)$ and $\ddot{f}(y|z;\theta)$ are all continuous in (z,θ). Next,

$$\sup_{\theta,G} \left|\ddot{\ell}_{\theta,G}^{Y}(y)\right| \le \sup_{z,\theta} \left|\ddot{f}(y|z;\theta)/f(y|z;\theta)\right| + \sup_{z.\theta} \left|\dot{\ell}_{\theta}^{ZY}(z,y)\right|^2$$

for all y. The right side is again bounded by a polynomial of $|y|$. In general, (4.17) will hold if there exists $\eta_G > 0$ such that $\eta_G \to 0$ as $\|G - G_0\|_{\mathcal{H}} \to 0$ and $\Delta_G \in \eta_G \mathcal{H}$ for every G, where

$$\Delta_G(z) := \int (1 - \pi_0(y))(\dot{\ell}_{\theta_0,G}^{Y} - \dot{\ell}_0^{Y})(y) f(y|z;\theta_0) d\mu(y).$$

With $\mathcal{H} = C_1^1(\mathcal{Z})$, this is equivalent to $\|\Delta_G\|_1 \to 0$ (where $\|\cdot\|_1$ is the $C^1(\mathcal{Z})$ norm not L_1). In this example, (6.1)–(6.3) and the dominated convergence theorem imply that $\Delta_G \to 0$ pointwise. In view of the boundedness of $\mathcal{Z}$, it suffices to show that the first-order Lipschitz norm of Δ_G tends to 0. This is bounded by

$$\int (1 - \pi_0(y))(\dot{\ell}_{\theta_0,G}^{Y} - \dot{\ell}_0^{Y})(y) \sup_z |df(y|z;\theta_0)/dz| d\mu(y),$$

which tends to 0 by (6.5) and the dominated convergence theorem. Condition (4.18) follows from a familiar dominated convergence argument. Condition (4.19) follows from the uniform boundedness of $\mathcal{H}$, the uniform convergence (6.4) and the dominated convergence theorem. Clearly, (4.20) will hold if

$$\sup_{h \in C_1^1(\mathcal{Z})} \|(B_{\theta,G}^* B_{\theta,G} - B_0^* B_0)h\|_1 \to 0,$$

that is, if $B_{\theta,G}^* B_{\theta,G} \to B_0^* B_0$ in $\mathcal{L}(C^1(\mathcal{Z}))$, the space of continuous linear maps from $C^1(\mathcal{Z})$ into itself. Write

$$\begin{aligned}(B_{\theta,G}^* B_{\theta,G} - B_0^* B_0)h(z) = h(z) &\int \pi_0(y)(f(y|z;\theta) - f(y|z;\theta_0)) d\mu(y) \\ &+ \int (1 - \pi_0(y)) \frac{\int h f(y|\cdot;\theta) dG}{f(y;G,\theta)} (f(y|z;\theta) - f(y|z;\theta_0)) d\mu(y) \\ &+ \int (1 - \pi_0(y)) \left(\frac{\int h f(y|\cdot;\theta) dG}{f(y;G,\theta)} - \frac{\int h f(y|\cdot;\theta_0) dG_0}{f(y;G_0,\theta_0)} \right) f(y|z;\theta_0) d\mu(y).\end{aligned}$$

Using the same arguments as before, it can be shown that each term on the right side tends to 0 in $C^1(\mathcal{Z})$, uniformly over $h \in C_1^1(\mathcal{Z})$. The details are omitted.

It remains to check the Donsker condition (D6). By Theorem 2.10.6 of VW, it suffices to show that for

some $\tau > 0$, these four classes are all P_0-Donsker:

$$C_1^1(\mathcal{Z}), \tag{6.6}$$

$$\{\dot{\ell}_\theta^{ZY} : |\theta - \theta_0| < \tau\}, \tag{6.7}$$

$$\{\Pi_{\theta,G}^Y h : \|(\theta - \theta_0, G - G_0)\| < \tau, h \in C_1^1(\mathcal{Z})\}, \tag{6.8}$$

$$\{\dot{\ell}_{\theta,G}^Y : \|(\theta - \theta_0, G - G_0)\| < \tau\}. \tag{6.9}$$

That (6.6) is Donsker is a direct application of Corollary 2.7.2 of VW; in fact, this is the main motivation for choosing $\mathcal{H} = C_1^1(\mathcal{Z})$. $\dot{\ell}_\theta^{ZY}(z, y)$ is clearly Lipschitz in θ, with Lipschitz constant bounded by a polynomial of $|y|$. By Theorem 2.7.11 of VW, (6.7) is Donsker (for any τ). For (6.8), note first that it is uniformly bounded. Next, from elementary calculus,

$$\frac{d}{dy}\Pi_{\theta,G}^Y h(y) = \operatorname{cov}\left(h(Z), \frac{d}{dy}\log f(y|Z;\theta)\middle| Y = y; \theta, G\right) \le a|y| + b$$

for some constants a and b. Thus for every θ, G, h and every $j \in \mathbb{N}$, the restriction of $\Pi_{\theta,G}^Y h$ to $(-j-1, -j] \cup [j, j+1)$ has C^1 norm bounded by a constant multiple of j. By the results of van der Vaart (1994b), (6.8) will be a Donsker class if Y has finite fourth moment, which is certainly true. (This argument works for any uniformly bounded class $\mathcal{H}$.) The Donsker property of (6.9) follows by a similar argument. First, $|\dot{\ell}_{\theta,G}^Y(y)|$ is bounded by a polynomial of $|y|$. Second,

$$\frac{d}{dy}\dot{\ell}_{\theta,G}^Y(y) = E\left(\frac{d}{dy}\dot{\ell}_\theta^{ZY}(Z, y)\middle| Y = y; \theta, G\right) + \operatorname{cov}\left(\dot{\ell}_\theta^{ZY}(Z, y), \frac{d}{dy}\log f(y|Z;\theta)\middle| Y = y; \theta, G\right).$$

The right side can be bounded by a polynomial of $|y|$. This completes the verification of the regularity conditions of Corollary 4.4.

Now consider the conditions of Theorem 5.2. The slight differences between (E0) and conditions in section 4 apparently cause no trouble; the arguments used in previous paragraphs remain applicable. (E1) can be established by showing that the classes $\{\Pi_{t,G_t(\theta,G)}^Y h_{t,\theta,G}\}$ and $\{h_{t,\theta,G}\}$ are Donsker for (t, θ, G) in a neighborhood of $(\theta_0, \theta_0, G_0)$. The first follows from the argument for (6.8) in the last paragraph, which requires only that $\mathcal{H}$ be uniformly bounded. For the second, recall that $h_{t,\theta,G} = h_G/\{1 - (t-\theta)^T h_G\}$. Both the numerators $\{h_G\}$ and the denominators $\{1 - (t-\theta)^T h_G\}$ are uniformly bounded Donsker classes, and the denominators can be uniformly bounded away from 0. By Theorem 2.10.6 of VW, the ratios form a Donsker class. In view of (D4′) and the weak compactness of $\mathcal{G}$ (Lemma 2.2), (E2) will follow from Lemma 3.3 if $\ddot{\ell}_{t,\theta,G}(z, y, r)$ can be shown to be continuous in (t, θ, G) for every (z, y, r). The latter can be argued the same way as for (5.13) if, for every sequence $(t_m, \theta_m, G_m) \to (t, \theta, G)$ and every y,

$$\left\|G_{t_m}(\theta_m, G_m) - G_t(\theta, G)\right\|_{\mathcal{H}} \to 0, \tag{6.10}$$

$$\left\|h_{t_m,\theta_m,G_m} - h_{t,\theta,G}\right\|_\infty \to 0, \tag{6.11}$$

$$\left(\Pi_{\theta_m,G_m}^Y - \Pi_{\theta,G}^Y\right)h(y) \to 0, \tag{6.12}$$

with $h = h_0$, $\dot{\ell}_0^{ZY}$, $\left(\dot{\ell}_0^{ZY}\right)^{\otimes 2}$, $\dot{\ell}_0^{ZY} h_0^T$, $\ddot{\ell}_0^{ZY}$. To see (6.10), write

$$(G_{t_m}(\theta_m, G_m) - G_t(\theta, G))h = (G_m - G)h - (G_m - G)\{(t_m - \theta_m)^T h_{G_m} h\} \\ - G\{(t_m - \theta_m)^T h_{G_m} h - (t - \theta)^T h_G h\}.$$

