Missing Data: Theory & Methods

Chapter 6

The Estimating Equation Approach

Contents

Basics of M Estimation

- Statistical analysis in the traditional parametric paradigm has been focused on the MLE, which requires specifying the distribution of the full data up to a finite-dimensional parameter.
- However, when the data structure is complex, e.g., multivariate, longitudinal, etc., it is difficult to correctly specify the joint distribution of the data.
- It is also sometimes unnecessary and/or uneconomical to fully model the whole distribution if we are only interested in some aspects of the distribution, e.g., the mean.
- As an alternative to MLE, **estimating equations** often serve us in terms of providing robust inference by obviating unnecessary and possibly faulty distributional assumptions.

Basics of M Estimation

- First, let's review the basics of MLE and its asymptotic properties, i.e., distribution of MLE when the sample size is large.
- As before, denote $p_Y(Y;\theta)$ as the (full-data) density function indexed by a Euclidean parameter θ, and write $l(Y;\theta) = \log p_Y(Y;\theta)$.
- For a random sample of n observations, the MLE is defined by

$$\begin{aligned}\widehat{\theta} &= \arg\max_{\theta} \sum_{i=1}^{n} l(Y_i;\theta) \\ &= \mathrm{Root}_{\theta} \sum_{i=1}^{n} \dot{l}(Y_i;\theta),\end{aligned}$$

where

$$\dot{l}(Y;\theta) = \frac{\partial}{\partial\theta} l(Y;\theta)$$

is the score function (for one observation).

Basics of M Estimation

Important notation and definitions

- To fix ideas, we need to be a bit more formal with the following notation that will be used extensively.
- First, from here on we will make a strict distinction between θ_0, the true value of the parameter, and θ, a generic symbol for the parameter. For example, the score function $\dot{l}(Y;\theta)$ is a (random) function of θ and $\dot{l}(Y;\theta_0)$, i.e., the score function evaluated at θ_0, is a random vector.
- When using $Eg(Y)$, the expectation is taken under the true distribution of Y, that is, under the density indexed by θ_0, i.e.,

$$Eg(Y) = \int g(y) p_Y(y;\theta_0) d\nu(y).$$

Likewise, if the function $g(Y;\theta)$ is indexed by θ,

$$Eg(Y;\theta) = \int g(y;\theta) p_Y(y;\theta_0) d\nu(y).$$

Basics of M Estimation

Important notation and definitions

- Recall that a sequence of random variables A_n converges in probability to a constant a, i.e.,

$$A_n \to_p a,$$

  if $\Pr(|A_n - a| > \epsilon) \to 0$ for every ϵ. In other words, A_n is eventually concentrated in arbitrarily small neighborhoods of a.
- An estimator $\widehat{\theta}_n$ is consistent, or asymptotically unbiased, if $\widehat{\theta}_n \to_p \theta_0$. Consistency is the basic requirement for validity of an estimator.
- Asymptotic order:
  - $A_n = o_p(B_n)$ if $A_n/B_n \to_p 0$.
  - $A_n = O_p(B_n)$ if A_n/B_n is asymptotically bounded (in probability), i.e., for every $\epsilon > 0$, there exists M such that $\liminf_n \Pr(|A_n/B_n| \leq M) > 1 - \epsilon$.

Basics of M Estimation

Important notation and definitions

- So, in particular, $\widehat{\theta}_n \to_p \theta_0$ is the same as $\widehat{\theta}_n = \theta_0 + o_p(1)$.
- We say A_n converges in distribution (or weakly) to a random variable A, denoted as

$$A_n \rightsquigarrow A,$$

  if the distribution function of A_n converges point-wise to that of A (at the continuity points of the latter). Thus, if the weak convergence holds, the distribution of A can be used to approximate that of A_n.
- In estimation of θ_0, if

$$\sqrt{n}(\widehat{\theta}_n - \theta_0) \rightsquigarrow N(0, \sigma^2), \tag{6.1}$$

  then, the distribution of $\widehat{\theta}_n$ can be approximated by

$$N(\theta_0, n^{-1}\sigma^2),$$

  where σ^2 is typically replaced by some estimator $\widehat{\sigma}_n^2$.

Basics of M Estimation

Linearization and influence functions

- The Slutsky's lemma says that if $A_n = B_n + o_p(1)$ and $B_n \rightsquigarrow A$, then $A_n \rightsquigarrow A$.
- This lemma is often exploited to establish the asymptotic normality (6.1) via an intermediate **linearization step**. Specifically, one can show that

$$\begin{aligned} \sqrt{n}(\widehat{\theta}_n - \theta_0) &= n^{-1/2} \sum_{i=1}^{n} \Psi(Y_i; \theta_0) + o_p(1) \\ &\rightsquigarrow N(0, E\Psi(Y; \theta_0)^{\otimes 2}), \end{aligned} \tag{6.2}$$

  where $\Psi(Y; \theta_0)$ is some mean-zero function and the second step is by the central limit theorem.
- We call $\Psi(Y; \theta_0)$ the **influence function** for $\widehat{\theta}_n$.

