GMM with Auxiliary Information

Bryan S. Graham, UC - Berkeley & NBER

January 21, 2021

Let $\{Z_i\}_{i=1}^{N}$ be a simple random sample from a target population of interest. Let $\psi_2(Z, \theta_0)$ be a $K \times 1$ vector of known functions of Z indexed by the unknown parameter of interest, θ_0. We assume that the moment restriction

$$\mathbb{E}\left[\psi_2(Z, \theta_0)\right] = 0 \tag{1}$$

just-identifies θ_0 (i.e, $K = \dim(\psi(Z, \theta)) = \dim(\theta)$; accommodating overidentification is straightforward). If the random sample and (1) were the only pieces of information available to the econometrician, then she could compute the estimate

$$\frac{1}{N}\sum_{i=1}^{N}\psi_2\left(Z_i, \hat{\theta}\right) = 0,$$

with, for $\Omega_0 = \mathbb{E}\left[\psi_2\psi_2'\right]$ and $\Gamma_0 = \mathbb{E}\left[\frac{\partial\psi_2}{\partial\theta'}\right]$, a limit distribution of

$$\sqrt{N}\left(\hat{\theta} - \theta_0\right) \xrightarrow{D} \mathcal{N}\left(0, \left(\Gamma_0'\Omega_0^{-1}\Gamma_0\right)^{-1}\right).$$

In some situations, however, additional information can be incorporated into an estimation procedure. As an example, let $h(Z)$ be a vector of known functions of Z. Assume that the population mean of this vector, $\mathbb{E}[h(Z)]$, is known. This might occur if information on a subset of the elements of of Z is available from register data. For example voter registration files might contain basic demographic information and historical voting behavior on the entire universe of registered voters in a given state. From such information the researcher could construct functions $h(Z)$ with known mean. In a separate probability sample of registered voters she might observe additional variables, such variables might enter the moment vector $\psi_2(Z, \theta_0)$. Register data are increasingly available to researchers, perhaps most famously in Nordic countries. Associated with such data are a number of interesting identification and estimation questions. In this note I focus on how such data can be used to improve parameter precision, assuming that the target parameter is identified without the auxiliary information (Later in the course we will study how auxiliary information can additionally facilitate identification).

We encode our auxiliary information in the moment restriction

$$\mathbb{E}\left[\psi_1\left(Z\right)\right]=0. \tag{2}$$

In the example $\psi_1\left(Z\right)=h\left(Z\right)-\mathbb{E}\left[h\left(Z\right)\right]$; more generally we assume that $\psi_1\left(Z\right)$ is a known mean-zero function, not depending on any unknown parameters.

Together (1) and (2) imply the following overidentified system of equations

$$\mathbb{E}\left[\begin{array}{c}\psi_1\left(Z\right)\\ \psi_2\left(Z,\theta_0\right)\end{array}\right]=0. \tag{3}$$

Applying an efficient GMM estimator to this system, for example the textbook two-step GMM estimator of Hansen (1982), yields an estimate with a limit distribution of

$$\sqrt{N}\left(\hat{\theta}-\theta_0\right)\xrightarrow{D}\mathcal{N}\left(0,\left(\Gamma_0'\left(\Omega_0-C_0\mathcal{I}_0^{-1}C_0'\right)^{-1}\Gamma_0\right)^{-1}\right)$$

where $C_0=\mathbb{E}\left[\psi_2\psi_1'\right]$ and $\mathcal{I}_0=\mathbb{E}\left[\psi_1\psi_1'\right]$. If $C_0\neq 0$, then incorporating the auxiliary moment (2) into our estimation procedure increases asymptotic precision relative to GMM based upon (1) alone.

Anatomy of the efficiency gain

Observe that the limit distribution of our $\sqrt{N}\left(\hat{\theta}-\theta_0\right)$ coincides with that of the method-of-moments estimate based upon the following just-identified system of equations:

$$\mathbb{E}\left[\psi_2^*\left(Z,\theta_0\right)\right]=\mathbb{E}\left[\psi_2\left(Z,\theta_0\right)-C_0\mathcal{I}_0^{-1}\psi_1\left(Z\right)\right]=0. \tag{4}$$

