STAT 542: Multivariate Analysis **Spring 2021**

Lecture 5: Empirical likelihood method

Instructor: Yen-Chi Chen

5.1 Empirical likelihood

The empirical likelihood (EL) is a nonparametric (though sometimes people viewed it as a semi-parametric) approach for computing an estimator. The idea is to find a 'maximum likelihood estimate' (MLE) of the distribution function F with some moment constraints. The main reference of EL is the following paper

> [O1990] Owen, A. (1990). Empirical Likelihood Ratio Confidence Regions. The Annals of Statistics, 18(1), 90-120.

Let $X_1, \cdots, X_n \sim F_0$, where F_0 is the underlying distribution function. Suppose that we know the mean of the distribution $\mu_0 = \int x dF_0(x)$ in advance. We want to compute the 'MLE' of the underlying distribution subject to this constraint. Note that here we do not specify any parametric model of F.

We consider a PMF P such that

$$P(X_i) = F(X_i) - F(X_i^-) = W_i.$$

Each P leads to a CDF F. Now we treat this PMF as if it is a multinomial distribution can consider the likelihood function

$$L_n(P) = \prod_{i=1}^n P(X_i)$$

or the log-likelihood function

$$\ell_n(P) = \sum_{i=1}^n \log P(X_i) = \sum_{i=1}^n \log W_i. \tag{5.1}$$

With this, we have specify a likelihood function of any PMF P. The idea of empirical likelihood is to find P such that equation (5.1) is maximized and the constraint $\mu_0 = \int x dF(x) = \sum_{i=1}^n P(X_i)X_i = \sum_{i=1}^n W_i X_i$ holds. Namely, we want to find

$$\widehat{W} = \mathsf{argmax}_{W_1,\cdots,W_n} \sum_{i=1}^n \log W_i.$$
$$\text{s.t. } \mu_0 = \sum_{i=1}^n W_i X_i, \quad \sum_{i=1}^n W_i = 1, \quad W_i \geq 0.$$

One can easily see that this procedure can be easily combined with any equation-type constraint. Specifically, if the constraint is $\mathbb{E}(g(X)) = 0$ for some given function $g \in \mathbb{R}^p$, then the empirical likelihood will be

$$\widehat{W} = \mathsf{argmax}_{W_1,\cdots,W_n} \sum_{i=1}^n \log W_i.$$
$$\text{s.t. } 0 = \sum_{i=1}^n W_i g(X_i), \quad \sum_{i=1}^n W_i = 1, \quad W_i \geq 0,$$

where $W = (W_1, \cdots, W_n)$ and $\widehat{W} = (\widehat{W}_1, \cdots, \widehat{W}_n)$.

Using the Lagrangian multiplier, we want to minimize the following quantity:

$$\sum_{i=1}^{n} \log W_i - \lambda^T \sum_{i=1}^{n} W_i g(X_i) - \mu(\sum_{i=1}^{n} W_i - 1)$$

with respect to $W_1, \cdots, W_n, \lambda, \mu$. Taking derivative with respect to W_i, we obtain

$$\frac{1}{W_i} = \lambda^T g(X_i) + \mu$$

or equivalently,

$$W_i = \frac{1}{\mu + \lambda^T g(X_i)}.$$

To solve μ, the above equation implies $1 = W_i(\mu + \lambda^T g(X_i))$ which further implies

$$n = \sum_{i=1}^{m} W_i(\mu + \lambda^T g(X_i)) = \mu + \lambda^T \underbrace{\sum_{i=1}^{n} W_i g(X_i)}_{=0} = \mu.$$

As a result, we obtain a closed-form solution as

$$W_i = \frac{1}{n + \widehat{\lambda}^T g(X_i)}$$

and the multiplier $\widehat{\lambda}$ satisfies the constraint

$$0 = \sum_{i=1}^{n} \frac{g(X_i)}{n + \widehat{\lambda}^T g(X_i)}.$$

Note that one can always absorb n into $\widehat{\lambda}$ and rewrite the above as

$$W_i = \frac{1}{1 + \widehat{\lambda}^T g(X_i)}, \quad 0 = \sum_{i=1}^{n} \frac{g(X_i)}{1 + \widehat{\lambda}^T g(X_i)}. \tag{5.2}$$

5.1.1 Estimating equation

The empirical likelihood method can be applied to various problems. Here we will show how it can be used to solve an estimating equation. We will briefly describe the procedure in

> [QL1994] Qin, J., & Lawless, J. (1994). Empirical likelihood and general estimating equations. the Annals of Statistics, 300-325.