Linearization and influence functions

- Another advantage of taking the linearization step is that it provides a natural estimator for the asymptotic variance, namely

$$n^{-1}\sum_{i=1}^{n}\Psi(Y_i;\widehat{\theta}_0)^{\otimes 2}.$$

- When $\widehat{\theta}_n$ can be linearized as in (6.2), we say $\widehat{\theta}_n$ is asymptotically linear.
- The MLE in a regular parametric model can be shown to be asymptotically linear and its influence functions can be exhibited.

Asymptotics for MLE

- Consider the function $g(D;\theta) \equiv n^{-1}\sum_{i=1}^{n} \dot{l}(Y_i;\theta)$ as a (data-dependent) function of θ. We use Taylor expansion of $g(D;\widehat{\theta}_n)$ around θ_0:

$$n^{-1}\sum_{i=1}^{n} \dot{l}(Y_i;\widehat{\theta}_n) = n^{-1}\sum_{i=1}^{n} \dot{l}(Y_i;\theta_0) + (\widehat{\theta}_n - \theta_0) n^{-1}\sum_{i=1}^{n} \ddot{l}(Y_i;\theta_0) + o_p(\widehat{\theta}_n - \theta_0).$$

- By definition of MLE, the left hand side is 0. So, after algebraic manipulation,

$$\begin{aligned}\sqrt{n}(\widehat{\theta}_n - \theta_0) &= -\left(n^{-1}\sum_{i=1}^{n} \ddot{l}(Y_i;\theta_0)\right)^{-1} \frac{1}{\sqrt{n}}\sum_{i=1}^{n} \dot{l}(Y_i;\theta_0) + o_p\left(\sqrt{n}(\widehat{\theta}_n - \theta_0)\right) \\ &= -\left\{E\ddot{l}(Y;\theta_0)\right\}^{-1} \frac{1}{\sqrt{n}}\sum_{i=1}^{n} \dot{l}(Y_i;\theta_0) + o_p\left(\left|\sqrt{n}(\widehat{\theta}_n - \theta_0)\right| + 1\right).\end{aligned} \tag{6.3}$$

Basics of M Estimation

Asymptotics for MLE

- Note that $E\dot{l}(Y;\theta_0) = 0$ and that

$$-E\ddot{l}(Y;\theta_0) = E\dot{l}(Y;\theta_0)^{\otimes 2}.$$

- Hence,

$$\sqrt{n}(\widehat{\theta}_n - \theta_0) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \widetilde{l}(Y_i;\theta_0) + o_p(1),$$

  where

$$\widetilde{l}(Y;\theta_0) = \left(E\dot{l}(Y;\theta_0)^{\otimes 2}\right)^{-1} \dot{l}(Y;\theta_0)$$

  is the **efficient influence function** for θ at θ_0.

- The reason $\widetilde{l}(Y;\theta_0)$ is called efficient influence function is because a regular asymptotically linear estimators for θ_0 is statistically efficient, i.e., having the smallest variance, if and only if its influence function is $\widetilde{l}(Y;\theta_0)$.

Asymptotics for MLE

- Denote

$$\mathcal{I}(\theta_0) = E\dot{l}(Y;\theta_0)^{\otimes 2}$$

  as the information of θ based on one observation.

- Then the asymptotic variance for the MLE $\widehat{\theta}_n$ is

$$\mathcal{I}(\theta_0)^{-1} E\dot{l}(Y;\theta_0)^{\otimes 2} \mathcal{I}(\theta_0)^{-1} = \mathcal{I}(\theta_0)^{-1},$$

  which can be estimated by

$$\left(n^{-1} \sum_{i=1}^{n} \dot{l}(Y;\widehat{\theta}_n)^{\otimes 2} \right)^{-1} \text{ or } -\left(n^{-1} \sum_{i=1}^{n} \ddot{l}(Y;\widehat{\theta}_n) \right)^{-1}$$

Asymptotics for MLE (Some technical aspects)

- In the second step we have replaced
  $-\left(n^{-1}\sum_{i=1}^{n}\ddot{l}(Y_i;\theta_0)\right)^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\dot{l}(Y_i;\theta_0)$ by
  $-\left\{E\ddot{l}(Y;\theta_0)\right\}^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\dot{l}(Y_i;\theta_0)+o_p(1)$.
- Denote this as replacing A_nB_n by $AB_n+o_p(1)$.
- This replacement is justified because by the law of large numbers (and continuous mapping theorem), $A_n=A+o_p(1)$, and also $B_n=O_p(1)$ because it is weakly convergent to a multivariate normal. So

$$A_nB_n=(A+o_p(1))B_n=AB_n+o_p(1)O_p(1)=AB_n+o_p(1).$$

- In general, if B_n is not asymptotically bounded (in probability), then,

$$A_nB_n\neq AB_n+o_p(1),$$

  even if the difference between A_n and A is $o_p(1)$.

