Simple nonlinear imputation

December 2020

1 Introduction

In this paper we consider estimation in the linear index model where given a function h, the relationship between the LHS variable (Y_i) and RHS variables ($X_i \in \mathbb{R}, \mathbf{Z_i} \in 1 \times \mathbb{R}^k$) is described by

$$E[Y_i|X_i, \mathbf{Z_i}, M_i] = h(\alpha X_i + \mathbf{Z_i}\beta), \tag{1}$$

and the variable X_i has missing values whenever M_i, the missingness indicator is 1. We are exploring the properties of a simple imputation GMM estimator under the further assumption that the conditional distribution of X_i given $\mathbf{Z}_i$ is independent of M_i:

$$P[X_i < t|\mathbf{Z}_i = \mathbf{z}, M_i = m] = P[X_i < t|\mathbf{Z}_i = \mathbf{z}, M_i = m] \ \forall t \in \mathbb{R}. \tag{2}$$

Example 1 (Probit model). The researcher has collected information for a binary outcome variable Y and the $\mathbf{Z}$ for a large sample, but only has information about a control variable X in a subsample.

$$\begin{aligned} Y_i &= \mathbf{1}[\alpha X_i + \mathbf{Z}_i\beta > \epsilon_i] \\ \epsilon_i &\sim N[0, 1] \end{aligned}$$

Example 2 (Baseline censored linear model). We have that

$$\begin{aligned} Y_i^* &= \alpha X_i + \mathbf{Z}_i\beta - \epsilon_i \\ Y_i &= \mathbf{1}[Y_i^* > 0]Y_i^* \\ \epsilon_i &\sim N[0,1] \end{aligned}$$

In both examples, we further assume that missingness is determined by the observable vector $\mathbf{Z}_i$.

LITERATURE REVIEW, SELLING MISSING HERE

2 A simple imputation GMM estimator

2.1 Model assumptions and definition

We collect our model assumptions for identification below. We denote the conditional pdf of X_i given $\mathbf{Z_i}, M_i$ as $f_{x|z,m}$.

Assumption 1 (Model)**.** *We assume that*

1. $E[Y_i|X_i = x, \mathbf{Z}_i = z, M_i = m] = h(\alpha x + z\beta)$*, where* h *is a known, strictly increasing and differentiable function,*
2. $f_{x|z,m}(x, z, m) = f_{x|z}(x, z)$*,*
3. *the support of the random variables* $X_i, \mathbf{Z}_i$ *does not lie in a proper subspace of* $\mathbb{R}^k$*,*
4. $P[M_i = 1|X_i, \mathbf{Z}_i] < 1 \ X_i, \mathbf{Z}_i - a.s.$

The first two conditions in Assumption 1 are exclusion restrictions and restrict the missingness structure, but they are substantially weaker than the often used missing-at-random (MAR) assumption. Since we assume that h is known, the first condition is also a functional form assumption in practice. The strict monotonicity of h is a simple condition that together with the sufficient variation required ensures identification of the coefficient vector. Under Assumption 1, the true (α, β) uniquely satisfies

$$E[\tilde{g}(y, z, x, m; a, b)] = E\begin{bmatrix} (1 - M_i)X_i(Y_i - h(\alpha X_i + Z_i\beta)) \\ (1 - M_i)Z_i(Y_i - h(\alpha X_i + Z_i\beta)) \end{bmatrix} = 0 \ a.s., \quad (3)$$

where the expected value of $\tilde{g}$ is the population moment that is the basis of a consistent GMM estimator using the fully observed part of the sample. The first four arguments of $\tilde{g}$ are from the support of the corresponding random variables in our model. The fifth and sixth arguments are elements from the (finite dimensional) parameter space of α, β.

In addition to these identifying moments we also add imputation moments to our GMM estimator in order to increase efficiency. Define the vector-valued function g as

$$g(y, z, x, m; a, b; E[y|z]) = \begin{bmatrix} (1-m)x(y - h(ax + zb)) \\ (1-m)z(y - h(ax + zb)) \\ mz(y - E[y|z](z; a, b)) \end{bmatrix}. \tag{4}$$

The function g has the same same first arguments as $\tilde{g}$, while the last argument is a function that is supposed to estimate the conditional expectation of Y_i given $\mathbf{Z}_i = z$. The function $E[y|z](.; \alpha, \beta) : \mathbf{R}^k \to \mathbf{R}$ at the true values is defined by

$$E[y|z](z, \alpha, \beta) = E[Y_i|\mathbf{Z}_i = z, M_i = 1].$$

Using Assumption 1,

$$E[Y_i|\mathbf{Z_i} = z, M_i = 1] = E[Y_i|\mathbf{Z_i} = z, M_i = 0] = \int h(\alpha x + z\beta) f_{x|z}(x, z) dx, \tag{5}$$

so the definition of the infinite dimensional nuisance parameter becomes

$$E[y|z](z; a, b) = \int h(ax + zb) f_{x|z}(x, z) dx. \tag{6}$$

Clearly, this function is identified, given the second exclusion restriction in Assumption 1. Moreover, note that

$$E[g(Y_i, \mathbf{Z}_i, X_i, M_i; a, b; E[y|z])] = 0 \iff a = \alpha, b = \beta. \tag{7}$$