The second component of $\psi_2^*\left(Z,\theta_0\right)$ coincides with the multivariate linear predictor:

$$\begin{aligned}\mathbb{E}^*\left[\left.\psi_2\left(Z,\theta_0\right)\right|\psi_1\left(Z\right)\right]&=\mathbb{E}\left[\psi_2\psi_1'\right]\times\mathbb{E}\left[\psi_2\psi_1'\right]^{-1}\psi_1\left(Z\right)\\&=C_0\mathcal{I}_0^{-1}\psi_1\left(Z\right),\end{aligned}$$

hence $\psi_2^*(Z,\theta_0)$ corresponds to the error associated with the (linear) projection of the identifying moment, $\psi_2(Z,\theta_0)$, onto the auxiliary one, $\psi_1(Z)$:

$$\psi_2^*(Z,\theta_0) = \psi_2(Z,\theta_0) - \mathbb{E}^*\left[\left.\psi_2(Z,\theta_0)\right|\psi_1(Z)\right].$$

Since projections are norm-reducing[1] we have that $\mathbb{V}(\psi_2^*(Z,\theta_0)) \leq \mathbb{V}(\psi_2(Z,\theta_0))$ and, consequently, greater asymptotic precision on account of utilizing the auxiliary moments.

There are a variety of intuitions for this result, the first is the one just described: we use the auxiliary information to reduce sampling variability in the identifying moment. This intuition gives some insight into what types of auxiliary moments are likely to be especially useful (see Imbens & Lancaster (1994) for useful discussion on this point).

A second intuition comes from the theory of efficient estimation of expectations (e.g., Imbens, 1997; Brown & Newey, 1998). Let$\hat{\mathcal{I}} = \frac{1}{N}\sum_{i=1}^N \psi_1(Z_i)\psi_1(Z_i)'$, $\hat{C}(\theta) = \left[\frac{1}{N}\sum_{i=1}^N \psi_2(Z_i,\theta)\psi_1(Z_i)'\right]$ and $\bar{\psi}_1 = \frac{1}{N}\sum_{i=1}^N \psi_1(Z_i)$. Let $\hat{\psi}_2^*(Z_i,\theta) = \psi_2(Z_i,\theta) - \hat{C}(\theta)\hat{\mathcal{I}}^{-1}\psi_1(Z_i)$ be an estimate of the modified moment introduced in (4) above. Next consider the estimator which chooses $\hat{\theta}$ as the solution to

$$\frac{1}{N}\sum_{i=1}^N \hat{\psi}_2^*\left(Z_i,\hat{\theta}\right) = 0. \tag{5}$$

Note that as we solve for $\hat{\theta}$ we continually recompute $\hat{C}(\theta)$ as θ various over the course of optimization. This feature of estimation also arises in the so-called continuously-updated GMM estimator (CUE) of Hansen et al. (1996); in fact there is a close connection between CUE and the solution to (5). The CUE estimator based upon (3) equals

$$\hat{\theta}_{\text{CUE}} = \arg\min_{\theta}\left[\frac{1}{N}\sum_{i=1}^N\left[\begin{array}{cc}\psi_1(Z_i)' & \psi_2(Z_i,\theta)'\end{array}\right]\right]\left[\begin{array}{cc}\hat{\mathcal{I}} & \hat{C}(\theta)' \\ \hat{C}(\theta) & \hat{\Omega}(\theta)\end{array}\right]^{-1}\left[\frac{1}{N}\sum_{i=1}^N\left[\begin{array}{c}\psi_1(Z_i) \\ \psi_2(Z_i,\theta)\end{array}\right]\right].$$

This differs from Hansen's (1982) two-step GMM in that the optimal weight matrix is "continuously-updated" as θ varies when optimizing. Some tedious matrix algebra shows that this CUE criterion function coincides with

$$\bar{\psi}_1'\hat{\mathcal{I}}^{-1}\bar{\psi}_1 + \left[\frac{1}{N}\sum_{i=1}^N \hat{\psi}_2^*(Z_i,\theta)\right]'\left[\hat{\Omega}(\theta) - \hat{C}(\theta)\hat{\mathcal{I}}^{-1}\hat{C}(\theta)'\right]^{-1}\left[\frac{1}{N}\sum_{i=1}^N \hat{\psi}_2^*(Z_i,\theta)\right],$$

but since the first term in the expression above doesn't vary with θ, while in the second $\dim\left(\hat{\psi}_2^*(Z_i,\theta)\right) = \dim(\theta) = K$, the solution $\hat{\theta}_{\text{CUE}}$ is identical to that of (5) above. Hence our estimator coincides with CUE as applied to the system of moments (3) above.