Of course, the first term on the right tends to 0 uniformly over $h \in \mathcal{H}$. The class $\{(t-\theta)^T h_G\}$ is bounded in $\mathbb{B} = C^1(\mathcal{Z})$ (of which $\mathcal{H}$ is the unit ball). The set of pairwise products of two bounded subsets of $C^1(\mathcal{Z})$ is again bounded in $C^1(\mathcal{Z})$. Hence the second term is negligible (uniformly in h). The last term is controlled using the dominated convergence theorem. For (6.11), observe that $\|h_{G_m} - h_G\|_\infty \to 0$, that $\|(t_m - \theta_m)^T h_{G_m} - (t-\theta)^T h_G\|_\infty \to 0$ and that $1 - (t_m - \theta_m)^T h_{G_m}$ is bounded away from 0 for large m. For (6.12), it suffices to note that all choices of h are, for fixed y, bounded and continuous in z. Thus all conditions of Theorem 5.2 are verified, and we have a consistent estimate of the asymptotic variance of $\hat{\theta}_n$ as well as a profile likelihood ratio test.

These arguments can in principle be extended to higher dimensions ($d > 1$). Then $C_1^1(\mathcal{Z})$ is not known to be a Donsker class. Instead, one may use as $\mathcal{H}$ the unit ball of $C^\alpha(\mathcal{Z})$ for some $\alpha > d/2$ or of a completely different Banach space.

6.2 POISSON REGRESSION

As a second example, consider a Poisson regression model, again with a one-dimensional missing covariate. Let $\mathcal{Z} = [l, u] \subset \mathbb{R}$, let $\mathcal{Y} = \{0\} \cup \mathbb{N}$ and let Θ be a compact subset of $\mathbb{R}^2$. With μ being the counting measure and $\theta = (\beta_0, \beta_1)$, suppose that

$$f(y|z;\theta) = \exp\{y(\beta_0 + \beta_1 z) - \exp(\beta_0 + \beta_1 z)\}/y!.$$

As before, assume that G_0 is nondegenerate, that $\theta_0 = (\beta_{00}, \beta_{10})$ is an interior point of Θ, and that A2 and A3 hold.

As in the previous example, conditions (C0)–(C5) are easily seen to hold; hence $(\hat{\theta}_n, \hat{G}_n)$ is strongly consistent for (θ_0, G_0) with respect to the topologies described in Theorem 3.5. We now use Corollary 4.4, again with $\mathbb{B} = C^1(\mathcal{Z})$ and $\mathcal{H} = C_1^1(\mathcal{Z})$, to show that $(\hat{\theta}_n, \hat{G}_n)$ is asymptotically normal and achieves the information bound for θ. (D1), (D2) and (D3′) do not take much work, and (D7) follows from exactly the same argument as in the previous example. In fact, the arguments therein can also be used to verify (D4), (D5) and (D8′), upon making the following observations. First, Y has finite moments of all orders. Second, for all $y \in \mathcal{Y}$,

$$\begin{aligned} \sup_{z,\theta} f(y|z;\theta) &\le \sup_{\lambda_1 \le \lambda \le \lambda_2} \exp(y \log \lambda - \lambda)/y! \\ &\le \exp(y \log \lambda_1 - \lambda_1)/y! + \exp(y \log \lambda_2 - \lambda_2)/y! + c 1_{[\lambda_1, \lambda_2]}(y)\}, \end{aligned} \tag{6.13}$$

where

$$\lambda_1 = \min_{z,\theta} \exp(\beta_0 + \beta_1 z),$$

$$\lambda_2 = \max_{z,\theta} \exp(\beta_0 + \beta_1 z),$$

$$c = \max_{\lambda_1 \le y \le \lambda_2} \exp(y \log y - y)/y!.$$

The right side of (6.13) as a function of y is certainly μ-integrable.

It remains to check the Donsker condition (D6). While (6.6) and (6.7) are easily seen to be Donsker, (6.8) and (6.9) require more effort. With $\mathcal{Y}$ nonconvex, the smoothness argument used in the last example is no longer viable. Here we take a different approach. Consider (6.8) first. Let $\underline{\mathcal{G}}$ denote the set of subprobablity measures (measures with total mass at most 1) on $\mathcal{Z}$, and note that $\int h(z) f(y|z;\theta) dG(z)$ with $\|h\|_\infty \le 1$ can be written as $f(y;\underline{G},\theta) := \int f(y|z;\theta) d\underline{G}(z)$ for some $\underline{G} \in \underline{\mathcal{G}}$. Hence (6.8) is contained in

$$\{f(Y;\underline{G}_1,\theta_1)/f(Y;G_2,\theta_2) : \theta_1,\theta_2 \in \Theta, \underline{G}_1 \in \underline{\mathcal{G}}, G_2 \in \mathcal{G}\}.$$

By Theorems 2.10.1 and 2.10.3 of VW, it suffices to show that

$$\{f(Y;G_1,\theta_1)/f(Y;G_2,\theta_2) : \theta_1,\theta_2 \in \Theta, G_1, G_2 \in \mathcal{G}\}.$$

is a Donsker class. Simple algebraic manipulation yields

$$\left| \frac{f(y;G_1,\theta_1)}{f(y;G_2,\theta_2)} - \frac{f(y;G_1',\theta_1')}{f(y;G_2',\theta_2')} \right| \le y! \exp(ay+b) \sum_{j=1}^{2} \left| f(y;G_j,\theta_j) - f(y;G_j',\theta_j') \right|$$

for some constants $a, b \in \mathbb{R}$. In view of Corollary 2.10.13 of VW, it now suffices to show that

$$\{Y! \exp(aY+b) f(Y;G,\theta) : \theta \in \Theta, G \in \mathcal{G}\} \tag{6.14}$$

is Donsker. To this end, observe that the class

$$\{\exp\{(a+\beta_0+\beta_1 z)Y + b - \exp(\beta_0+\beta_1 z)\} : z \in \mathcal{Z}, \theta \in \Theta\} \tag{6.15}$$

is VC (cf. proof of Lemma 3.2) and pointwise separable. Furthermore, it is dominated by a square integrable function because the conditional moment generating function of Y given Z exists everywhere and is continuous in Z. Theorem 2.5.2 of VW then says that (6.15) is Donsker. Its convex hull contains

$$\{Y! \exp(aY+b) f(Y;G,\theta) : \theta \in \Theta, G \in \mathcal{G}_d\}, \tag{6.16}$$

where $\mathcal{G}_d$ is defined in Lemma 3.2 as the collection of finitely discrete probability measures on $\mathcal{Z}$. So (6.16) is Donsker too. We now show that (6.14) is contained in the pointwise sequential closure of (6.16); by Theorem 2.10.2 of VW, this would imply the Donsker property of (6.14). Fix (θ, G). We need to find a sequence $(G_m) \subset \mathcal{G}_d$ such that $f(y;G_m,\theta) \to f(y;G,\theta)$ for every y. (Since (6.14) has a square integrable envelope, this pointwise convergence would imply L_2 convergence.) That such a sequence exists can be seen by the

following argument. By Kolmogorov's existence theorem, there is on some probability space a sequence (U_m) of independent random variables identically distributed as G. Let $\mathbb{D}_m$ denote the empirical distribution of $U_1, \ldots, U_m$. For each y, the function $z \mapsto f(y|z;\theta)$ is continuous and therefore bounded. By the law of large numbers,

$$f(y;\mathbb{D}_m,\theta) = \mathbb{D}_m f(y|Z;\theta) \to Gf(y|Z;\theta) = f(y;G,\theta) \tag{6.17}$$

almost surely for every y. Since $\mathcal{Y}$ is countable, we have that, almost surely, (6.17) holds for every y. Pick an ω in the aforementioned probability space for which this is the case, and set $G_m = \mathbb{D}_m(\omega)$.