Suppose that the parameter of interest $\beta \in \mathbb{R}^p$ is derived from solving the (population) estimating equation

$$\beta_0 : 0 = \mathbb{E}(g(X; \beta_0)).$$

This implies that for any given β, we can try to find $W = W(\beta)$ and $\widehat{\lambda} = \widehat{\lambda}(\beta)$ such that

$$W_i(\beta) = \frac{1}{1 + \widehat{\lambda}^T g(X_i; \beta)}, \quad 0 = \sum_{i=1}^{n} \frac{g(X_i; \beta)}{1 + \widehat{\lambda}^T(\beta) g(X_i; \beta)}. \tag{5.3}$$

Thus, the empirical log-likelihood as a function of β is

$$\ell_{n,\mathsf{EL}}(\beta) = -\sum_{i=1}^{n} \log\left(1 + \widehat{\lambda}^T(\beta) g(X_i;\beta)\right).$$

So we can estimate β via maximizing the empirical log-likelihood or equivalently,

$$\widehat{\beta}_{\mathsf{EL}} = \mathsf{argmin}_\beta \sum_{i=1}^{n} \log\left(1 + \widehat{\lambda}^T(\beta) g(X_i;\beta)\right). \tag{5.4}$$

This is the maximum empirical likelihood estimator (MELE) described in [QL1994].

Interestingly, there is another way of expressing the MELE as a saddle point problem. Under smoothness conditions, $\widehat{\lambda}(\beta)$ can be expressed as

$$\widehat{\lambda}(\beta) = \mathsf{argmax}_{\lambda \in \Lambda_n(\beta)} \sum_{i=1}^{n} \log\left(1 + \lambda^T g(X_i;\beta)\right),$$

where the feasible region

$$\Lambda_n(\beta) = \{1 + \lambda^T g(X_i;\beta) \geq 0 : i = 1, \cdots, n\}$$

is to ensure that the weight in equation (5.3) is non-negative. One can easily verify that the first-order condition (gradient set to 0) leads to the usual constraint to define $\widehat{\lambda}(\beta)$. As a result, we can rewrite equation (5.4) as

$$\widehat{\beta}_{\mathsf{EL}} = \mathsf{argmin}_\beta \sup_{\lambda \in \Lambda_n(\beta)} \sum_{i=1}^{n} \log\left(1 + \lambda^T g(X_i;\beta)\right). \tag{5.5}$$

One can show that under appropriate conditions (β_0 solves the population equation $0 = \mathbb{E}(g(X;\beta_0))$),

$$\sqrt{n}(\widehat{\beta}_{\mathsf{EL}} - \beta_0) \xrightarrow{D} N(0,\sigma^2), \quad \sqrt{n}(\widehat{\lambda}(\widehat{\beta}_{\mathsf{EL}}) - 0) \xrightarrow{D} N(0,\sigma_\lambda^2). \tag{5.6}$$

This result can be found in Theorem 1 of [QL1994]. Moreover, $(\widehat{\beta}_{\mathsf{EL}}, \widehat{\lambda})$ converges jointly to a multivariate normal distribution. So the usual inference on the parameter can be applied. Note that the population version of $\widehat{\lambda}(\widehat{\beta}_{\mathsf{EL}})$ is 0 from equation (5.6); one can easily see this because in the population case, the maximal value of $\mathbb{E}[\log\left(1 + \lambda^T g(X_i;\beta)\right)]$ occurs when $\lambda = 0$. Also, the empirical likelihood value can also be used to perform hypothesis test, see Theorem 2 of [QL1994].

5.2 Generalized empirical likelihood

Equation (5.5) shows an elegant form of writing an estimator in terms of a saddle point problem. This form further motivates the generalized empirical likelihood method. The main reference of the generalized empirical likelihood is the following paper:

> [NS2004] Newey, W. K., & Smith, R. J. (2004). Higher order properties of GMM and generalized empirical likelihood estimators. Econometrica, 72(1), 219-255.

To be consistent with the notation in above paper, we re-write equation (5.5) as

$$\widehat{\beta}_{\mathsf{EL}} = \mathsf{argmin}_\beta \sup_{\lambda \in \Lambda_n(\beta)} \sum_{i=1}^{n} \log\left(1 - \lambda^T g(X_i;\beta)\right). \tag{5.7}$$

Since we are taking the supremum over λ, this will lead to the same result. Now we define $\rho_{EL}(\lambda^T g(X_i;\beta)) = \log\big(1-\lambda^T g(X_i;\beta)\big)$. Then equation (5.7) can be written as

$$\widehat{\beta}_{\mathsf{EL}} = \mathsf{argmin}_\beta \sup_{\lambda\in\Lambda_n(\beta)} \sum_{i=1}^n \rho_{EL}(\lambda^T g(X_i;\beta)). \tag{5.8}$$

The function $\rho_{EL}(x)$ has a feature that $\rho_{EL}(0)=0, \rho'_{EL}(1)=\rho''_{EL}(0)=-1$.