Define the sample analogues (for a sample of size n)

$$\hat{g}(a,b;\hat{E}[y|z]) = n^{-1}\sum_{i=1}^{n} g(y_i,z_i,x_i,m_i;a,b;\hat{E}[y|z]) = \tag{8}$$
$$= n^{-1}\sum_{i=1}^{n}\begin{bmatrix} (1-m_i)x_i(y_i - h(ax_i+z_ib)) \\ (1-m_i)z_i(y_i - h(ax_i+z_ib)) \\ m_iz_i(y_i - \hat{E}[y|z](z_i;a,b)) \end{bmatrix},$$

where $\hat{E}_{y|z}$ is an estimator of the conditional expectation $E(y|z)$. At the end of this section we give the Nadaraya-Watson estimator as a specific example for a viable $\hat{E}(y|z)$. The imputation GMM estimator is minimizing

$$\hat{Q}_n(a,b) = \hat{g}(a,b;\hat{E}[y|z])'\hat{W}\hat{g}(a,b;\hat{E}[y|z]) \tag{9}$$

with respect to a, b, where the $\hat{W}$ is a symmetric weighting matrix such that for some W (positive definite)

$$\hat{W} \xrightarrow{p} W. \tag{10}$$

The fully observed GMM estimator that is based on the moment

$$\hat{\tilde{g}}(a,b) = n^{-1}\sum_{i=1}^{n} g(y_i,z_i,x_i,m_i;a,b) = \tag{11}$$
$$= n^{-1}\sum_{i=1}^{n}\begin{bmatrix} (1-m_i)x_i(y_i - h(ax_i+z_ib)) \\ (1-m_i)z_i(y_i - h(ax_i+z_ib)) \end{bmatrix},$$

and defined as the M-estimator minimizing

$$\hat{\hat{Q}}_n(a,b) = \hat{\tilde{g}}(a,b)'\hat{\tilde{W}}\hat{\tilde{g}}(a,b) \tag{12}$$

with respect to a, b, where the $\hat{\tilde{W}}$ is a symmetric weighting matrix such that for some $\tilde{W}$ (positive definite)

$$\hat{\tilde{W}} \xrightarrow{p} \tilde{W}. \tag{13}$$

2.2 Asymptotic properties

In the following arguments we closely follow Ichimura and Newey (2015) and Chernozhukov et al. (2018). We denote the estimator as a random variable by $\hat{\theta}_n$, while the true value $\theta = (\alpha, \beta)$. We also introduce the notation $\hat{G}(\theta; \hat{E}[y|z])$ for the derivative of $\hat{g}$ with respect to the parameter vector. Correspondingly, the derivative of $E[g]$ is denoted by G.

Assumption 2. *We place several smoothness assumptions on the structural functions.*

- $f_{x|z}(x, z)$ *is continuously differentiable and uniformly bounded,*
- h *is continuously differentiable with bounded derivatives.*

Further usual assumptions:

1. *random sampling,*
2. $\hat{E}[y|z](a, b) \xrightarrow{p} E[y|z](a, b)$ *uniformly over* z *and* $(a, b) \in A \times B$*,*
3. $\alpha, \beta \in A \times B$*, a compact set,*
4. $G'WG$ *is a.s. an invertible matrix.*

Writing up the first order Taylor expansion of $\hat{g}$ around θ gives

$$\begin{aligned} 0 &= \hat{G}'(\hat{\theta}_n; \hat{E}[y|z])\hat{W}\hat{g}(\hat{\theta}_n; \hat{E}[y|z]) = \\ &= \hat{G}'(\hat{\theta}_n; \hat{E}[y|z])\hat{W}\hat{g}(\theta; \hat{E}[y|z]) + \hat{G}'(\hat{\theta}_n; \hat{E}[y|z])\hat{W}\hat{G}(\bar{\theta}_n; \hat{E}[y|z])(\hat{\theta}_n - \theta) = \\ &= \hat{G}'\hat{W}\hat{g}_0 + \hat{G}'\hat{W}\bar{G}(\hat{\theta}_n - \theta). \end{aligned} \tag{14}$$