[1] See the Projection Theorem notes for a review.

In turns out there is a further recasting of this optimization problem. One that has computational advantages and provides considerable insight into what drives the efficiency gains associated with the auxiliary moments. Interchanging the order of summation we have that (see the second and third equalities below):

$$\begin{aligned}
\frac{1}{N}\sum_{i=1}^{N}\hat{\psi}_2^{*}\left(Z_i,\theta\right) &= \frac{1}{N}\sum_{i=1}^{N}\left[\psi_2\left(Z_i,\theta\right)-\hat{C}\left(\theta\right)\hat{\mathcal{I}}^{-1}\psi_1\left(Z_i\right)\right] \\
&= \frac{1}{N}\sum_{i=1}^{N}\left[\psi_2\left(Z_i,\theta\right)-\left[\frac{1}{N}\sum_{j=1}^{N}\psi_2\left(Z_j,\theta\right)\psi_1\left(Z_j\right)'\right]\times\hat{\mathcal{I}}^{-1}\psi_1\left(Z_i\right)\right] \\
&= \frac{1}{N}\sum_{i=1}^{N}\left[\psi_2\left(Z_i,\theta\right)-\psi_2\left(Z_i,\theta\right)\psi_1\left(Z_j\right)'\hat{\mathcal{I}}^{-1}\bar{\psi}_1\right] \\
&= \sum_{i=1}^{N}\left[\psi_2\left(Z_i,\theta\right)\frac{1}{N}\left\{1-\psi_1\left(Z_i\right)'\hat{\mathcal{I}}^{-1}\bar{\psi}_1\right\}\right] \\
&= \sum_{i=1}^{N}\hat{\pi}_i\psi_2\left(Z_i,\theta\right) \qquad (6)
\end{aligned}$$

where we define, for $i=1,\ldots,N$, the probability weights

$$\hat{\pi}_i=\frac{1}{N}\left\{1-\psi_1\left(Z_i\right)'\hat{\mathcal{I}}^{-1}\bar{\psi}_1\right\}.$$

Representation (6) shows that our estimator can be viewed as replacing the empirical measure of random sample (i.e., the measure which places weight $1/N$ on each observation) with one which places weights $\{\hat{\pi}_i\}_{i=1}^{N}$ on the sample points. Instead of choosing $\hat{\theta}$ to set the sample mean of $\psi_2\left(Z_i,\theta\right)$ equal to zero, we choose it to set a weighted average of these moments equal to zero. In the population we know that $\mathbb{E}\left[\psi_1\left(Z_i\right)\right]=0$, however in the sample it is unlikely that $\frac{1}{N}\sum_{i=1}^{N}\psi_1\left(Z_i\right)=0$. However it is the case that the weighted average

$$\begin{aligned}
\sum_{i=1}^{N}\hat{\pi}_i\psi_1\left(Z_i\right) &= \frac{1}{N}\sum_{i=1}^{N}\left\{\psi_1\left(Z_i\right)-\psi_1\left(Z_i\right)\psi_1\left(Z_i\right)'\hat{\mathcal{I}}^{-1}\bar{\psi}_1\right\}. \\
&= \bar{\psi}_1-\hat{\mathcal{I}}\hat{\mathcal{I}}^{-1}\bar{\psi}_1=0,
\end{aligned}$$

is exactly mean zero.

This formulation provides a second intuition for how the auxiliary moments improve precision: we use them to construct a more efficient estimate of the distribution function of the data. In a just-identified GMM problem, the efficient “estimate” of the distribution function for the data is simply the empirical measure. However the addition of the auxiliary moments

(2) adds extra information that can be used to construct a more efficient estimate of the distribution function (cf., Imbens, 1997). We can then use this more efficient estimate to find the solution to the sample analog of the identifying moment (1). We'll put some of these ideas on firmer footing after we study semiparametric efficiency in more detail later in the course.

One advantage of representation (6) is that the weights can be constructed pre-estimation (and have a closed-form). Once these weights are constructed any M-estimation program which accepts user-inputed weights can be used to compute $\hat{\theta}$. Incorporating the auxiliary information (2) improves efficiency and involves negligible additional effort in terms of computation.