Basics of M Estimation

M-Estimation by estimating equations

- We call a function $m(Y;\theta)$ a valid **estimating function** if

$$Em(Y;\theta_0) = 0.$$

- Given an estimating function $m(Y;\theta)$, we can derive an M estimator $\widehat{\theta}_n$ based on solving

$$n^{-1}\sum_{i=1}^{n} m(Y_i;\widehat{\theta}_n) = 0,$$

that is,

$$\widehat{\theta}_n = \text{Root}_\theta \left\{ n^{-1}\sum_{i=1}^{n} m(Y_i;\theta) \right\}.$$

- Clearly, the MLE corresponds to the M estimator with $m(Y;\theta) = \dot{l}(Y;\theta)$.

Basics of M Estimation

M-Estimation by estimating equations

- Valid estimating functions (i.e., mean zero at true value of parameter) generally lead to valid M estimators. The reasoning is as follows.
- By the law of large numbers, we have that, for all θ,

$$n^{-1}\sum_{i=1}^{n} m(Y_i;\theta) \rightarrow_p Em(Y;\theta).$$

- Therefore one expects that

$$\widehat{\theta}_n = \mathrm{Root}_\theta \left\{ n^{-1}\sum_{i=1}^{n} m(Y_i;\theta) \right\} \rightarrow_p \mathrm{Root}_\theta Em(Y;\theta) = \theta_0.$$

- Hence, an M estimator is guaranteed to be consistent as long as its estimating function has mean zero at the true parameter regardless of other aspects of the underlying distribution.

Basics of M Estimation

Example 1: Linear regression with unspecified error distribution

- Let Y and Z denote the response and the covariates, respectively.
- Suppose

$$Y = \beta^{\mathrm{T}} Z + \epsilon, \qquad E[\epsilon|Z] = 0. \tag{6.4}$$

  But we do not specify the conditional distribution of ϵ.
- Then, a moment's thought shows that the following estimating function

$$m_h(Y, Z; \beta) = h(Z)(Y - \beta^{\mathrm{T}} Z)$$

  is valid for any function h of Z.
- Choice of the weight function $h(Z)$ does not affect the validity of the M estimator as long as the model assumption (6.4) is satisfied.

Basics of M Estimation

Example 1: Linear regression with unspecified error distribution

- However, it does affect the asymptotic efficiency of the estimator through its influence function, as we will see.
- The particular choice of $h_0(Z) = Z$ make $m_{h_0}(Y, Z; \beta)$ proportional to the score function of the regression model with independent normal errors, so that this M estimator is equivalent to the MLE and is thus efficient when the underlying error distribution is normal.
- When the error distribution is not normal, the M estimator associated with $m_{h_0}(Y, Z; \beta)$ is still valid, but need not be efficient. Hence, it is called **locally efficient** at normal errors.

Basics of M Estimation

Asymptotics for the M estimator

- The derivation of asymptotics for the M estimator is essentially the same as the MLE, only with $\dot{l}(Y;\theta)$ replaced by $m(Y;\theta)$.
- Specifically, we have

$$\sqrt{n}(\widehat{\theta}_n - \theta_0) = -\left\{E\dot{m}(Y;\theta_0)\right\}^{-1} \frac{1}{\sqrt{n}} \sum_{i=1}^{n} m(Y_i;\theta_0) + o_p(1),$$

  where $\dot{m}(Y;\theta) = \frac{\partial}{\partial\theta} m(Y;\theta)$.
- So the influence function for $\widehat{\theta}_n$ is

$$-\left\{E\dot{m}(Y;\theta_0)\right\}^{-1} m(Y;\theta_0).$$

- The asymptotic variance is

$$\left\{E\dot{m}(Y;\theta_0)\right\}^{-1} Em(Y;\theta_0)^{\otimes 2} \left\{E\dot{m}(Y;\theta_0)^{\mathrm{T}}\right\}^{-1}.$$

Asymptotics for the M estimator

- Note that in this case, there is no reason to believe there is any relationship between $-E\dot{m}(Y;\theta_0)$ and $Em(Y;\theta_0)^{\otimes 2}$.
- So, an estimator for the asymptotic variance is the following **sandwich estimator**

$$\left\{n^{-1}\sum_{i=1}^{n}\dot{m}(Y_i;\widehat{\theta}_n)\right\}^{-1}\left\{n^{-1}\sum_{i=1}^{n}m(Y_i;\widehat{\theta}_n)^{\otimes 2}\right\}\left\{n^{-1}\sum_{i=1}^{n}\dot{m}(Y_i;\widehat{\theta}_n)^{\mathrm{T}}\right\}^{-1}.$$