For legibility, we abbreviated the notation for the various matrices from the second row. Here $\bar{\theta}$ is a vector of convex combinations of θ and $\hat{\theta}_n$, but this value can (and generally should) be different for different rows of $\hat{G}$. In this sense we abuse the notation for $\hat{G}$ somewhat when we say it is evaluated at $\bar{\theta}$. We also introduce

$$\hat{g}_0 = \hat{g}(\theta; \hat{E}[y|z]), \tag{15}$$

$$g_0 = g(Y_i, X_i, \mathbf{Z}_i, M_i; \theta; E[y|z]). \tag{16}$$

Given Assumption 2, we can prove that $\hat{\theta}_n$ consistently estimates θ, as $||\hat{G}'\hat{W}\bar{G}||$ is going to be bounded and $\hat{g}_0 \to E[g_0] = 0$ with probability approaching 1. This

in turn yields

$$\begin{aligned}\hat{\theta}_n - \theta &= -(\hat{G}'\hat{W}\bar{G})^{-1}\hat{G}'\hat{W}\hat{g}_0 \\ &= -(G'WG)^{-1}G'W\hat{g}_0 + o_p(\hat{g}_0).\end{aligned} \tag{17}$$

The following is a consequence of Theorem 7 of Ichimura and Newey (2015), and our calculations in the previous subsection.

Proposition 1. *Under Assumption 1-2, if* $||E[y|z](z_i, a, b) - \hat{E}[y|z](z_i, a, b)|| = o_p(n^{-1/4})$ $Z_i - a.s.$ *and for every* a, b*, then*

$$\sqrt{n}(\theta_n - \theta) \xrightarrow{d} N[0, (G'WG)^{-1}G'W\Omega WG(G'WG)^{-1}],$$

with $\Omega = E[g_0'g_0]$.

We also note that the fully observed GMM estimator has the same properties under our assumptions.

Convergence rate of $\hat{E}[y|z]$ *for the Nadaraya-Watson estimator* We are going to estimate the conditional expectation by a Nadaraya-Watson type estimator.

$$\hat{E}[y|z](y_i, z_i; a, b) = \frac{\sum_j K[H^{-1}(z_i - z_j)]h(ax_j + bz_i)}{\sum_j K[H^{-1}(z_i - z_j)]}. \tag{18}$$

For the sake of simplicity, we will assume that H is a diagonal matrix with positive diagonal entries. Let us have the entry that decreases to zero at the slowest rate denoted by h_{max}, moreover let us write $\prod h_k = h$.

Assumption 3. *Our estimation assumptions:*

1. $h_{max} \to 0$
2. $nh_{max}^{k_z} \to \infty$
3. *K is a Parzen-Rosenblatt kernel (second order)*
4. *$Supp(Z_i)$ is compact, with the strong pdf assumption*
5. *the pdf for Z_i is twice differentiable*
6. *the conditional distribution function $f_{x|z}(x, z)$ is twice differentiable (with bounded Hessian)*

Some of these assumptions are stronger than necessary (notably, conditions number 2 and number 4). We conclude that as long as the bandwidth h_{max} is $o(n^{-1/4})$, we only have to worry about the contribution of the variances, under the restriction that the $\hat{E}[y|z]$ converges uniformly to the conditional expectation as a function, which gives the restriction that $nh \to \infty$. For these conditions we need that the $k < 4$ that is involved in the calculations of the conditional expectations. In addition, we note that discrete z_i-s are allowed with simple modifications of the estimator and theory.

Corollary 1. *If $\hat{E}[y|z]$ is the Nadaraya-Watson estimator with $k < 4$, under Assumption 1-3 the imputation GMM estimator $\hat{\theta}_n$ is such that*

$$\sqrt{n}(\theta_n - \theta) \xrightarrow{d} N[0, (G'WG)^{-1}G'W\Omega WG(G'WG)^{-1}],$$

with $\Omega = E[g_0'g_0]$.

The corollary suggests that the dimensionality of $\mathbf{Z}_i$ included in the is crucial for imputation to work. There are two reasons to include an (always observed) RHS variable into the imputation moments:

- Weakening of the missing-at-random assumption: we think the variable is related to missingness,
- Predictive power for X_i: observing the variable gives information about the missing RHS variable

Even if the second point would not warrant an inclusion of a particular element of $\mathbf{Z}_i$ into the group of conditioning variables in $E[y|z]$, if we think that it may be related to missingness, it needs to be included in the estimator.