The Donsker property of (6.9) follows from a similar argument. It can be shown that

$$\left| \frac{\int \dot{f}(y|z;\theta_1)dG_1(z)}{f(y;G_2,\theta_2)} - \frac{\int \dot{f}(y|z;\theta_1')dG_1'(z)}{f(y;G_2',\theta_2')} \right| \leq y! \exp(ay+b)$$
$$\times \left\{ \left| \int \dot{f}(y|z;\theta_1)dG_1(z) - \int \dot{f}(y|z;\theta_1')dG_1'(z) \right| + (cy+d) \left| f(y;G_2,\theta_2) - f(y;G_2',\theta_2') \right| \right\},$$

for some constants $a, b, c, d \in \mathbb{R}$. As argued in the last paragraph, it suffices to show that the classes

$$\big\{(cY+d)\exp\{(a+\beta_0+\beta_1 z)Y + b - \exp(\beta_0+\beta_1 z)\} : z \in \mathcal{Z}, \theta \in \Theta\big\}, \tag{6.18}$$

$$\big\{\{Y - \exp(\beta_0+\beta_1 z)\}\exp\{(a+\beta_0+\beta_1 z)Y + b - \exp(\beta_0+\beta_1 z)\} : z \in \mathcal{Z}, \theta \in \Theta\big\}, \tag{6.19}$$

$$\big\{z\{Y - \exp(\beta_0+\beta_1 z)\}\exp\{(a+\beta_0+\beta_1 z)Y + b - \exp(\beta_0+\beta_1 z)\} : z \in \mathcal{Z}, \theta \in \Theta\big\} \tag{6.20}$$

are all Donsker. The VC property of (6.15) and Lemma 2.6.18(vi) of VW together imply the VC property of (6.18). From here it is a small step to deduce that (6.18) is Donsker. It follows from the same argument that

$$\big\{Y\exp\{(a+\beta_0+\beta_1 z)Y + b - \exp(\beta_0+\beta_1 z)\} : z \in \mathcal{Z}, \theta \in \Theta\big\} \tag{6.21}$$

is Donsker too. The class

$$\big\{-\exp(\beta_0+\beta_1 z)\exp\{(a+\beta_0+\beta_1 z)Y + b - \exp(\beta_0+\beta_1 z)\} : z \in \mathcal{Z}, \theta \in \Theta\big\} \tag{6.22}$$

is contained in the convex hull of a multiple of (6.15), hence Donsker also. The Donsker property of (6.19) now follows from Theorem 2.10.6 of VW and the Donsker properties of (6.21) and (6.22). Lastly, (6.20) is contained in the convex hull of a multiple of (6.19).

We now turn to Theorem 5.2. With the machinery we have developed, this does not take much work. Arguments in the above paragraph, which require only that $\mathcal{H}$ be uniformly bounded, can be used to show that $\{\Pi^Y_{t,G_t(\theta,G)} h_{t,\theta,G}\}$ is Donsker. Condition (E2) and the Donsker property of $\{h_{t,\theta,G}\}$ follow from exactly the same arguments as in the last example.

7.0 COMPUTATION AND SIMULATIONS

This chapter is concerned with the implementation and finite-sample performance of the proposed semiparametric MLE. Since most of this discussion is for fixed sample size, the subscript n, which was essential in the asymptotic analysis, is now dropped. Also suppressed is the subscript 0 indicating the "truth". The computation of $(\hat{\theta}, \hat{G})$, a finite-dimensional maximization problem, is apparently easier than that of $(\tilde{\theta}, \tilde{G})$; we do not yet know of a general algorithm for computing $(\tilde{\theta}, \tilde{G})$. In fact, the asymptotic equivalence result (Theorem 3.6) even calls into question the motivation for finding $(\tilde{\theta}, \tilde{G})$. Therefore we shall focus on the restricted MLE $(\hat{\theta}, \hat{G})$ in the following discussion. Actually, with θ being the primary inferential target, numerical results are reported only for $\hat{\theta}$, although $\hat{\theta}$ and $\hat{G}$ are obtained simultaneously. In what follows, we propose an EM algorithm for computing $(\hat{\theta}, \hat{G})$ and for estimating the asymptotic variance of $\hat{\theta}$, and carry out simulation experiments comparing the proposed method with standard maximum likelihood methods based on (correct and incorrect) parametric models for G.

7.1 THE EM ALGORITHM

Given a realized sample, let z_j, n_j, (j,l), $l = 1, \ldots, n_j$, $j = 0(1), \ldots, k$ be defined as in Chapter 2. Denote $p_j = G\{z_j\}$, $j = 1, \ldots, k$, so that $\sum_{j=1}^{k} p_j = 1$. In terms of θ and $p := (p_1, \ldots, p_k)^T$, the likelihood (2.1) can be rewritten as

$$\left[\prod_{l=1}^{n_0} \left\{ \sum_{j=1}^{k} p_j f(Y_{(0,l)} | z_j; \theta) \right\} \right] \prod_{j=1}^{k} \left\{ p_j^{n_j} \prod_{l=1}^{n_j} f(Y_{(j,l)} | z_j; \theta) \right\}. \tag{7.1}$$

Maximizing (7.1) with respect to (θ, p) is a $q+k$-dimensional constrained maximization problem which, under a suitable transformation of p, can be transformed into a $q + k - 1$-dimensional unconstrained maximization problem. Therefore a Newton-type algorithm is applicable, at least in principle. But note that, as the sample size n increases, k increases at the same rate, unless Z has a finite support. Thus in a relatively large sample, Newton's method can be inefficient and/or unstable, if feasible at all.

Expression (7.1) can be considered as a parametric likelihood under the *working* assumption that G is concentrated on $\{z_j : j = 1, \ldots, k\}$. As such it can be maximized by using an EM algorithm which, naturally,

treats $\{(Z_i, Y_i) : i = 1, ..., n\}$ as complete data. The complete-data log-likelihood is simply

$$l_c(\theta, p) := \sum_{i=1}^{n} \left\{ \log f(Y_i|Z_i;\theta) + \sum_{j=1}^{k} 1_{Z_i=z_j} \log p_j \right\}.$$

Let $(\theta^{(0)}, p^{(0)})$ be an initial guess. For example, one may take as $\theta^{(0)}$ an estimate obtained from a complete-case analysis, and set $p_j^{(0)} = n_j/(n - n_0)$, $j = 1, \ldots, k$. Given $(\theta^{(m)}, p^{(m)})$, $m \geq 0$, we seek to maximize

$$\begin{aligned}
&E\Big\{ l_c(\theta, p) \Big| (R_i, R_i Z_i, Y_i)_{i=1}^n; \theta^{(m)}, p^{(m)} \Big\} \\
&= \sum_{i=1}^{n} \left[R_i \left\{ \log f(Y_i|Z_i;\theta) + \sum_{j=1}^{k} 1_{Z_i=z_j} \log p_j \right\} \right. \\
&\quad \left. + (1 - R_i) \sum_{j=1}^{k} P\Big(Z_i = z_j \Big| Y_i; \theta^{(m)}, p^{(m)} \Big) \Big\{ \log f(Y_i|z_j;\theta) + \log p_j \Big\} \right] \\
&= \sum_{j=1}^{k} \left\{ \sum_{l=1}^{n_j} \log f(Y_{(j,l)}|z_j;\theta) + \sum_{l=1}^{n_0} q_{jl} \log f(Y_{(0,l)}|z_j;\theta) + (n_j + q_{j\cdot}) \log p_j \right\},
\end{aligned}$$

where

$$\begin{aligned}
q_{jl} &:= P\Big(Z_{(0,l)} = z_j \Big| Y_{(0,l)}; \theta^{(m)}, p^{(m)} \Big) = \frac{f(Y_{(0,l)}|z_j;\theta^{(m)}) p_j^{(m)}}{\sum_{j'=1}^{k} f(Y_{(0,l)}|z_{j'};\theta^{(m)}) p_{j'}^{(m)}}, \\
q_{j\cdot} &:= \sum_{l=1}^{n_0} q_{jl}.
\end{aligned}$$