- In Example 1, the influence function for the M estimator is

$$-\left\{Eh(Z)Z^{\mathrm{T}}\right\}^{-1}h(Z)(Y-\beta_0^{\mathrm{T}}Z).$$

The sandwich estimator for asymptotic variance is

$$\left\{n^{-1}\sum_{i=1}^{n}h(Z_i)Z_i^{\mathrm{T}}\right\}^{-1}n^{-1}\sum_{i=1}^{n}(Y_i-\widehat{\beta}_n^{\mathrm{T}}Z_i)^2h(Z)^{\otimes 2}\left\{n^{-1}\sum_{i=1}^{n}Z_ih(Z_i)^{\mathrm{T}}\right\}^{-1}.$$

Example 2: The Generalized Estimating Equation (GEE)

- Suppose each $Y_i = (Y_{i1}, \cdots, Y_{iq})^{\mathrm{T}}$ consists of responses repeatedly measured on one subject. The covariates are a $q \times p$ matrix $Z_i = (Z_{i1}, \cdots, Z_{iq})^{\mathrm{T}}$, where each Z_{ij} is a p-dimensional vector.
- We want to model the conditional mean of Y_{ij} as a function of Z_{ij} using the generalized linear models:

$$E[Y_{ij}|Z_{ij}] = \mu(\beta_0^{\mathrm{T}} Z_{ij}),$$

  where μ is a known link function.
- In matrix form, this is

$$E[Y_i|Z_i] = \mu(Z_i\beta_0), \tag{6.5}$$

  where the function $\mu(\cdot)$ operates component-wise.

Example 2: The Generalized Estimating Equation (GEE)

- However, the marginal regression models (6.5) do not full specify the joint distribution of Y_i, whose components are likely to be correlated since they are measured on one individual.
- One option is include random effects (see GLMM in §3.4) to induce correlation and derive the MLE.
- If the random effects model are wrongly specified, inference can be invalid.
- An alternative is to use estimating equations based on the marginal regression models (6.5) only. This approach is called the generalized estimating equation (GEE).
- Specifically, note that for any $p \times q$ weight matrix $H(Z)$, the estimating function

$$H(Z)(Y - Z\beta)$$

is valid.

Example 2: The Generalized Estimating Equation (GEE)

- Although the M estimator is valid no matter what H is, it is most efficient if H is chosen to be

$$Z^{\mathrm{T}}\mathsf{Diag}\{\dot{\mu}(Z\beta)\}\mathsf{Var}(Y|Z;\beta)^{-1}.$$

- Since the structure $\mathsf{Var}(Y|Z;\beta)$ is unknown, it is convenient to choose a working model assuming independent components of Y. So the estimating equation becomes

$$\sum_{i=1}^{n}\sum_{j=1}^{p} Z_{ij}\dot{\mu}(\beta^{\mathrm{T}}Z_{ij})\mathsf{Var}(Y_{ij}|Z_{ij};\beta)^{-1}\left(Y_{ij}-\mu(\beta^{\mathrm{T}}Z_{ij})\right)=0.$$

- Note that the left hand side is equal to the score function under the independence working model (see §3.1).

Example 2: The Generalized Estimating Equation (GEE)

- Hence the estimator from GEE with the independence working assumption is valid as long as the regression models (6.5) hold, that is, even when the independence assumption is not true, and is locally efficient when the independence assumption does hold.
- The sandwich estimator for the asymptotc variance is $A_n^{-1}\Omega_n A_n^{-1}$, where

$$A_n = n^{-1}\sum_{i=1}^{n}\sum_{j=1}^{p} Z_{ij}^{\otimes 2}\dot{\mu}(\widehat{\beta}_n^{\mathrm{T}} Z_{ij})^2 \mathsf{Var}(Y_{ij}|Z_{ij};\widehat{\beta}_n)^{-1},$$

and

$$\Omega_n = n^{-1}\sum_{i=1}^{n}\left\{\sum_{j=1}^{p} Z_{ij}\dot{\mu}(\widehat{\beta}_n^{\mathrm{T}} Z_{ij})\mathsf{Var}(Y_{ij}|Z_{ij};\widehat{\beta}_n)^{-1}\left(Y_{ij}-\mu(\widehat{\beta}_n^{\mathrm{T}} Z_{ij})\right)\right\}^{\otimes 2}.$$

Contents

Inverse Probability Weighting

- Suppose $m(Y;\theta)$ is a valid estimating function. With missing data, the full-data estimating function is not applicable because Y is not completely observed.
- Let $R = 1$ if Y is fully observed and $R = 0$ if otherwise. In this chapter, we will always assume that data are MAR.
- The complete case analysis using the estimating function $m(Y;\theta)$ is to solve

$$n^{-1}\sum_{i=1}^{n} R_i m(Y_i;\theta) = 0$$

  for θ.
- However, the CC analysis is generally biased, i.e., $E\{Rm(Y;\theta_0)\} \neq 0$, because the complete cases need not be a random sample from the population unless the data are MCAR.