This form hints on the fact that we may replace ρ_{EL} by some other functions ρ satisfying

$$\rho(0)=0, \rho'(1)=\rho''(1)=-1, \rho \text{ is concave} \tag{5.9}$$

and $\rho(s)$ is only defined on $s\in(-\infty,1]$. With a function satisfies the above constraint, we define the generalized empirical likelihood (GEL) estimator of a generalized estimating equation as

$$\widehat{\beta}_{\mathsf{GEL}} = \mathsf{argmin}_\beta \sup_{\lambda\in\Lambda_n(\beta)} \sum_{i=1}^n \rho(\lambda^T g(X_i;\beta)). \tag{5.10}$$

Moreover, the fact that when using the empirical likelihood, the weights W_i can be expressed as (from equation (5.3))

$$W_i = \frac{(1+\widehat{\lambda}^T g(X_i))^{-1}}{\sum_{j=1}^n (1+\widehat{\lambda}^T g(X_j))^{-1}} = \frac{\rho'_{EL}(\widehat{\lambda}^T g(X_i))}{\sum_{j=1}^n \rho'_{EL}(\widehat{\lambda}^T g(X_j))}.$$

So we can interpret

$$W_{\rho,i} = \frac{\rho'(\widehat{\lambda}^T(\beta) g(X_i;\beta))}{\sum_{j=1}^n \rho'(\widehat{\lambda}^T(\beta) g(X_j;\beta))}, \tag{5.11}$$

where

$$\widehat{\lambda}(\beta) = \mathsf{argmax}_{\lambda\in\Lambda_n(\beta)} \sum_{j=1}^n \rho(\lambda^T g(X_j;\beta)). \tag{5.12}$$

Under suitable conditions, one can show the result similar to equation (5.13) as

$$\sqrt{n}(\widehat{\beta}_{\mathsf{GEL}}-\beta_0) \xrightarrow{D} N(0,\sigma^2), \quad \sqrt{n}(\widehat{\lambda}(\widehat{\beta}_{\mathsf{GEL}})-0) \xrightarrow{D} N(0,\sigma_\lambda^2). \tag{5.13}$$

This result can be found in Theorem 3.2 of [NS2004].

5.2.1 A brief derivation of the asymptotic normality

Here is a brief derivation of the asymptotic normality in equation (5.13). The formal assumptions and proofs can be found in [NS2004] (Theorem 3.1 and 3.2). The key idea is that we choose ρ to be a smooth function so that the minimum $\widehat{\lambda}(\widehat{\beta}_{\mathsf{GEL}})$ and the maximum $\widehat{\beta}_{\mathsf{GEL}}$ satisfies the first order condition (gradient being 0).

Let $G(x;\beta) = \nabla_\beta g(x;\beta) \in \mathbb{R}^{p\times p}$ be the derivative of $g(x;\beta)$ with respect to β. Then $\widehat{\lambda}(\widehat{\beta}_{\mathsf{GEL}})$ and $\widehat{\beta}_{\mathsf{GEL}}$ satisfy

$$\begin{aligned}
0 &= \frac{\partial}{\partial\beta}\frac{1}{n}\sum_{i=1}^n \rho(\widehat{\lambda}^T(\widehat{\beta}_{\mathsf{GEL}}) g(X_i;\widehat{\beta}_{\mathsf{GEL}})) = \frac{1}{n}\sum_{i=1}^n \rho'(\widehat{\lambda}^T(\widehat{\beta}_{\mathsf{GEL}}) g(X_i;\widehat{\beta}_{\mathsf{GEL}})) G(X_i;\widehat{\beta}_{\mathsf{GEL}})\widehat{\lambda}(\widehat{\beta}_{\mathsf{GEL}}),\\
0 &= \frac{\partial}{\partial\lambda}\frac{1}{n}\sum_{i=1}^n \rho(\widehat{\lambda}^T(\widehat{\beta}_{\mathsf{GEL}}) g(X_i;\widehat{\beta}_{\mathsf{GEL}})) = \frac{1}{n}\sum_{i=1}^n \rho'(\widehat{\lambda}^T(\widehat{\beta}_{\mathsf{GEL}}) g(X_i;\widehat{\beta}_{\mathsf{GEL}})) g(X_i;\widehat{\beta}_{\mathsf{GEL}}).
\end{aligned} \tag{5.14}$$

We will abbreviate the above two equations as $\frac{1}{n}\sum_{i=1}^n \Psi(\widehat{\beta}_{\mathsf{GEL}}, \widehat{\lambda}(\widehat{\beta}_{\mathsf{GEL}})|X_i) = 0$.