As a result,

$$\theta^{(m+1)} = \arg\max_{\theta} \sum_{j=1}^{k} \left\{ \sum_{l=1}^{n_j} \log f(Y_{(j,l)}|z_j;\theta) + \sum_{l=1}^{n_0} q_{jl} \log f(Y_{(0,l)}|z_j;\theta) \right\}, \tag{7.2}$$

$$p_j^{(m+1)} = (n_j + q_{j\cdot})/n, \qquad j = 1, \ldots, k. \tag{7.3}$$

In many examples, $\theta^{(m+1)}$ can be found by solving

$$\sum_{j=1}^{k} \left\{ \sum_{l=1}^{n_j} \dot{\ell}_\theta^{ZY}(z_j, Y_{(j,l)}) + \sum_{l=1}^{n_0} q_{jl} \dot{\ell}_\theta^{ZY}(z_j, Y_{(0,l)}) \right\} = 0 \tag{7.4}$$

for θ. Note that, if $(\theta^{(m+1)}, p^{(m+1)}) = (\theta^{(m)}, p^{(m)})$, then (7.3) and (7.4) are equivalent to the likelihood equations (2.5) and (4.5), respectively. (7.4) can be solved analytically for the normal linear model. In general, a Newton-type algorithm can be used. This application of Newton's method differs from the one mentioned earlier in that the dimension of the current problem is q, regardless of k.

A slightly modified version of this EM algorithm can be used to evaluate the profile likelihood for θ. For a given θ, $\hat{G}(\theta)$ (defined in Chapter 5) can be found by simply iterating (7.3), with $q_{j\cdot}$ replaced by $q'_{j\cdot} := \sum_{l=1}^{n_0} q'_{jl}$ where $q'_{jl} := P(Z_{(0,l)} = z_j | Y_{(0,l)}; \theta, p^{(m)})$, until convergence. In light of (5.18) in Theorem 5.2, a consistent estimate of the efficient information I_e is now available. Consider first the diagonal elements $I_e(s, s)$, $s = 1, \ldots, q$. Let e_s be a q-vector with 1 as the sth element and 0 everywhere else. Set $v_n \equiv v = e_s$

and $u_n = an^{-1/2}$ for some constant $a > 0$. Then (5.18) says that

$$2\{pl_n(\hat{\theta}) - pl_n(\hat{\theta} + an^{-1/2}e_s)\}/a^2 = I_e(s,s) + o_p(1).$$

The left side can be interpreted as a numerical second-order partial derivative. Naturally the desired derivative can be approximated from the opposite direction as well. In other words, e_s can be replaced by its negative to yield

$$2\{pl_n(\hat{\theta}) - pl_n(\hat{\theta} - an^{-1/2}e_s)\}/a^2 = I_e(s,s) + o_p(1).$$

Common wisdom then suggests taking the average of the two and estimating $I_e(s,s)$ by

$$\{2pl_n(\hat{\theta}) - pl_n(\hat{\theta} + an^{-1/2}e_s) - pl_n(\hat{\theta} - an^{-1/2}e_s)\}/a^2.$$

For an off-diagonal element $I_e(s,t)$, $s \neq t$, let $e_{st} = e_s + e_t$. Then a consistent estimate of $e_{st}^T I_e e_{st} = I_e(s,s) + I_e(t,t) + 2I_e(s,t)$ is given by

$$\{2pl_n(\hat{\theta}) - pl_n(\hat{\theta} + an^{-1/2}e_{st}) - pl_n(\hat{\theta} - an^{-1/2}e_{st})\}/a^2.$$

It follows that $I_e(s,t)$ can be consistently estimated by

$$\frac{1}{2a^2}\Big\{pl_n(\hat{\theta} + an^{-1/2}e_s) + pl_n(\hat{\theta} - an^{-1/2}e_s) + pl_n(\hat{\theta} + an^{-1/2}e_t) + pl_n(\hat{\theta} - an^{-1/2}e_t) \\ - 2pl_n(\hat{\theta}) - pl_n(\hat{\theta} + an^{-1/2}e_{st}) - pl_n(\hat{\theta} - an^{-1/2}e_{st})\Big\}.$$

Inverting the estimate of I_e gives a consistent estimate of the asymptotic variance of $\hat{\theta}$.

7.2 SIMULATION EXPERIMENTS

Our simulation experiments are conducted under the two models considered in Chapter 6: a normal linear model and a Poisson regression model. Let us start with the linear model. Data are generated according to the following mechanism:

$$Z \sim \text{Beta}\,(\alpha, 1), \tag{7.5}$$

$$Y|Z = z \sim N(\beta_0 + \beta_1 z, \sigma^2), \tag{7.6}$$

$$\text{logit}\{\pi(y)\} = y + \gamma, \tag{7.7}$$

where $\alpha \in \{0.5, 1, 2\}$, $\beta_0 = 0$, $\beta_1 \in \{0, 5\}$, $\sigma^2 = 1$, and γ is adjusted to bring the overall missing proportion $\kappa := 1 - E(R)$ to the desired level (0.2 or 0.5). A *sample* consists of $n = 100$ or 200 independent copies of (Z, Y, R). For each sample size, 1000 *replicates* are generated under each of the 12 *scenarios* (combinations of parameter values).

Given a sample, $\theta = (\beta_0, \beta_1, \sigma^2)$ is estimated using the following five methods. FD (full data) is the usual least squares procedure applied to $\{Z_i, Y_i) : i = 1, \ldots, n\}$ (as if they were all observed). This is

not a competitor method for missing covariates. Rather, it serves as an indicator for the total amount of information about θ contained in the data generated. CC (complete case) is the least squares procedure applied to $\{(Z_i, Y_i) : R_i = 1\}$ (as if they were the original sample). Under an outcome-dependent selection mechanism, this approach is invalid (see Chapter 1). Yet it is still commonly used in practice, regardless of the selection mechanism. We include it in our simulation studies as an illustration of the potential bias and loss of efficiency and also as a source of initial parameter values for the iterative procedures. ML0 is the standard maximum likelihood procedure under the parametric model defined by (7.5) and (7.6). The relative (in)efficiency of ML0 to FD indicates the amount of information lost due to missing values of Z, with G known up to a finite-dimensional paramter. On the other hand, the Fisher information for θ in this model (or any other correct parametric model) is larger (in the sense of nonnegative-definiteness) than the efficient Fisher information for θ in the semiparametric model where G is left unspecified. Therefore ML0 is expected to be more efficient than our (or any other) semiparametric method. Of interest to us is the amount of efficiency gain that comes with a detailed knowledge of the covariate distribution. In practice, it is often difficult to specify a parametric model that is (nearly) correct. In the present setting, a data analyst without sufficient information about G might simply specify a normal model:

$$Z \sim N(\nu, \tau^2), \tag{7.8}$$

which may be called common practice. Denote by ML1 the maximum likelihood procedure under (7.6) and (7.8). We would like to quantify the bias of ML1 due to model misspecification and hence the robustness we have achieved by sparing a parametric specification of G. Lastly, SPML is the proposed semiparametric maximum likelihood method.

ML0 and ML1 are both (conveniently) implemented using an EM algorithm similar to the one described in the preceding subsection. It appears that, even for ML0 and ML1, the EM algorithm is more stable than a quasi-newton algorithm where variable scaling can be a serious problem. We did not investigate this issue further because it is not the focus of this dissertation. Regardless of the algorithm chosen, the implementation of ML0 requires numerical integration, which adds to the computational burden and introduces some arbitrariness. This is, in fact, a general problem with parametric modeling, ML1 being an exception. In contrast, an EM iteration for SPML involves only well-defined summations (and matrix inversion, which is common to all five methods here).

Each method gives for each regression parameter a point estimate, a standard error (standard deviation estimate) and a Wald confidence interval. The only exception here is that, under the least squares approach, inference about σ^2 is based on a chi-squared distribution and does not involve variance estimation. Empirical bias and standard deviation (SD) of a point estimate are calculated using knowledge of the true parameter value and standard formulas applied to the different replicates. Standard errors (SE) are averaged across replicates and compared with the empirical standard deviation. Empirical coverage probabilities (CP) are calculated for (intended) 95% confidence intervals.