Modified CUE weights

In practice we might want to use a slightly different set of probability weights for computation. Observe that

$$\sum_{i=1}^{N}\hat{\pi}_i = 1-\bar{\psi}_1'\hat{\mathcal{I}}^{-1}\bar{\psi}_1.$$

Next observe that

$$N\bar{\psi}_1'\hat{\mathcal{I}}^{-1}\bar{\psi}_1 \xrightarrow{D} \chi_J^2$$

such that, for N large enough,

$$\begin{aligned}\mathbb{E}\left[\bar{\psi}_1'\hat{\mathcal{I}}^{-1}\bar{\psi}_1\right] &\approx \frac{J}{N}\\ \mathbb{V}\left(\bar{\psi}_1'\hat{\mathcal{I}}^{-1}\bar{\psi}_1\right) &\approx \frac{2J}{N^2}.\end{aligned}$$

Together these results imply that

$$\sum_{i=1}^{N}\hat{\pi}_i = 1+o_p\left(N^{-\delta}\right),\ \delta\in[0,1).$$

Distributions integrate/sum to one, so we would expect a consistent distribution function estimate to also consist of probabilities summing to one. This is true asymptotically, but need not hold in finite samples. An easy fix is to normalize the weights:

$$\tilde{\pi}_i = \frac{1\text{-}\psi_1\left(Z_i\right)\text{'}\hat{\mathcal{I}}^{-1}\bar{\psi}_1}{\sum_{j=1}^{N}1-\psi_1\left(Z_j\right)'\hat{\mathcal{I}}^{-1}\bar{\psi}_1} \tag{7}$$

such that they sum to one by construction (note that many weighted M-estimation programs will automatically normalize the weights). In practice basing estimation on $\{\tilde{\pi}\}_{i=1}^{N}$instead of $\{\hat{\pi}_i\}_{i=1}^{N}$ leads to better finite sample performance.

Given the prior knowledge that the population mean of $\psi_1(Z_i)$ is mean zero, if $\hat{\mathcal{I}}^{-1}\bar{\psi}_1 > 0$, than large realizations of $\psi_1(Z_i)$ are over-represented in the sample. The weights (7) effectively down-weight units with $\psi_1(Z_i) > 0$, while up-weighting those with $\psi_1(Z_i) < 0$(cf., Back & Brown, 1993). One unattractive feature of the weights (7) is that they may be negative. If the sample is large, and is also a genuine random sample from the population of interest, this will likely not be a major problem; but in small samples – or biased ones – it can be. Imposing positivity on the probability weights can be done using generalized empirical likelihood methods (e.g., Newey & Smith, 2004), but we will not pursue this extension here.

Examples

Linear regression

Let $Z = (W', X', Y)$. The researcher is interested in the linear regression of Y onto X and W. From register data she knows the mean of Y in J subpopulations defined in terms of X. For example information on earnings, Y, and basic demographic attributes, X (age, gender, ethnicity, place-of-birth, etc.), may be available from the US Census. In the National Longitudinal Survey of Youth (NLSY) additional covariates, W (AFQT, parental education, etc.) may be observed in addition to X and Y. The researcher, using her NLSY sample, bases estimation on the pair of moment functions

$$\psi_2(Z,\theta) = (Y - X'\beta - W'\gamma)\begin{pmatrix} X \\ W \end{pmatrix}$$

$$\psi_1(Z) = \begin{bmatrix} \mathbf{1}(X \in \mathbb{X}_1)(Y - \mu_1) \\ \vdots \\ \mathbf{1}(X \in \mathbb{X}_J)(Y - \mu_J) \end{bmatrix},$$

where $\{\mu_j\}_{j=1}^{J}$are the known conditional means of Y garnished from the register data.

Conditional likelihood

In some likelihood settings additional structure can be exploited to construct more efficient estimators. Such settings are studied in an elegant paper by Imbens & Lancaster (1994).