We will use the regular approach to derive the asymptotic normality–compare this to the 'population version' of $\widehat{\beta}_{\mathsf{GEL}}, \widehat{\lambda}(\widehat{\beta}_{\mathsf{GEL}})$. The population version of them will be $(\beta_0, \lambda_0) = (\beta_0, 0)$ with $\mathbb{E}(g(X_i;\beta_0)) = 0$. To see why they solve the population version of the two equations, the population version of two first order equations are

$$0 = \mathbb{E}(\rho'(\lambda_0^T g(X_i;\beta_0))G(X_i;\beta_0)\lambda_0),$$
$$0 = \mathbb{E}(\rho'(\lambda_0^T g(X_i;\beta_0))g(X_i;\beta_0)).$$

When $\lambda_0 = 0$, the first equation automatically satisfies and the second equation becomes (using $\rho'(0) = -1$) $-\mathbb{E}(g(X_i;\beta_0)) = 0$ by the definition of β_0. Thus, the choice $(\beta_0, \lambda_0) = (\beta_0, 0)$ satisfies the population first order conditions. We abbreviate the population equations as $\Psi_0(\beta_0, 0) = 0$

By definition, one can easily see that

$$\mathbb{E}(\Psi(\beta, \lambda|X_i)) = \Psi_0(\beta, \lambda).$$

As a result, using the Taylor's expansion and the fact that $\frac{1}{n}\sum_{i=1}^n \Psi(\widehat{\beta}_{\mathsf{GEL}}, \widehat{\lambda}(\widehat{\beta}_{\mathsf{GEL}})|X_i) = 0$:

$$\begin{aligned}
\frac{1}{n}\sum_{i=1}^n \Psi(\beta_0, \lambda_0|X_i) &= \frac{1}{n}\sum_{i=1}^n \Psi(\beta_0, \lambda_0|X_i) - \underbrace{\mathbb{E}(\Psi(\beta_0, \lambda_0|X_i))}_{=0} \\
&= \frac{1}{n}\sum_{i=1}^n \left(\Psi(\beta_0, \lambda_0|X_i) - \Psi(\widehat{\beta}_{\mathsf{GEL}}, \widehat{\lambda}(\widehat{\beta}_{\mathsf{GEL}})|X_i)\right) \\
&\approx \frac{1}{n}\sum_{i=1}^n \left[\frac{\partial}{\partial\beta}\Psi(\beta_0, \lambda_0|X_i)\right](\beta_0 - \widehat{\beta}_{\mathsf{GEL}}) + \left[\frac{\partial}{\partial\lambda}\Psi(\beta_0, \lambda_0|X_i)\right](\lambda_0 - \widehat{\lambda}(\widehat{\beta}_{\mathsf{GEL}})).
\end{aligned}$$

Thus, we conclude that the vector

$$\begin{pmatrix} \beta_0 - \widehat{\beta}_{\mathsf{GEL}}) \\ \lambda_0 - \widehat{\lambda}(\widehat{\beta}_{\mathsf{GEL}}) \end{pmatrix} \approx \begin{bmatrix} \frac{1}{n}\sum_{i=1}^n \frac{\partial}{\partial\beta}\Psi(\beta_0, \lambda_0|X_i) \\ \frac{1}{n}\sum_{i=1}^n \frac{\partial}{\partial\lambda}\Psi(\beta_0, \lambda_0|X_i) \end{bmatrix}^{-1} \frac{1}{n}\sum_{i=1}^n \Psi(\beta_0, \lambda_0|X_i).$$

Apparently, $\frac{1}{\sqrt{n}}\sum_{i=1}^n \Psi(\beta_0, \lambda_0|X_i)$ has asymptotic normality since it is the summation of IID random elements. Also, by the law of large number,

$$\begin{aligned}
\frac{1}{n}\sum_{i=1}^n \frac{\partial}{\partial\beta}\Psi(\beta_0, \lambda_0|X_i) &\xrightarrow{p} \mathbb{E}\left[\frac{\partial}{\partial\beta}\Psi(\beta_0, \lambda_0|X_i)\right] \\
\frac{1}{n}\sum_{i=1}^n \frac{\partial}{\partial\lambda}\Psi(\beta_0, \lambda_0|X_i) &\xrightarrow{p} \mathbb{E}\left[\frac{\partial}{\partial\lambda}\Psi(\beta_0, \lambda_0|X_i)\right].
\end{aligned}$$

So we conclude that by the Slutsky's theorem (and the fact that $\lambda_0 = 0$),

$$\sqrt{n}\begin{pmatrix} \widehat{\beta}_{\mathsf{GEL}}) - \beta_0 \\ \widehat{\lambda}(\widehat{\beta}_{\mathsf{GEL}}) - 0 \end{pmatrix} \xrightarrow{d} N(0, \Sigma)$$

for some covariance matrix Σ.