Inverse Probability Weighting

- The inverse probability weighting (IPW) technique is a general approach to correcting the selection bias in the complete cases. It originates from the Horvitz-Thompson estimator in survey analysis to account for different sampling proportions.
- Specifically, suppose for each subject i, we know the probability of observing the full data Y_i, that is, the selection probability, denoted as $\pi_i = \Pr(R_i = 1|Y_i)$. Then, the bias can be corrected by weighting each complete case with π_i^{-1}, so the estimating equation becomes

$$n^{-1} \sum_{i=1}^{n} \pi_i^{-1} R_i m(Y_i; \theta) = 0.$$

- The IPW estimating function is valid because

$$E\left\{\pi_i^{-1} R_i m(Y_i; \theta_0)\right\} = E\left\{\pi_i^{-1} E\left[R_i|Y_i\right] m(Y_i; \theta_0)\right\} = E m(Y_i; \theta_0) = 0.$$

Inverse Probability Weighting

- The derivation of asymptotics of the resulting M estimator follows the full-data case closely except for weighting each subject by $\pi_i^{-1} R_i$.
- If missingness is not by design but by chance, as is the case in observational studies, then we typically do not know the selection probabilities.
- The obvious solution is to use a parametric model to estimate the selection probabilities and then weight each complete case by the inverse of the estimated selection probability.
- For simplicity, we first look at the case with two-levels of missingness, that is, one level is the full data Y, indicated by $R = 1$; the other is the partial data Y_{obs}, indicated by $R = 0$. We use X to denote Y_{obs}, that is, $X = Y_{\text{obs}}$.

Inverse Probability Weighting

- By the MAR assumption, the selection probability $\Pr(R=1|Y)$ is a function only of X. Denote this function as $\pi(X)$.
- In order to estimate the function $\pi(\cdot)$, we postulate a parametric model

$$\Pr(R=1|X) = \pi(X;\psi_0), \tag{6.6}$$

  where ψ is a finite-dimensional parameter.
- Note that $(R_i, X_i), i = 1, \cdots, n$ are fully observed (keep in mind that X is a component, or more generally a function, of Y, so X is observed in both incomplete and complete cases).
- So, an estimator $\widehat{\psi}_n$ of ψ_0 can be easily computed using the MLE based on the binary regression model (6.6).
- Popular choices for model (6.6) include logistic regression and probit regression models.

Inverse Probability Weighting

- After obtaining the estimate $\widehat{\psi}_n$, the selection probability for subject i can be estimated by $\pi(X_i; \widehat{\psi}_n)$, giving rise to the following IPW estimating equation

$$n^{-1} \sum_{i=1}^{n} \frac{R_i}{\pi(X_i; \widehat{\psi}_n)} m(Y_i; \theta) = 0. \tag{6.7}$$

  for θ.

- Two natural questions arise:
  1. Is the M estimator $\widehat{\theta}_n$ resulting from solving (6.7) consistent to θ_0?
  2. If yes, is the asymptotic distribution (or equivalently, the influence function) of $\widehat{\theta}_n$ the same as the M estimator from the estimating equation

$$n^{-1} \sum_{i=1}^{n} \frac{R_i}{\pi(X_i; \psi_0)} m(Y_i; \theta) = 0?$$

Inverse Probability Weighting

- The answer to the first question is yes, and that to the second is no.
- First, it is easy to see that the estimating function $\frac{R}{\pi(X;\psi_0)} m(Y;\theta)$ is valid in the sense that

$$E\left\{\frac{R}{\pi(X;\psi_0)} m(Y;\theta_0)\right\} = 0$$

  by iterative conditional expectation.
- Then, under some extra regularity conditions, we expect that

$$\begin{aligned} n^{-1}\sum_{i=1}^{n} \frac{R_i}{\pi(X_i;\widehat{\psi}_n)} m(Y_i;\theta) &= n^{-1}\sum_{i=1}^{n} \frac{R_i}{\pi(X_i;\psi_0)} m(Y_i;\theta) + o_p(1) \\ &\to_p E\left\{\frac{R}{\pi(X;\psi_0)} m(Y;\theta)\right\}, \end{aligned}$$

  so that $\widehat{\theta}_n$, the root for the left hand side goes in probability to θ_0, the root for the far right hand side.