As above assume the researcher is interested in the regression of Y onto W and X, but now assume that Y is binary and the regression takes the probit form. Furthermore assume that this regression function is correctly specified. In this setting we can work with the pair of moment functions

$$\psi_2(Z,\theta) = \frac{(Y - \Phi(X'\beta + W'\gamma))}{\Phi(X'\beta + W'\gamma)\left[1 - \Phi(X'\beta + W'\gamma)\right]}\phi(X'\beta + W'\gamma)\begin{bmatrix} X \\ W \end{bmatrix}$$

$$\psi_1(Z,\theta) = \begin{bmatrix} \mathbf{1}(X \in \mathbb{X}_1)(\Phi(X'\beta + W'\gamma) - \mu_J) \\ \vdots \\ \mathbf{1}(X \in \mathbb{X}_J)(\Phi(X'\beta + W'\gamma) - \mu_J) \end{bmatrix}.$$

The first moment is just the score vector associated with a probit log-likelihood function. The second vector is similar to the one in our first example, but instead of Y entering we use the (correctly specified) conditional mean function of Y given W and X. This uses more structure and is generally more efficient. Note that now θ also enters the auxiliary moment.

Average Treatment Effect (ATE)

A final example comes from the theory of covariate adjustment (see Imbens & Wooldridge (2009) for a review). Let Y_1 be some potential outcome under treatment and Y_0 the corresponding potential outcome under control. Let $D = 1$ if a unit is treated and zero otherwise. Finally let $e_0(X) = \Pr(D = 1|X)$ be a *known* propensity score. Here X correspond to a vector of pre-treatment assignment/conditioning variables.

The conditional probability of treatment would be known, for example, in the context of a stratified random experiment; or in settings where lotteries are used to assign treatment (or even in a setting where program participation is known from administrative/register data and the researcher is willing to make an unconfoundedness assumption).

Say we wish to estimate $\alpha_1 = \mathbb{E}[Y_1]$ the mean potential outcome under treatment. Assuming unconfoundedness such that $(Y_0, Y_1) \perp D|X = x$ for all $x \in \mathbb{X}$ and that $e_0(X)$ is bounded away from zero, the inverse probability weighted (IPW) moment restriction

$$\mathbb{E}[\psi_2(Z,\theta_0)] = \mathbb{E}\left[\frac{D}{e_0(X)}(Y_1 - \alpha_{10})\right] = 0. \tag{8}$$

identifies α_{10}(see, for example, Hirano et al. (2003) or Wooldridge (2007) among others). Given that the propensity score is known, we can augment our problem with the additional restriction

$$\mathbb{E}\left[\psi_1(Z)\right] = \mathbb{E}\left[\left(\frac{D}{e_0(X)} - 1\right) t(X)\right] = 0, \tag{9}$$

with $t(X)$ a vector of known functions of X. Using the general notation established above we have, using iterated expectations,

$$\mathcal{I}_0 = \mathbb{E}\left[\psi_1(Z)\psi_1(Z)'\right] = \mathbb{E}\left[\frac{1-e_0(X)}{e_0(X)} t(X)t(X)'\right]$$

and also, using conditional independence of treatment assignment,

$$\begin{aligned} C_0 = \mathbb{E}\left[\psi_2(Z,\beta)\psi_1(Z)'\right] &= \mathbb{E}\left[\frac{DY_1}{e_0(X)}\left(\frac{D}{e_0(X)} - 1\right) t(X)'\right] \\ &= \mathbb{E}\left[\left(\frac{1-e_0(X)}{e_0(X)^2}\right) DY_1 t(X)'\right] \\ &= \mathbb{E}\left[\frac{1-e_0(X)}{e_0(X)} q_1(X)t(X)'\right], \end{aligned}$$

with $q_1(X) = \mathbb{E}\left[Y_1 | X, D=1\right] = \mathbb{E}\left[Y_1 | X\right]$. The linear projection of the identifying IPW moment (8) onto the auxiliary propensity score moment (9) is therefore

$$\begin{aligned} \mathbb{E}^*\left[\psi_2(Z,\theta_0) | \psi_1(Z)\right] =& C_0\mathcal{I}_0^{-1}\psi_1(Z) \\ =& \mathbb{E}\left[\frac{1-e_0(X)}{e_0(X)} q_1(X)t(X)'\right] \\ & \times \mathbb{E}\left[\frac{1-e_0(X)}{e_0(X)} t(X)t(X)'\right]^{-1}\left(\frac{D}{e_0(X)} - 1\right) t(X) \\ =& \frac{q_1^*(X)}{e_0(X)}\left(D - e_0(X)\right) \end{aligned}$$

where

$$q_1^*(X) = \mathbb{E}\left[\frac{1-e_0(X)}{e_0(X)} q_1(X)t(X)'\right] \times \mathbb{E}\left[\frac{1-e_0(X)}{e_0(X)} t(X)t(X)'\right]^{-1} t(X)$$

equals a *weighted* projection of $q_1(X)$ onto $t(X)$. If the set of basis functions included in $t(X)$ is rich enough it will typically be the case that $q_1^*(X) \approx q_1(X)$, but the two objects will not formally coincide unless $q_1(X)$ - the true conditional expectation function of Y_1given X – can be expressed as a linear combination of $t(X)$ (i.e., unless $q_1(x) = \Pi_0 t(x)$ for all