Tables 7.1–7.4 summarize numerical results obtained under different scenarios (described earlier) at $n = 100, 200$. In all scenarios studied here, CC is associated with a large bias. In the presence of a strong regression relationship ($\beta_1 = 5$), it also tends to have a large standard deviation. In contrast, all three methods (ML0, ML1, SPML) that explicitly adjust for missing data generally perform better, at least in terms of bias. These observations highlight the relevance of missing data as a methodological issue. We now turn to the comparison of ML0, ML1 and SPML, with ML0 being an ideal that cannot be achieved (without a good knowledge of G). It appears that, under weak regression ($\beta_1 = 0$), the three methods are nearly equivalent in terms of the few criteria considered here. In that case, it does not seem to matter how to deal with the covariate distribution—parametrically or nonparametrically, correctly or incorrectly—as long as we do deal with it. In the case of strong regression ($\beta_1 = 5$), however, the ML1 estimates can be seen to carry a significant bias. The magnitude of this bias depends on the true distribution of Z, which is determined by α. In the current situation, the normal model (7.8) seems to approximate the uniform distribution ($\alpha = 1$) better than the skewed ones ($\alpha = 0.5, 2$) in that the bias is smaller in the former case. For each fixed α, increasing the missing proportion κ (from 0.2 to 0.5) magnifies the misspecification bias of ML1. In fact, increasing κ and/or β_1 also has the effect of setting a higher sample size requirement for our asymptotic results about SPML to take effect. Indeed, for each fixed n, one can make ML1 and SPML perform arbitrarily poorly by choosing large values of κ and β_1. Note, for example, the biases of ML1 and SPML in the scenario where $\kappa = 0.5$, $\beta_1 = 5$ and $\alpha = 2$, at a sample size of $n = 100$. On the other hand, in each fixed scenario, the bias of SPML eventually vanishes with increasing n, whereas that of ML1 does not. In the same scenario as described above, but at $n = 200$, ML1 remains severely biased while SPML becomes much less so. The (in)efficiency of SPML relative to ML0 quantifies the statistical buying power of an accurate knowledge of G in the presence of missing values of Z. The parameters κ and β_1 are again important factors in this assessment: the larger they are, the less efficient SPML is relative to ML0. In most cases, the SPML standard errors estimate the true standard deviations reasonably well and the associated confidence intervals enjoy good coverage probabilities.

The simulation experiments for Poisson regression are conducted in a similar fashion and yield similar results. Data are generated according (7.5), (7.7) and, of course, a Poisson regression model:

$$Y|Z = z \sim \text{POI}\,(\exp(\beta_0 + \beta_1 z)), \tag{7.9}$$

where $\beta_0 = 0$ and $\beta_1 \in \{0, 2\}$. Again, 1000 replicates are generated in each scenario at each sample size. Here FD and CC refer to the standard maximum likelihood procedure applied to $\{(Z_i, Y_i) : i = 1, \ldots, n\}$ and $\{(Z_i, Y_i) : R_i = 1\}$ respectively. ML0 is the maximum likelihood method under (7.5) and (7.9). ML1 is the maximum likelihood method under (7.8) and (7.9). SPML is the proposed method. FD and CC are implemented using a Newton-Raphson algorithm. The other three methods are doubly iterative, with an outer EM loop and an inner Newton-Raphson loop. ML0 and ML1 also involve numerical integration whereas SPML does not. Numerical results are reported in Tables 7.5–7.8. All of the qualitative remarks in the above paragraph remain valid here, except that the bias of ML1 now seems to increase with α.

Table 7.1: Linear Regression with $n = 100$ and $\kappa = 0.2$

Scenario		Method	Bias (×1000)			SD (×1000)			SE (×1000)			CP (×100)		
β_1	α		β_0	β_1	σ^2	β_0	β_1	σ^2	β_0	β_1	σ^2	β_0	β_1	σ^2
0	.5	FD	6	−18	−4	151	342	138	150	337		94	95	96
		CC	179	−23	−113	154	346	137	158	356		80	96	92
		ML0	12	−32	−27	164	388	134	164	391	138	94	95	94
		ML1	10	−25	−27	163	389	135	164	395	138	95	95	94
		SPML	10	−26	−27	163	388	134	165	395	138	95	95	94
	1	FD	14	−28	−1	203	338	149	201	349		95	96	94
		CC	181	−18	−113	218	367	152	211	368		84	95	87
		ML0	17	−29	−25	234	408	146	227	408	139	94	94	91
		ML1	13	−20	−25	235	411	146	227	409	139	94	94	91
		SPML	13	−20	−25	235	411	146	227	409	139	94	94	91
	2	FD	−2	2	−2	307	439	139	302	427		94	94	96
		CC	170	−2	−111	327	469	138	319	451		90	94	90
		ML0	6	−7	−26	364	528	136	347	498	139	93	93	93
		ML1	3	−2	−26	364	528	136	347	499	139	93	93	92
		SPML	2	−1	−26	365	529	136	347	499	139	93	93	92
5	.5	FD	−3	−1	6	155	357	141	151	339		94	94	95
		CC	284	−358	−83	171	374	143	173	356		63	81	93
		ML0	−2	2	−19	159	364	145	153	342	148	94	93	93
		ML1	107	−200	−75	153	344	145	159	333	149	90	89	87
		SPML	−2	2	−19	160	364	146	153	341	148	94	93	92
	1	FD	−11	18	0	201	347	140	201	348		95	95	95
		CC	375	−421	−82	241	391	140	245	393		66	80	94
		ML0	−13	21	−22	209	353	149	211	359	154	94	95	93
		ML1	37	−65	−34	213	352	154	222	366	159	95	96	92
		SPML	−15	25	−21	212	356	150	212	360	155	94	95	94
	2	FD	−7	14	4	312	436	141	303	429		95	95	95
		CC	497	−493	−79	367	500	146	383	515		73	84	93
		ML0	−10	15	−16	340	472	162	330	460	161	94	94	92
		ML1	−113	131	26	373	507	174	348	484	172	93	94	95
		SPML	−29	40	−9	363	497	165	334	465	162	94	94	93

Table 7.2: Linear Regression with $n = 100$ and $\kappa = 0.5$

Scenario		Method	Bias (×1000)			SD (×1000)			SE (×1000)			CP (×100)		
β_1	α		β_0	β_1	σ^2	β_0	β_1	σ^2	β_0	β_1	σ^2	β_0	β_1	σ^2
0	.5	FD	−4	5	−2	149	339	134	150	337		95	94	96
		CC	416	8	−170	197	457	166	196	443		43	94	89
		ML0	24	−52	−31	203	522	141	194	495	141	93	92	93
		ML1	8	7	−38	207	552	133	199	520	141	92	91	93
		SPML	7	8	−37	209	555	133	200	521	141	93	91	93
	1	FD	1	5	−3	202	347	144	201	347		94	95	95
		CC	408	7	−172	265	470	173	261	451		64	94	86
		ML0	18	−5	−34	294	547	142	278	518	141	93	93	91
		ML1	9	11	−39	297	560	143	284	531	141	92	92	91
		SPML	10	10	−38	298	562	143	285	532	140	92	92	91
	2	FD	−4	8	−4	302	428	140	300	425		94	94	96
		CC	410	7	−171	406	572	168	392	555		80	94	88
		ML0	18	−7	−26	461	665	137	447	652	143	93	93	94
		ML1	12	3	−39	475	695	140	446	651	141	92	93	92
		SPML	11	4	−39	473	694	140	445	651	141	92	93	92
5	.5	FD	1	1	−7	141	340	142	149	338		96	94	95
		CC	644	−683	−132	257	473	172	252	445		29	65	92
		ML0	−4	14	−30	169	370	165	175	380	170	96	95	92
		ML1	178	−353	−100	200	375	182	215	385	187	86	86	87
		SPML	−8	13	−24	176	382	172	177	374	170	95	94	93
	1	FD	7	−19	4	203	358	149	201	348		95	95	94
		CC	795	−752	−134	383	551	189	371	536		42	68	89
		ML0	−12	7	−28	260	402	193	260	418	192	95	96	91
		ML1	−70	52	−4	349	484	239	314	464	219	93	94	91
		SPML	−47	54	−9	332	482	217	272	431	191	89	92	90
	2	FD	−14	15	5	304	428	149	303	429		95	95	94
		CC	950	−793	−144	584	736	182	576	730		61	79	89
		ML0	−31	44	−20	414	562	219	422	565	212	96	95	90
		ML1	−438	502	89	565	710	259	482	652	242	85	87	95
		SPML	−177	197	33	568	713	240	438	598	210	86	89	91