For the completeness, we derive the closed-form of $\mathbb{E}\left[\frac{\partial}{\partial\beta}\Psi(\beta_0, \lambda_0|X_i)\right]$ and $\mathbb{E}\left[\frac{\partial}{\partial\lambda}\Psi(\beta_0, \lambda_0|X_i)\right]$. Note that $\Psi = (\Phi_\beta, \Phi_\lambda) = (\frac{\partial}{\partial\beta}\Phi, \frac{\partial}{\partial\lambda}\Phi)$ consists of two functions in (5.14) and $\Phi(\beta, \lambda|x) = \rho(\lambda^T g(x;\beta))$. Thus, denoting $\Phi_{a,b}$ as the partial derivatives with respect to a and partial derivative with respect to b, we can express

$$\begin{bmatrix} \mathbb{E}\left(\frac{\partial}{\partial\beta}\Psi(\beta_0, \lambda_0|X_i)\right) \\ \mathbb{E}\left(\frac{\partial}{\partial\lambda}\Psi(\beta_0, \lambda_0|X_i)\right) \end{bmatrix} = \begin{bmatrix} \mathbb{E}[\Phi_{\beta,\beta}(\beta, \lambda|X_i)] & \mathbb{E}[\Phi_{\beta,\lambda}(\beta, \lambda|X_i)] \\ \mathbb{E}[\Phi_{\lambda,\beta}(\beta, \lambda|X_i)] & \mathbb{E}[\Phi_{\lambda,\lambda}(\beta, \lambda|X_i)] \end{bmatrix},$$

and

$$\begin{aligned}
\Phi_{\beta,\beta}(\beta_0,\lambda_0|X_i)] &= 0 \quad (\text{since } \lambda_0 = 0)\\
\Phi_{\beta,\lambda}(\beta_0,\lambda_0|X_i)] &= \mathbb{E}\left[\rho'(0)G(X_i;\beta_0)\right]\\
&= -\mathbb{E}\left[G(X_i;\beta_0)\right]\\
\Phi_{\lambda,\beta}(\beta_0,\lambda_0|X_i)] &= \Phi_{\beta,\lambda}(\beta_0,\lambda_0|X_i)]\\
&= -\mathbb{E}\left[G(X_i;\beta_0)\right]\\
\Phi_{\lambda,\lambda}(\beta_0,\lambda_0|X_i)] &= \mathbb{E}\left[\rho''(0)G(X_i;\beta_0)G^T(X_i;\beta_0)\right]\\
&= -\mathbb{E}\left[G(X_i;\beta_0)G^T(X_i;\beta_0)\right]
\end{aligned}$$

5.3 Calibration

The idea of GEL approach can also be applied to missing data and causal inference problems. In this case, the idea is called *calibaration*. In what follows, we will describe how this idea is carried out in the missing data problem. The main reference is the following paper

> [CY2014] Chan, K. C. G., & Yam, S. C. P. (2014). Oracle, multiple robust and multipurpose calibration in a missing response problem. Statistical Science, 29(3), 380-396.

The application of calibration in a causal inference problem can be seen in the following paper:

> [CY2016] Chan, K. C. G., Yam, S. C. P., & Zhang, Z. (2016). Globally efficient non-parametric inference of average treatment effects by empirical balancing calibration weighting. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 78(3), 673-700.

The calibration is a synthesis of GEL and the inverse probability weighting (IPW) approach and the regression adjustment (RA) approach. Consider a simple missing data problem where the outcome $Y \in \mathbb{R}$ may be missing and let R be the binary response pattern of Y in the sense that if $R = 1$, we observe Y and if $R = 0$, we do not. In addition to these two variables, even individual has a covariate X that is always observed. We assume the simple missing at random (MAR) condition that

$$Y \perp R|X.$$

Let $\pi(X) = P(R = 1|X)$ be the probability of observing Y given X. Then one can easily see that if we want to estimate the marginal mean of Y, we have

$$\theta \equiv \mathbb{E}(Y) = \mathbb{E}\left(\frac{YR}{\pi(X)}\right).$$

So a consistent estimator of θ (known as the IPW estimator) is

$$\widehat{\theta}_{\mathsf{IPW}} = \frac{1}{n}\sum_{i=1}^{n}\frac{Y_i R_i}{\widehat{\pi}(X_i)},$$

where $\widehat{\pi}$ is an estimate of π.