Inverse Probability Weighting

M estimation in the presence of a nuisance parameter

- We now treat the second question about asymptotic distribution of $\widehat{\theta}_n$, which we frame in a more general setting of M estimation in the presence of an estimated nuisance parameter.
- Let $\{m(Y;\theta,\eta):\eta\}$ be a class of estimating functions for θ_0 indexed by an unknown nuisance parameter η. We call it a valid class of estimating functions for θ_0 if

$$Em(Y;\theta_0,\eta_0)=0.$$

- In our problem, $\eta=\psi$, and the class of estimating functions for θ is

$$\left\{\frac{R}{\pi(X;\psi)}m(Y;\theta):\psi\right\}.$$

  Clearly, this is a valid class by previous arguments.

Inverse Probability Weighting

M estimation in the presence of a nuisance parameter

- Now, we consider the following estimation equation for θ

$$n^{-1}\sum_{i=1}^{n} m(Y_i;\theta,\widehat{\eta}_n) = 0, \tag{6.8}$$

  where $\widehat{\eta}_n$ is a consistent estimator for η_0.
- Denote the M estimator from (6.8), the solution to (6.8), as $\widehat{\theta}_n$.
- We want to derive the asymptotic distribution, or equivalently, the influence function, of $\widehat{\theta}_n$.
- To this aim, we treatment $n^{-1}\sum_{i=1}^{n} m(Y_i;\theta,\eta)$ as a function of the bivariate (θ,η) and use (first-order) Taylor expansion around (θ_0,η_0) as similarly conducted in the standard case.

Inverse Probability Weighting

M estimation in the presence of a nuisance parameter

- We have that

$$\begin{aligned} n^{-1}\sum_{i=1}^{n} m(Y_i;\widehat{\theta}_n,\widehat{\eta}_n) = n^{-1}\sum_{i=1}^{n} m(Y_i;\theta_0,\eta_0) \\ + (\widehat{\theta}_n-\theta_0)n^{-1}\sum_{i=1}^{n}\dot{m}_1(Y_i;\theta_0,\eta_0) \\ + (\widehat{\eta}_n-\eta_0)n^{-1}\sum_{i=1}^{n}\dot{m}_2(Y_i;\theta_0,\eta_0) \\ + o_p\left(|\widehat{\theta}_n-\theta_0|+|\widehat{\eta}_n-\eta_0|\right), \end{aligned}$$

where $\dot{m}_1(Y;\theta,\eta)=\partial m(Y;\theta,\eta)/\partial\theta$ and $\dot{m}_2(Y;\theta,\eta)=\partial m(Y;\theta,\eta)/\partial\eta$.

Inverse Probability Weighting

M estimation in the presence of a nuisance parameter

- Note that, by definition, $n^{-1}\sum_{i=1}^{n} m(Y_i;\widehat{\theta}_n,\widehat{\eta}_n)=0$. By similar re-arrangements and arguments as in the standard case, we have that

$$\sqrt{n}(\widehat{\theta}_n-\theta_0)=-\left\{E\dot{m}_1(Y;\theta_0,\eta_0)\right\}^{-1}\left\{\frac{1}{\sqrt{n}}\sum_{i=1}^{n} m(Y_i;\theta_0,\eta_0)\right.$$
$$\left.+E\dot{m}_2(Y;\theta_0,\eta_0)\sqrt{n}(\widehat{\eta}_n-\eta_0)\right\}+o_p(1). \qquad (6.9)$$

- Suppose $\widehat{\eta}_n$ is linearized as

$$\sqrt{n}(\widehat{\eta}_n-\eta_0)=\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\Psi_\eta(Y_i;\eta_0)+o_p(1).$$

M estimation in the presence of a nuisance parameter

- Then, (6.9) can be further linearized as

$$\sqrt{n}(\widehat{\theta}_n - \theta_0) = -\left\{E\dot{m}_1(Y;\theta_0,\eta_0)\right\}^{-1} \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left\{ m(Y_i;\theta_0,\eta_0) + V(\theta_0,\eta_0)\Psi_\eta(Y_i;\eta_0) \right\} + o_p(1). \tag{6.10}$$

  where $V(\theta_0,\eta_0) = E\dot{m}_2(Y;\theta_0,\eta_0)$.

- So the influence function for $\widehat{\theta}_n$ is

$$-\left\{E\dot{m}_1(Y;\theta_0,\eta_0)\right\}^{-1} \left\{ m(Y;\theta_0,\eta_0) + V(\theta_0,\eta_0)\Psi_\eta(Y_i;\eta_0) \right\}.$$