Table 7.3: Linear Regression with $n = 200$ and $\kappa = 0.2$

Scenario		Method	Bias (×1000)			SD (×1000)			SE (×1000)			CP (×100)		
β_1	α		β_0	β_1	σ^2	β_0	β_1	σ^2	β_0	β_1	σ^2	β_0	β_1	σ^2
0	.5	FD	0	2	1	106	239	100	106	238		95	95	95
		CC	172	2	−111	111	252	99	112	251		67	95	84
		ML0	2	−3	−11	116	282	99	116	279	99	95	94	94
		ML1	1	1	−11	117	284	99	117	281	99	95	94	94
		SPML	1	1	−11	117	284	99	117	281	99	95	94	94
	1	FD	−2	3	−1	139	241	102	142	246		96	94	95
		CC	168	4	−111	149	257	99	149	259		79	95	83
		ML0	0	0	−13	158	287	101	160	288	99	96	95	93
		ML1	−2	4	−13	159	289	101	161	289	99	96	95	93
		SPML	−2	4	−13	159	289	101	161	289	99	96	95	93
	2	FD	9	−11	0	211	304	98	213	302		95	95	95
		CC	182	−13	−109	230	325	98	225	319		86	94	83
		ML0	15	−18	−12	252	364	97	247	354	99	94	94	93
		ML1	12	−13	−12	254	366	97	247	355	99	95	94	93
		SPML	12	−13	−12	253	365	97	247	355	99	95	94	93
5	.5	FD	−4	−2	−1	111	242	101	106	238		94	95	96
		CC	284	−364	−87	122	255	102	122	250		34	68	88
		ML0	−3	−2	−11	114	246	105	109	243	106	93	94	95
		ML1	106	−204	−67	110	233	106	112	236	106	85	87	87
		SPML	−3	−2	−11	114	246	106	109	241	106	93	94	95
	1	FD	−1	5	4	139	246	103	142	246		95	95	94
		CC	380	−427	−77	174	289	104	173	278		40	66	88
		ML0	−4	9	−7	146	254	111	151	256	111	95	95	93
		ML1	45	−76	−17	151	256	114	157	260	114	94	94	94
		SPML	−5	11	−6	150	258	112	150	255	111	95	95	93
	2	FD	2	−2	1	212	296	98	212	300		94	94	96
		CC	513	−514	−91	259	349	105	266	358		51	70	88
		ML0	−5	10	−13	229	317	115	234	326	114	95	95	94
		ML1	−102	117	25	249	339	122	242	337	122	93	94	95
		SPML	−13	20	−11	240	331	116	234	324	115	94	94	94

Table 7.4: Linear Regression with $n = 200$ and $\kappa = 0.5$

Scenario		Method	Bias (×1000)			SD (×1000)			SE (×1000)			CP (×100)		
β_1	α		β_0	β_1	σ^2	β_0	β_1	σ^2	β_0	β_1	σ^2	β_0	β_1	σ^2
0	.5	FD	2	−3	1	108	234	101	106	239		95	95	94
		CC	419	−13	−168	136	309	123	137	309		13	95	76
		ML0	3	−5	−14	140	355	98	137	350	100	94	94	93
		ML1	11	−15	−17	142	371	101	141	368	100	94	94	93
		SPML	11	−15	−17	143	372	101	142	368	100	94	94	93
	1	FD	−2	−3	2	141	249	101	142	246		96	94	95
		CC	416	−5	−167	190	331	116	184	319		38	94	79
		ML0	22	−25	−21	211	396	99	197	367	99	92	92	93
		ML1	6	−7	−16	212	401	99	202	379	100	93	93	94
		SPML	6	−7	−16	212	401	99	202	379	100	93	93	94
	2	FD	−5	9	3	222	316	101	213	301		94	93	95
		CC	416	−1	−170	277	388	122	276	390		66	95	76
		ML0	20	−20	−25	328	483	100	310	454	99	93	92	93
		ML1	7	−3	−14	326	474	100	320	467	100	94	94	94
		SPML	7	−3	−14	325	474	100	319	467	100	94	94	94
5	.5	FD	−3	6	−6	106	235	102	106	237		94	95	94
		CC	631	−657	−127	170	299	123	178	312		7	42	87
		ML0	−4	5	−10	119	258	126	129	284	124	97	97	94
		ML1	171	−339	−81	138	253	132	153	272	134	81	77	86
		SPML	−8	22	−20	123	258	121	125	261	120	95	95	93
	1	FD	−4	6	0	146	249	102	142	246		95	94	95
		CC	777	−718	−130	269	396	123	262	379		17	52	84
		ML0	−12	13	−12	185	289	139	193	311	139	96	97	93
		ML1	−81	72	14	239	341	158	223	329	157	93	95	94
		SPML	−32	44	−7	215	328	142	197	308	138	94	94	93
	2	FD	−1	−3	2	212	296	101	213	302		95	95	94
		CC	948	−789	−146	398	503	127	399	506		34	65	81
		ML0	−2	5	−17	263	360	154	322	433	151	97	98	93
		ML1	−403	463	92	383	483	178	331	448	170	79	82	95
		SPML	−82	88	16	384	487	160	312	418	149	89	92	93

Table 7.5: Poisson Regression with $n = 100$ and $\kappa = 0.2$

Scenario		Method	Bias (×1000)		SD (×1000)		SE (×1000)		CP (×100)	
β_1	α		β_0	β_1	β_0	β_1	β_0	β_1	β_0	β_1
0	.5	FD	−12	5	156	343	152	340	95	95
		CC	121	2	159	346	159	357	86	97
		ML0	−10	−3	165	372	160	370	95	95
		ML1	−10	1	167	377	160	371	94	95
		SPML	−10	2	166	376	160	372	94	95
	1	FD	−12	4	203	355	203	352	95	95
		CC	123	−5	204	363	213	369	90	95
		ML0	−4	−12	217	392	217	383	95	94
		ML1	−6	−7	218	394	216	383	95	94
		SPML	−6	−7	217	393	217	384	95	94
	2	FD	−30	24	305	433	307	433	95	95
		CC	102	18	306	434	323	456	95	96
		ML0	−23	14	324	467	332	472	96	95
		ML1	−27	19	325	468	329	468	95	95
		SPML	−26	19	325	467	332	473	96	95
2	.5	FD	−2	−2	128	208	128	208	94	95
		CC	196	−218	129	211	137	219	69	85
		ML0	−2	−3	130	214	130	214	95	95
		ML1	0	−13	131	214	133	218	95	95
		SPML	−2	−2	130	214	130	213	94	95
	1	FD	0	−1	152	216	152	215	95	95
		CC	258	−273	149	211	166	231	65	79
		ML0	−2	0	154	221	156	222	95	95
		ML1	−11	9	158	224	162	231	96	96
		SPML	−3	1	155	221	156	222	95	95
	2	FD	−5	0	196	253	195	249	95	94
		CC	349	−357	197	254	220	276	65	76
		ML0	−6	1	204	263	205	263	95	95
		ML1	−42	42	216	276	227	291	95	95
		SPML	−8	3	206	266	205	263	95	95