The IPW estimator is often not an efficient estimator. The idea of calibration is an attempt to improve the efficiency of the IPW estimator while preserving its beautiful form.

The insight of calibration is from the observation that we can rewrite the IPW estimator as

$$\widehat{\theta}_{\mathsf{IPW}} = \sum_{i=1}^{n} Y_i R_i W_i, \tag{5.15}$$

where W_i is the weight distributed on the i-th observation. And since both IPW and RA estimator are estimating the same quantity, we can *adjust* the weights using the idea of GEL so that

Before we formally introduce the idea of calibration, observe the fact that the MAR assumption implies

$$m_1(x) = \mathbb{E}(Y|X = x, R = 1) = \mathbb{E}(Y|X = x, R = 0) = m_0(x)$$

so we can estimate the same quantity using

$$\theta \equiv \mathbb{E}(Y) = \mathbb{E}(m_1(X)), \tag{5.16}$$

leading to the RA (regression adjustment) estimator

$$\widehat{\theta}_{\mathsf{RA},1} = \frac{1}{n}\sum_{i=1}^{n} \widehat{m}_1(X_i),$$

where $\widehat{m}_1$ is an estimate of m_1.

In addition to the above form, one can show that we can combine the IPW and RA estimates as

$$\theta \equiv \mathbb{E}(Y) = \mathbb{E}\left(\frac{m_1(X)R}{\pi(X)}\right), \tag{5.17}$$

leading to the estimator

$$\widehat{\theta}_{\mathsf{RA},2} = \frac{1}{n}\sum_{i=1}^{n} \frac{\widehat{m}_1(X_i)R_i}{\widehat{\pi}(X_i)}.$$

So the idea of calibration is: given that equations (5.16) and (5.17) are targeting at the same quantity, we can choose $W_1, \cdots, W_n$ to satisfy the constraint that equations (5.16) and (5.17) are the same. Specifically, we want to choose W such that

$$\begin{aligned} \bar{m}_1 \equiv \frac{1}{n}\sum_{i=1}^{n} \widehat{m}_1(X_i) &= \sum_{i=1}^{n} W_i \widehat{m}_1(X_i) R_i \\ \Leftrightarrow \quad 0 &= \sum_{i=1}^{n} W_i R_i (\widehat{m}_1(X_i) - \bar{m}_1). \end{aligned} \tag{5.18}$$

Note that the equivalence follows from the fact that $1 = \sum_{i=1}^{n} \frac{R_i}{\widehat{\pi}(X_i)} = \sum_{i=1}^{n} W_i R_i$.

To determine the weight, we use the expression in equations (5.11) and (5.12). First, the Lagrangian multiplier will be chosen as

$$\widehat{\lambda}_{cal} = \mathsf{argmax}_{\lambda \in \Lambda_n} \sum_{i=1}^{n} \frac{R_i}{\widehat{\pi}(X_i)} \rho(\lambda(\widehat{m}_1(X_i) - \bar{m}_1)).$$

This is based on the IPW form of re-writing equation (5.12). One may be wondering why we use the IPW form in the above; technically speaking, without the IPW terms (replacing $\frac{R_i}{\widehat{\pi}(X_i)}$ by 1), this quantity will be estimating the same quantity. The major reason of writing it in this form is to focus on the weights for those $R_i = 1$. Later when defining the calibration estimator (equation (5.20)), one will see that we only need the weights for those $R_i = 1$, similar to the case of usual IPW estimator.

With this choice of λ, we then define the weights by generalizing (5.11):

$$W_{i,\rho,cal} = \frac{\widehat{\pi}^{-1}(X_i)\rho'(\widehat{\lambda}_{cal}(\widehat{m}_1(X_i) - \bar{m}_1))}{\sum_{j=1}^n \widehat{\pi}^{-1}(X_j)\rho'(\widehat{\lambda}_{cal}(\widehat{m}_1(X_j) - \bar{m}_1))}. \tag{5.19}$$

With this result, the *calibration* estimator is

$$\widehat{\theta}_{cal} = \sum_{i=1}^n R_i W_{i,\rho,cal} Y_i. \tag{5.20}$$

In [CY2014], they proved that $\widehat{\theta}_{cal}$ has asymptotic normality and achieve the semi-parametric efficient bound (assuming that $\widehat{m}_1(x)$ and $\widehat{\pi}(x)$ are using the correct model and converges at the $\sqrt{n}$ rate). The technical challenge of deriving the asymptotic normality is to bound $\widehat{\lambda}_{cal} = O_P(n^{-1/2})$; see the proof of Lemma 1 of [CY2014] . Note that similar to Section 5.2, the population quantity of $\widehat{\lambda}_{cal}$ is $\lambda_0 = 0$.