$x \in \mathbb{X}$). Regardless, we have that

$$\sqrt{N}\left(\hat{\alpha}_1-\alpha_{10}\right)=\frac{1}{N}\sum_{i=1}^{N}\frac{D_i}{e_0\left(X_i\right)}\left(Y_1-\alpha_{10}\right)-\frac{q_1^*\left(X_i\right)}{e_0\left(X_i\right)}\left(D_i-e_0\left(X_i\right)\right)+o_p\left(1\right). \qquad (10)$$

Our augmented GMM estimator is asymptotically equivalent to a particular Augmented Inverse Probability Weighting (AIPW) estimator (cf., Robins et al., 1994). Graham et al. (2012) exploit this connection to develop estimators for a class of missing data problems with very good properties.

1 Further reading

Imbens & Lancaster (1994) and Hailong & Schmidt (1999) and are two references quite closely connected to the material covered here. The general underlying theory is that of efficient estimation of expectations, which is treated elegantly by Brown & Newey (1998). The particular problem of incorporating auxiliary information belongs to a general class of data combinations problems. Ridder & Moffitt (2007) survey the econometrics of such problems (see also Graham et al. (2016)). The discussion of CUE and distribution function estimation also connects to the literature on generalized empirical likelihood (see, especially, Imbens (1997) and Newey & Smith (2004)).

References

Back, K. & Brown, D. P. (1993). Implied probabilities in gmm estimators. *Econometrica*, 61(4), 971 – 75.

Brown, B. W. & Newey, W. K. (1998). Efficient semiparametric estimation of expectations. *Econometrica*, 66(2), 453 – 464.

Graham, B. S., Pinto, C., & Egel, D. (2012). Inverse probability tilting for moment condition models with missing data. *Review of Economic Studies*, 79(3), 1053 – 1079.

Graham, B. S., Pinto, C., & Egel, D. (2016). Efficient estimation of data combination models by the method of auxiliary-to-study tilting (ast). *Journal of Business and Economic Statistics*, 31(2), 288 – 301.

Hailong, Q. & Schmidt, P. (1999). Improved instrumental variables and generalized method of moments estimators. *Journal of Econometrics*, 91(1), 145 – 169.

Hansen, L. P. (1982). Large sample properties of generalized method of moments estimators. *Econometrica*, 50(4), 1029 – 1054.

Hansen, L. P., Heaton, J., & Yaron, A. (1996). Finite-sample properties of some alternative gmm estimators. *Journal of Business and Economic Statistics*, 14(3), 262 – 280.

Hirano, K., Imbens, G. W., & Ridder, G. (2003). Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica*, 71(4), 1161 – 1189.

Imbens, G. W. (1997). One-step estimators for over-identified generalized method of moments models. *Review of Economic Studies*, 64(3), 359 – 383.

Imbens, G. W. & Lancaster, T. (1994). Combining micro and macro data in microeconometric models. *Review of Economic Studies*, 61(4), 665 – 680.

Imbens, G. W. & Wooldridge, J. M. (2009). Recent developments in the econometrics of program evaluation. *Journal of Economic Literature*, 47(1), 5 – 86.

Newey, W. K. & Smith, R. J. (2004). Higher order properties of gmm and generalized empirical likelihood estimators. *Econometrica*, 72(1), 219 – 255.

Ridder, G. & Moffitt, R. (2007). *Handbook of Econometrics*, volume 6B, chapter The econometrics of data combination, (pp. 5469 – 5547). North-Holland.

Robins, J. M., Rotnitzky, A., & Zhao, L. P. (1994). Estimation of regression coefficients when some regressors are not always observed. *Journal of American Statistical Association*, 89(427), 846 – 866.

Wooldridge, J. M. (2007). Inverse probability weighted estimation for general missing data problems. *Journal of Econometrics*, 141(2), 1281 – 1301.