Table 7.6: Poisson Regression with $n = 100$ and $\kappa = 0.5$

Scenario		Method	Bias (×1000)		SD (×1000)		SE (×1000)		CP (×100)	
β_1	α		β_0	β_1	β_0	β_1	β_0	β_1	β_0	β_1
0	.5	FD	−2	−32	154	340	152	343	94	95
		CC	331	−27	175	399	184	417	53	97
		ML0	7	−57	186	464	178	439	93	94
		ML1	3	−36	192	485	179	447	94	93
		SPML	2	−33	190	481	182	455	93	93
	1	FD	−12	3	210	376	202	351	95	95
		CC	328	2	233	407	245	425	70	97
		ML0	−2	−14	265	487	250	456	94	94
		ML1	−9	3	271	499	251	459	95	95
		SPML	−9	3	270	496	255	466	94	94
	2	FD	−14	10	309	438	304	430	95	95
		CC	328	2	337	475	370	523	85	98
		ML0	7	−18	396	570	393	567	95	95
		ML1	−9	6	400	578	387	560	95	95
		SPML	−7	4	398	575	396	573	95	95
2	.5	FD	−2	−4	126	208	128	208	95	96
		CC	492	−512	147	225	170	254	16	46
		ML0	−3	−9	135	226	141	232	96	97
		ML1	−6	−15	144	230	161	259	97	98
		SPML	−3	−7	136	225	138	225	95	96
	1	FD	0	−3	148	215	151	215	95	95
		CC	607	−594	184	251	218	288	17	45
		ML0	−8	3	168	240	175	247	96	96
		ML1	−75	88	193	266	221	310	96	95
		SPML	−10	8	173	245	174	245	95	95
	2	FD	−4	1	200	254	196	249	95	95
		CC	767	−715	262	321	306	369	25	48
		ML0	−7	−2	240	307	237	301	95	95
		ML1	−195	223	308	381	318	407	69	69
		SPML	−42	42	279	348	245	311	92	93

Table 7.7: Poisson Regression with $n = 200$ and $\kappa = 0.2$

Scenario		Method	Bias (×1000)		SD (×1000)		SE (×1000)		CP (×100)	
β_1	α		β_0	β_1	β_0	β_1	β_0	β_1	β_0	β_1
0	.5	FD	−2	−7	105	240	106	239	96	95
		CC	129	−11	106	237	112	251	78	96
		ML0	1	−14	110	257	112	260	95	95
		ML1	1	−13	111	258	112	260	95	95
		SPML	0	−12	111	257	112	261	95	95
	1	FD	−3	−7	144	253	142	246	94	95
		CC	127	−9	145	252	149	259	86	95
		ML0	−1	−11	155	273	152	269	94	95
		ML1	−2	−10	155	274	151	268	94	94
		SPML	−2	−10	155	273	152	270	94	94
	2	FD	−24	28	217	304	215	304	95	95
		CC	110	21	210	293	226	319	92	96
		ML0	−19	19	225	317	233	332	95	96
		ML1	−21	23	226	318	231	330	95	95
		SPML	−21	22	226	318	233	332	95	95
2	.5	FD	−2	−5	92	149	90	146	95	95
		CC	197	−221	91	148	96	153	46	70
		ML0	−2	−6	93	150	92	150	95	96
		ML1	1	−17	94	151	95	154	96	96
		SPML	−1	−6	93	150	91	149	95	96
	1	FD	−4	7	106	150	107	151	95	96
		CC	253	−264	111	156	117	163	41	61
		ML0	−6	9	109	156	110	157	95	96
		ML1	−15	17	112	158	117	166	96	96
		SPML	−6	9	109	156	110	156	95	96
	2	FD	−3	4	139	176	138	175	94	94
		CC	352	−356	145	182	155	194	36	54
		ML0	−3	2	146	187	145	185	94	94
		ML1	−38	43	155	195	178	229	95	95
		SPML	−3	3	147	187	144	185	94	94

Table 7.8: Poisson Regression with $n = 200$ and $\kappa = 0.5$

Scenario		Method	Bias (×1000)		SD (×1000)		SE (×1000)		CP (×100)	
β_1	α		β_0	β_1	β_0	β_1	β_0	β_1	β_0	β_1
0	.5	FD	−3	−4	107	230	107	239	95	96
		CC	338	−11	114	255	127	287	23	98
		ML0	5	−25	124	300	124	307	94	96
		ML1	2	−13	127	313	126	312	95	96
		SPML	1	−12	127	311	127	317	95	95
	1	FD	−6	−2	147	254	142	247	95	94
		CC	334	−5	167	284	172	298	48	97
		ML0	4	−20	188	341	176	322	93	94
		ML1	−2	−7	190	346	178	326	94	95
		SPML	−2	−7	190	344	179	328	93	94
	2	FD	−9	8	214	297	214	302	95	96
		CC	329	10	229	321	257	364	76	98
		ML0	−3	1	277	396	276	399	95	95
		ML1	−13	15	279	400	272	394	95	95
		SPML	−11	14	278	398	277	402	95	94
2	.5	FD	−11	13	90	142	90	147	96	96
		CC	480	−489	101	151	119	179	2	17
		ML0	−12	13	96	153	102	168	97	97
		ML1	−14	4	102	156	126	203	99	99
		SPML	−12	15	97	154	97	158	95	96
	1	FD	−2	3	107	151	107	151	95	95
		CC	608	−594	130	179	153	202	2	14
		ML0	−3	2	121	170	126	178	96	96
		ML1	−68	83	139	189	187	263	95	95
		SPML	−5	6	126	175	122	172	95	95
	2	FD	−1	−2	141	178	138	175	95	94
		CC	769	−719	181	221	213	257	5	18
		ML0	−5	−1	167	209	171	217	97	96
		ML1	−181	211	210	257	162	206	33	32
		SPML	−17	15	183	227	171	217	95	94

8.0 APPLICATIONS IN RELATED PROBLEMS

Discussed in this chapter are potential applications of the proposed method in statistical problems that, strictly speaking, do not fall into the category of parametric regression with missing covariates. Interestingly, such a problem can often be modified into one to which our method is applicable, under assumptions whose plausibility depends, as always, on the scientific problem at hand. If this is the case, applying the proposed method often yields a novel approach to the original problem. Two examples are given below.

8.1 MEASUREMENT ERROR MODELS WITH VALIDATION DATA

As before, let Y given $X = (W, Z)$ follow a parametric regression model, with W finitely discrete, (W, Y) always observed and Z possibly missing. Of primary interest is the regression parameter vector. Suppose we also (always) observe a variable V, often called a surrogate variable, that is known to be related to Z but not necessarily to (W, Y). For example, V may be an inexpensive (and imprecise) measurement of Z. Let $R = 1$ if Z is observed; 0 otherwise. Assume that Z is missing at random, that is,

$$E(R|X, Y, V) = E(R|W, Y, V), \qquad \text{almost surely.}$$

This is weaker than assumption (1.7). Hence our method is not immediately applicable. Even if it is, it may be inefficient without making use of V. We now consider how V can be incorporated into the regression model while fulfilling our key requirements.

Depending on the situation, it may be appropriate to postulate a parametric model for the conditional distribution of V given Z. If, in addition, V is conditionally independent of (W, Y) given Z, then we have a parametric regression model for $\tilde{Y} := (Y, V)$ given X:

$$[\tilde{Y}|X] = [Y|X][V|X, Y] = [Y|X][V|Z].$$

This permits application of the proposed method, provided the regularity conditions can be verified. A parametric model for V given Z may be readily available if V is a physical measurement of Z. In the simplest case, it may be reasonable to assume that $V = Z + \varepsilon$ where $\varepsilon \sim \mathrm{N}(0, \varsigma^2)$ is the measurement error. If V is a measurement of Z, then the above conditional independence requirement simply says that the measurement process is independent of the other regression variables, which may be the case in many applications.

At times the assumptions made in the above paragraph may not be realistic. But the proposed method may still be applicable, via a different route. Assume that V is finitely discrete or can be discretized without much loss of information. Assume also that V is conditionally independent of Y given X. Then V can be "absorbed" into W, written $\tilde{W} = (W, V)$, and the regression model is unchanged:

$$[Y|\tilde{W}, Z] = [Y|W, Z].$$

Note that the conditional independence condition here is weaker than the one in the above paragraph. Moreover, this approach requires no parametric assumptions (in addition to the original regression model).