5.3.1 Robustness against propensity score mis-specification

A powerful feature of the calibration estimator is its robustness against the mis-specification of the propensity score. Suppose that $\widehat{\pi}(x)$ is mis-specified so it converges to $\pi_0(x) \neq P(R = 1|X = x)$. We will argue that $\widehat{\theta}_{cal}$ is still a consistent estimator.

To see this, denote $\tilde{W}(X_i) = \frac{\pi_0^{-1}(X_i)}{\sum_{j=1}^n \pi_0^{-1}(X_j)}$. It is not hard to see that when n is large, $W_{i,\rho,IPW} \approx \tilde{W}(X_i)$ because $\widehat{\lambda}_{cal} \xrightarrow{P} 0$ so the contribution from ρ' disappears (using the fact that $\rho'(0) = -1$).

As a result, we rewrite equation (5.20) as

$$\begin{aligned}
\widehat{\theta}_{cal} &= \sum_{i=1}^n R_i W_{i,\rho,cal} Y_i \\
&= \sum_{i=1}^n R_i W_{i,\rho,cal}(Y_i - \widehat{m}_1(X_i)) + \sum_{i=1}^n R_i W_{i,\rho,cal}\widehat{m}_1(X_i) \\
&\approx \sum_{i=1}^n R_i \tilde{W}(X_i)(Y_i - \widehat{m}_1(X_i)) + \sum_{i=1}^n R_i W_{i,\rho,cal}\widehat{m}_1(X_i) \\
&\approx \sum_{i=1}^n R_i \tilde{W}(X_i)(Y_i - \widehat{m}_1(X_i)) + \sum_{i=1}^n R_i W_{i,\rho,cal}\widehat{m}_1(X_i).
\end{aligned}$$

Using the assumption that $Y \perp R|X$ and the fact that $\widehat{m}_1(X_i) \approx m_1(X_i)$ (the outcome regression model is correct), one can easily see that the first quantity will be approaching

$$\sum_{i=1}^n R_i \tilde{W}(X_i)(Y_i - \widehat{m}_1(X_i)) \approx \mathbb{E}[R_i \pi_0^{-1}(X_i)(m_1(X_i) - \widehat{m}_1(X_i))] \approx 0.$$

For the second quantity, the calibration equation (5.18) implies that

$$\sum_{i=1}^n R_i W_{i,\rho,cal}\widehat{m}_1(X_i) = \bar{m}_1 = \frac{1}{n}\sum_{i=1}^n \widehat{m}_1(X_i) \approx \mathbb{E}(m_1(X)) = \theta.$$

As a result, $\widehat{\theta}_{cal}$ is still a consistent estimator as long as the outcome regression is correctly specified.

5.3.2 Oracle property when outcome regression may be mis-specified

Another powerful property of $\widehat{\theta}_{cal}$ is its oracle property when the outcome regression is mis-specified (propensity score is correctly specified). We denote the propensity score as $P(R=1|X=x)=\pi(x)=\pi(x;\beta_0)$, where β_0 is the true parameter of the propensity score.

In particular, suppose that we fit p distinct parametric models of the outcome regression, leading to

$$m_{1,1}(x;\gamma_1),\cdots m_{1,p}(x;\gamma_p).$$

The calibration equation (5.18) will be a system of p equations, i.e.,

$$\bar{m}_{1,\ell}\equiv\frac{1}{n}\sum_{i=1}^{n}m_{1,\ell}(X_i;\widehat{\gamma}_\ell)=\sum_{i=1}^{n}W_i m_{1,\ell}(X_i;\widehat{\gamma}_\ell)R_i \tag{5.21}$$

for $\ell=1,\cdots,p$ and $\widehat{\gamma}_\ell$ is the estimator of the parameter of the ℓ-th model. In this case, the Lagrangian multiplier will be a vector of p element and the weights can be defined in a similar manner as equation (5.19).