Inverse Probability Weighting

M estimation in the presence of a nuisance parameter

- The first term of the previous display is the influence function of the M estimator with estimating equation $m(Y;\theta,\eta_0)$, and the second term represents the "influence" of estimating η_0.
- A sandwich estimator for the asymptotic variance can be similarly constructed as

$$\left\{ n^{-1}\sum_{i=1}^{n} \dot{m}_1(Y_i;\widehat{\theta}_n,\widehat{\eta}_n) \right\}^{-1} n^{-1}\sum_{i=1}^{n} \Big\{ m(Y_i;\widehat{\theta}_n,\widehat{\eta}_n) + \widehat{V}_n(\widehat{\theta}_n,\widehat{\eta}_n)\Psi_\eta(Y_i;\widehat{\eta}_n) \Big\}^{\otimes 2} \left\{ n^{-1}\sum_{i=1}^{n} \dot{m}_1(Y_i;\widehat{\theta}_n,\widehat{\eta}_n)^{\mathrm{T}} \right\}^{-1},$$

where

$$\widehat{V}(\theta,\eta) = n^{-1}\sum_{i=1}^{n} \dot{m}_2(Y_i;\theta,\eta).$$

Asymptotics for the IPW estimator

- In our IPW estimation equation setting, we have that $\eta = \psi$,

$$m(Y;\theta,\psi) = \frac{R}{\pi(X;\psi)} m(Y;\theta),$$

$\dot{m}_1(Y;\theta,\psi) = \frac{R}{\pi(X;\psi)}\dot{m}(Y;\theta)$, and

$$\dot{m}_2(Y;\theta,\psi) = -\frac{R}{\pi(X;\psi)^2} m(Y;\theta)\dot{\pi}(X;\psi)^{\mathrm{T}}.$$

- If the model $\pi(X;\psi)$ is assumed to be a logistic regression model, then

$$\pi(X;\psi) = \frac{\exp(\psi^{\mathrm{T}}X)}{1+\exp(\psi^{\mathrm{T}}X)},$$

and the (efficient) influence function of $\widehat{\psi}_n$ is given by

$$\Psi_\psi(R_i, X_i;\psi) = \left\{ E\frac{\exp(\psi^{\mathrm{T}}X)}{(1+\exp(\psi^{\mathrm{T}}X))^2} X^{\otimes 2} \right\}^{-1} \left(R - \frac{\exp(\psi^{\mathrm{T}})X}{1+\exp(\psi^{\mathrm{T}}X)} X \right).$$

Inverse Probability Weighting

Example: Estimating the mean of a partially observed outcome

- In Chapter 1 we looked at an example of estimating the mean of an outcome, i.e., weight, after an intervention program.
- The post-intervention weight was not observed on those who did not participate in the program. But it was assumed that whether the subject chose to participate, and thus, whether the variable of interest was observed, depends totally on the person's pre-intervention weight.
- Denote $Y = (X, Y_2)$, where X and Y_2 are pre- and post-intervention weights, respectively. We thus want to estimate $\theta := EY_2$. Let $R = 1$ if Y_2 is observed, and $R = 0$ if otherwise. By MAR assumption, $\Pr(R = 1|Y)$ is only a function of X.

Inverse Probability Weighting

Example: Estimating the mean of a partially observed outcome

- For the purpose of estimating the selection probability, we use a logistic regression model

$$\Pr(R = 1|X) = \frac{\exp(\psi_0^{\mathrm{T}} \widetilde{X})}{1 + \exp(\psi_0^{\mathrm{T}} \widetilde{X})},$$

  where $\widetilde{X} = (1, X)$. Denote $\widehat{\psi}_n$ as the MLE for ψ_0.
- A natural estimator for θ had the full data been available is $n^{-1} \sum_{i=1}^{n} Y_{2i}$, which the M estimator for the estimating function

$$m(Y; \theta) = Y_2 - \theta.$$

- Thus the IPW estimating equation is

$$n^{-1} \sum_{i=1}^{n} \frac{R_i}{\pi(X_i; \widehat{\psi}_n)} (Y_{2_i} - \theta) = 0.$$

Inverse Probability Weighting

Example: Estimating the mean of a partially observed outcome

- This estimating equation has an explicit solution

$$\widehat{\theta}_n = \frac{n^{-1}\sum_{i=1}^n \frac{R_i}{\pi(X_i;\widehat{\psi}_n)} Y_{2i}}{n^{-1}\sum_{i=1}^n \frac{R_i}{\pi(X_i;\widehat{\psi}_n)}}.$$

- By the theory developed above, the influence function can be shown to be

$$\frac{R}{\pi(X;\psi_0)}(Y_{2i}-\theta_0) - \frac{R}{\pi(X;\psi_0)^2}(Y_{2i}-\theta_0)\dot{\pi}(X;\psi_0)^{\mathrm{T}}$$
$$\times \left\{ E\frac{\exp(\psi_0^{\mathrm{T}}\widetilde{X})}{(1+\exp(\psi_0^{\mathrm{T}}\widetilde{X}))^2} X^{\otimes 2} \right\}^{-1} \left(R - \frac{\exp(\psi_0^{\mathrm{T}}\widetilde{X})}{1+\exp(\psi_0^{\mathrm{T}}\widetilde{X})} \right) \widetilde{X}.$$

- A variance estimator for $\widehat{\theta}_n$ can be constructed accordingly.