8.2 AUXILIARY VARIABLES

Suppose now that we would like to estimate the distribution of some variable Y. Actually we may be only interested in some aspects of it such as the mean, variance or certain quantiles. But all of these can be easily obtained from an estimate of the distribution; so let us focus on the latter. Unfortunately Y cannot be ascertained for all subjects in a random sample. What we can observe for all subjects is an auxiliary variable X which forms the basis for the (known or unknown) selection process. With R being the usual selection indicator, this means

$$E(R|X, Y) = E(R|X), \qquad \text{almost surely.} \tag{8.1}$$

Examples of this type of problems include survey nonresponse and imprecise measurement.

When the selection mechanism is known or can be modeled, a simple estimator is the Horvitz-Thompson estimator, weighting each observed value of Y by the inverse of the (estimated) conditional probability of observing Y given X. A more common, and often more efficient, approach is regression imputation, described as follows. In view of (8.1), the conditional distribution of Y given X can be estimated from the complete cases using any standard method. (This is usually based on a parametric or semiparametric regression model. But it could be done nonparametrically with some discretization or smoothing.) With this estimate and the observed value of X, one can then impute the missing value of Y for an incomplete case. See Little and Rubin (2002) for various implementations of this simple idea.

In some applications, it may be more natural to regress X on Y. This would be the case if, for example, X is a physical measurement of Y and the error distribution is well understood. Then the method developed in this dissertation would be applicable. The semiparametric MLE of the distribution of Y is strongly consistent and asymptotically Gaussian in a suitable sense. Aside from modeling considerations, the proposed estimator may also have some efficiency advantage. This is because the regression parameter and the distribution of Y are now estimated simultaneously, rather than sequentially as in regression imputation. We do not yet have results to support this conjecture. Further investigation is warranted.

9.0 DISCUSSION

To help put this work in proper perspective, we now give a brief discussion on the pros and cons of the proposed method relative to the existing methods in the literature. More details about the competing methods can be found in Chapter 1 and, of course, the original papers. The main advantages of the proposed method over the complete-case analysis are consistency under outcome-dependent selection and efficiency under outcome-independent selection. The former is well established, theoretically and numerically. The latter is intuitively clear but has not been studied in great details. It would be interesting to calculate the asymptotic variance of the CC estimator and compare it with that of SPML. The proposed method achieves the semiparametric information bound, which represents a major advantage over the methods based on estimated likelihoods or scores (pseudolikelihood, mean score and pseudoscore). In principle, semiparametric efficiency can also be achieved under the approach of Robins, Hsieh and Newey (1995) (RHN). However, the RHN methodology requires modeling the selection mechanism and is generally difficult to implement. To the best of our knowledge, it has been implemented only when W and Y are both discrete. In contrast, the proposed method leaves the selection mechanism unspecified and is relatively easy to implement, as shown in Chapter 7. Comparing with the parametric modeling approach of Ibrahim, Chen and Lipsitz (1999), the proposed method is more robust in that G is treated nonparametrically.

Semiparametric MLEs have been studied in related but different contexts, such as two-phase sampling (Wild, 1991; Lawless, Kalbfleisch and Wild, 1999; van der Vaart and Wellner, 2001; Braslow, McNeney and Wellner, 2003), mixture models (van der Vaart, 1996; Roeder, Carroll and Lindsay, 1996; Murphy and van der Vaart, 2001) and deconvolution (van der Vaart and Wellner, 1992; van der Vaart, 1994a). In spirit, the proposed method is similar to the method of Lawless, Kalbfleisch and Wild (1999) for two-phase sampling. A defining feature of the two-phase sampling problem is a finite stratification of the sample space which determines the second-phase sampling. This limits the applicability of their method to the general missing covariates problem that we consider. It also makes the analysis here somewhat more complicated.

A serious limitation of the proposed method is the requirement that W be finitely discrete. Although this is often true or can be made true by discretizing, there are certainly situations where W is and should be treated as a continuous variable. In the latter case, it is not clear how to handle G nonparametrically within the maximum likelihood framework. Full robustness with respect to G, though absolutely desirable, is difficult if not impossible to achieve.

BIBLIOGRAPHY

Bickel, P. J., Klaassen, C. A. J., Ritov, Y. and Wellner, J. A. (1993). *Efficient and Adaptive Estimation for Semiparametric Models*. Johns Hopkins University Press, Baltimore.

Breslow, N. E., McNeney, B. and Wellner, J. A. (2003). Large sample theory for semiparametric regression models with two-phase, outcome dependent sampling. *Ann. Statist.*, 31, 1110–1139.

Carroll, R. J. and Wand, M. P. (1991). Semiparametric estimation in logistic measurement error models. *JRSS-B*, 53, 573–585.

Chatterjee, N., Chen, Y. H. and Breslow, N. E. (2003). A pseudoscore estimator for regression problems with two-phase sampling. *JASA*, 98, 158–168.

Chen, H. Y. (2002). Double-semiparametric method for missing covariates in Cox regression models. *JASA*, 97, 565–576.

Cox, D. R. and Hinkley, D. V. (1974). *Theoretical Statistics*. Chapman and Hall, London.

Ibrahim, J. G., Chen, M. H. and Lipsitz, S. R. (1999). Monte Carlo EM for missing covariates in parametric regression models. *Biometrics*, 55, 591–596.

Lawless, J. F., Kalbfleisch, J. D. and Wild, C. J. (1999). Semiparametric methods for response-selective and missing data problems in regression. *JRSS-B*, 61, 413–438.

Little, R. J. A. and Rubin D. B. (2002). *Statistical Analysis with Missing Data*, 2nd ed. Wiley, New York.

Murphy, S. A. and van der Vaart, A. W. (2000). On profile likelihood (with discussion). *JASA*, 95, 449–465.

Murphy, S. A. and van der Vaart, A. W. (2001). Semiparametric mixtures in case-control studies. *J. Mult. Anal.*, 79, 1–32.

Pepe, M. S. and Fleming, T. R. (1991). A nonparametric method for dealing with mismeasured covariate data. *JASA*, 86, 108–113.

Reilly, M. and Pepe, M. S. (1995). A mean score method for missing and auxiliary covariate data in regression models. *Biometrika*, 82, 299–314.

Robins, J. M., Hsieh, F. and Newey, W. (1995). Semiparametric efficient estimation of a conditional density with missing or mismeasured covariates. *JRSS-B*, 57, 409–424.

Roeder, K., Carroll, R. J. and Lindsay, B. G. (1996). A semiparametric mixture approach to case-control studies with errors in covariables. *JASA*, 91, 722–732.

Rubin, D. B. (1976). Inference and missing data. *Biometrika*, 63, 581–592.

Rudin, W. (1973). *Functional Analysis*. McGraw-Hill, New York.

van der Vaart, A. W. (1994a). Maximum likelihood estimation with partially censored data. *Ann. Statist.*, 22, 1896–1916.

van der Vaart, A. W. (1994b). Bracketing smooth functions. *Stochastic Processes and Their Applications*, 52, 93–105.

van der Vaart, A. W. (1996). Efficient maximum likelihood estimation in semiparametric mixture models. *Ann. Statist.*, 24, 862–878.

van der Vaart, A. W. (1998). *Asymptotic Statistics*. Cambridge University Press, New York.

van der Vaart, A. W. and Wellner, J. A. (1992). Existence and consistency of maximum likelihood in upgraded mixture models. *J. Mult. Anal.*, 43, 133–146.

van der Vaart, A. W. and Wellner, J. A. (1996). *Weak Convergence and Empirical Processes with Applications to Statistics*. Springer-Verlag, New York.

van der Vaart, A. W. and Wellner, J. A. (2000). Preservation theorems for Glivenko-Cantelli and uniform Glivenko-Cantelli classes. In *High Dimensional Probability II* (E. Giné, D. M. Mason and J. A. Wellner, eds.). Birkäuser, Boston.

van der Vaart, A. W. and Wellner, J. A. (2001). Consistency of semiparametric maximum likelihood estimators for two-phase sampling. *Canad. J. Statist.*, 29, 269–288.

Wild, C. J. (1991). Fitting prospective regression models to case-control data. *Biometrika*, 78, 705–717.