Let $\gamma_0=(\gamma_{1,0},\cdots,\gamma_{p,0})$ be the collection of parameter where $\widehat{\gamma}_\ell-\gamma_{\ell,0}=O_P(1/\sqrt{n})$. Namely, γ_0 is the quantity that the estimated parameter is converging to. We define the 'best linear predictor' of Y using these p models as $m^*(x;\gamma_0)$

$$m^*(x;\gamma_0)=c_0^*+\sum_{\ell=1}^{p}c_\ell^* m(x;\gamma_{\ell,0})$$

$$(c_0^*,\cdots,c_p^*)=\mathsf{argmin}_{(c_0,\cdots,c_p)}\mathbb{E}\left(\left(Y-c_0-\sum_{\ell=1}^{p}c_\ell m(X;\gamma_{\ell,0})\right)^2\right).$$

The above least square property imply

$$\mathbb{E}\left[m_\ell(X;\gamma_{\ell,0})(Y-m^*(X;\gamma_0))\right]=0 \tag{5.22}$$

for each $\ell=1,\cdots,p$.

Then Lemma 1 of [CY2014] showed that

$$\widehat{\theta}_{cal}-\theta=\frac{1}{n}\sum_{i=1}^{n}\left(\frac{R_i}{\pi(X_i;\beta_0)}(Y_i-m^*(X_i;\gamma_0))+(m^*(X_i;\gamma_0)-\theta)\right)+\Psi_n+o_P(1/\sqrt{n}), \tag{5.23}$$

where $\sqrt{n}\Psi_n$ is asymptotically normal (mean 0) and is from the estimation of the propensity score. The power of equation (5.23) is that the first asymptotic linear term follows the usual doubly-robust form and is centered at the outcome regression being specified as $m^*(X_i;\gamma_0)$, the best linear predictor.

Remark.

- Even if the outcome regression model is incorrect, the calibration estimator is still approximating the best linear predictor. Moreover, as long as one of the outcome regression model is correct (this would lead to other coefficients $c_\ell^*=0$ when ℓ is not the correct model), $m^*(x;\gamma_0)=m_1(x)=\mathbb{E}(Y|X=x,A=1)$ will be the correct outcome regression model and $\widehat{\theta}_{cal}$ achieves semi-parametric efficiency.
- From equation (5.23), we see that the asymptotic behavior of the outcome regression does not directly impact the asymptotic performance of $\widehat{\theta}_{cal}$. Here is a simple reason of why this occurs. Note that we can rewrite

$$\widehat{\theta}_{cal}=\sum_{i=1}^{n}R_i W_{i,\rho,cal}Y_i=\sum_{i=1}^{n}R_i\frac{\pi^{-1}(X_i;\widehat{\beta})\rho'(\widehat{\lambda}_{cal}^T(m(X_i;\widehat{\gamma}_0)-\bar{m}(\widehat{\gamma}_0)))}{\sum_{j=1}^{n}\pi^{-1}(X_j;\widehat{\beta})\rho'(\widehat{\lambda}_{cal}^T(m(X_j;\widehat{\gamma}_0)-\bar{m}(\widehat{\gamma}_0)))}Y_i.$$

We use the expression $m(X_i;\widehat{\gamma}_0) = (m_1(X_i;\widehat{\gamma}_1), \cdots, m_p(X_i;\widehat{\gamma}_p))^T \in \mathbb{R}^p$. When taking derivative with respect to γ and evaluate at the convergence point $(\beta_0, \gamma_0, \lambda_0 = 0)$, one would notice that there will be a λ_0 term multiplying everything because $\widehat{\theta}_{cal}$ depends on γ via

$$\rho'(\widehat{\lambda}_{cal}^T(m(X_j;\widehat{\gamma}_0) - \bar{m}(\widehat{\gamma}_0))).$$

When evaluating at $\lambda_0 = 0$, this term is gone. So the first derivative is asymptotically negligible, and hence the asymptotic behavior of the outcome regression does not impact the asymptotic behavior of $\widehat{\theta}_{cal}$.

- Some people may be wondering how does the best linear predictor comes in. It turns out that it will cancel out the contribution from $\widehat{\lambda}_{cal}$ in the asymptotic behavior of $\widehat{\theta}_{cal}$. The rough idea is that when taking derivative of $\widehat{\theta}_{cal}$ with respect to λ and evaluate at $(\beta_0, \gamma_0, \lambda_0 = 0)$, it will be approaching

$$\mathbb{E}\left[(m(X;\gamma_0) - \bar{m}(\gamma_0))(Y - m^*(X;\gamma_0))\right];$$

see the term A_1 in page 18 of the supplementary material of [CY2014]. By the least square property in equation (5.22), this term will be 0 so the contribution from $\widehat{\lambda}_{cal}$ will not be in the first order.