Inverse Probability Weighting

IPW for Monotone Missing Data

- For monotone missingness, the IPW M estimator works similarly as in the two-level case, i.e., by inversely weighting each complete case by its selection probability.
- The extra difficulty lies in modeling the missingness mechanism, and hence the selection probability.
- Let the full data be $Y = (Y_1, \cdots, Y_q)$. Let $R = j$, $j = 1, \cdots, q$, if

$$Y_{\text{obs}} = (Y_1, \cdots, Y_j).$$

- So, the indicator for a complete case is $I(R = q)$. We want to model this probability given the full data while assuming MAR.
- This can be completed in the following ways.

IPW for Monotone Missing Data

- We model the conditional selection probabilities

$$\Pr(R = j | R \geq j, Y) = \lambda_j(\widetilde{Y}_j; \psi_{j0}), \tag{6.11}$$

  where $\widetilde{Y}_j = (Y_1, \cdots, Y_j)$, $j = 1, \cdots, q-1$. Note that the left hand side being only a function of $\widetilde{Y}_j$ is justified by MAR.
- Denote the selection probability as $\pi(Y; \psi_0) = \Pr(R = q|Y)$, where $\psi_0 = (\psi_{10}, \cdots, \psi_{q0})$. It can be shown that

$$\pi(Y; \psi_0) = \prod_{j=1}^{q-1} \Big\{ 1 - \lambda_j(\widetilde{Y}_j; \psi_{j0}) \Big\}$$

- It is convenient to use binary regression models such as logistic regression to model (6.11). The MLEs for $\widehat{\psi}_{jn}$ are based on "complete cases", i.e., subjects for which $\widetilde{Y}_j$ is observed.

IPW for Monotone Missing Data

- To see this, note that those with some components of $\widetilde{Y}_j$ missing are excluded by the conditioning event $\{R \geq j\}$.
- Then, the selection probability can be estimated by

$$\pi(Y; \widehat{\psi}_n) = \prod_{j=1}^{q-1} \left\{ 1 - \lambda_j(\widetilde{Y}_j; \psi_{jn}) \right\}.$$

- Then, the IPW M estimator goes by solving

$$n^{-1} \sum_{i=1}^{n} \frac{I(R_i = q)}{\pi(Y; \widehat{\psi}_n)} m(Y_i; \theta) = 0$$

  for θ.
- Though more complicated, the asymptotic variance for $\widehat{\theta}_n$ can be derived similarly as in the two-level case.

Contents

Doubly Robust Estimation with Two-Levels of Missingness

- We use Example 1 of §6.2 to compare the IPW approach with the MLE in dealing with missing data.
- Recall that the full data are $Y = (X, Y_2)$, where X is always observed and Y_2 is possibly missing ($R = 1$: Y_2 is observed; $R = 0$, Y_2 is missing). We have assumed that the Y_2 are MAR, that is, the probability of missingness depends on X. The interest is in estimating the mean of Y_2. So we call X an **auxiliary** variable.
- The MLE approach requires a joint model for the full data $Y = (X, Y_2)$, say, bivariate normal, and EY_2 is computed based on this model.
- The IPW approach requires a selection model, i.e., $\Pr(R = 1|X)$, and EY_2 is computed based on inversely weighting the complete cases with the estimated selection probabilities.

Doubly Robust Estimation with Two-Levels of Missingness

- When the joint model for (X, Y_2) is the bivaraite normal, then it can be shown that the MLE (derived via an EM algorithm in chapter 1) for EY_2 is equivalent to

$$n^{-1}\left\{R_i Y_{2_i} + (1 - R_i)\widehat{E}(Y_{2i}|X_i)\right\},$$

  where $\widehat{E}(Y_{2i}|X_i)$ is based on a bivariate normal distribution with parameters estimated from the complete cases. In other words, the MLE seeks to "impute" the missing values of Y_2 using the auxiliary variable X based on the their assumed joint distribution.
- The MLE would be biased if the posited joint model is incorrect.
- The IPW estimation differs from MLE in two respects, one in terms of modeling assumptions and the other in terms of the data used.

Doubly Robust Estimation with Two-Levels of Missingness

- For modeling assumptions, instead of the joint distribution of (X, Y_2), the IPW specifies the conditional distribution $[R|X]$. The IPW estimator would be biased if the the selection model is wrongly specified.
- Also, as a result of not positing any relationship between X and Y_2, the IPW essentially uses information in the complete cases.
- There are two ways to improve the IPW estimator.
- First, we want to construct an estimator that is valid when either the model $[Y_2|X]$ or the model $[R|X]$ is true.
- Second, we want to exploit the association between X and Y_2 to gain efficiency with compromising robustness.

Doubly Robust Estimation with Two-Levels of Missingness

- Suppose

  Consider the following estimator for EY_2:

$$n^{-1} \sum_{i=1}^{n} \frac{R_i}{\pi}$$

Contents