2.3 The role of the weighting matrix and efficiency

Our goal is to minimize the Mean Squared-Error (MSE). It can be calculated as the expected value of the diagonal of the matrix

$$\begin{aligned}(\hat{\theta}_n - \theta)(\hat{\theta}_n - \theta)' \\ = ((G'WG)^{-1}G'W\hat{g}_0 + o_p(\hat{g}_0))((G'WG)^{-1}G'W\hat{g}_0 + o_p(\hat{g}_0))' = \\ = (G'WG)^{-1}G'W\hat{\Omega}_0WG(G'WG)^{-1} + o_p(\hat{g}_0\hat{g}_0').\end{aligned} \tag{19}$$

Now let us set

$$W^{-1} = \hat{\Omega}_0 = n^{-1} \sum_i g(y_i, z_i, x_i, m_i; \alpha, \beta; \hat{E}[y|z]) g(y_i, z_i, x_i, m_i; \alpha, \beta; \hat{E}[y|z])',$$

then we get that

$$diag\left((\hat{\theta}_n - \theta)(\hat{\theta}_n - \theta)'\right) = diag\left((G'\hat{\Omega}G)^{-1} + o_p(\hat{\Omega}^{-1})\right). \tag{20}$$

We note that

$$diag\left((G'WG)^{-1}G'W\hat{\Omega}WG(G'WG)^{-1} - (G'\hat{\Omega}G)^{-1}\right) + o_p(\hat{g}_0\hat{g}_0') > 0 \; ev.,$$

which means this is the infeasible optimal weighting for large samples. This optimal weighting matrix can be estimated by the inverse of $(g\hat{g}')$, which is a block-diagonal matrix.

$$\hat{\Omega} = \begin{bmatrix} \hat{\hat{\Omega}} & 0 \\ 0 & \hat{B} \end{bmatrix}. \tag{21}$$

The block matrix corresponding to the imputation moments ($\hat{B}$) is positive definite if the additional moments do not have a zero optimal weight as n tends to infinity. In this case $diag(G\hat{W}G)^{-1}$ is smaller or equal than the diagonal of the optimal covariance matrix of the estimator that does not contain the added moments (which is $(\tilde{G}'^{-1}\hat{\hat{\Omega}}\tilde{G})$).

Assumption 4. $\hat{W} = \hat{\Omega}^{-1}$.

Proposition 2. *Under Assumption 1-2 and 4, $MSE(\hat{\theta}_n) \leq MSE(\tilde{\theta}_n)$ ev. for any admissible weighting of the $\tilde{\theta}_n$ estimator. The inequality is strict for the β-coefficients corresponding to the fully observed variables (z_i).*

Remark 1. Under our assumptions, the relative optimal weight for the qth and rth element of $\hat{g}$ (denoted by superscripts) has the same order as

$$\frac{L_2(\hat{g}_0^q - g_0^q)}{L_2(\hat{g}_0^r - g_0^r)}.$$

So if an element of $\hat{g}_0$ does not converge with a $\sqrt{n}$ rate to zero due to the estimates nuisance parameter, its relative weight is set to be arbitrarily close to zero, eventually. The optimal weighting matrix selects the moments auto-

matically so that the estimator is always root-n consistent with an asymptotic variance-covariance matrix that is the same as for the fully observed GMM estimator.
However, this inference introduces additional noise and loss of degrees of freedom in finite samples, so if the applied researcher implements the imputation method for $k \geq 4$, imputation will *increase* standard errors.

Remark 2. Our estimation method and theory can be easily extended to the case when there are more than one RHS variables are missing. However, we do not pursue the full imputation estimator where the researcher uses two variables with missing values to predict each other ('Swiss cheese case').

3 Monte Carlo simulation

4 Application

5 Conclusion

6 References

Abrevaya and Donald (2017) Ichimura and Newey (2015) Pakes and Pollard (1989) Chernozhukov et al. (2018)

7 Appendix

7.1 Proposition 1

Theorem 7 checking the assumptions?

7.2 Corollary 1

$$\hat{E}(zy|z = z_i) = (nh)^{-1} \frac{\sum_j K[H^{-1}(z_i - z_j)] z_i g(\alpha x_j + \beta z_i)}{(nh)^{-1} \sum_j K[H^{-1}(z_i - z_j)]}, \tag{22}$$

where the denominator clearly converges in probability to $f(z_i)$, uniformly, so we are going to ignore it, and focus on the expected value of

$$(nh)^{-1}\sum_j K[H^{-1}(z_i - z_j)]z_i g(\alpha x_j + \beta z_i) - E[zg(\alpha x + \beta z)|z = z_i]. \quad (23)$$

First, let us calculate

$$\begin{aligned} &E[(nh)^{-1}\sum_j K[H^{-1}(z_i - z_j)]z_i g(\alpha x_j + \beta z_i)|\mathbf{z}] = \quad (24)\\ &= (nh)^{-1}\sum_j K[H^{-1}(z_i - z_j)]z_i E[g(\alpha x_j + \beta z_i)|\mathbf{z}] = \\ &= (nh)^{-1}\sum_j K[H^{-1}(z_i - z_j)]z_i \int g(\alpha x + \beta z_i) f_{x|z}(x, z_j)dx, \end{aligned}$$

which gives

$$\begin{aligned} &E[(nh)^{-1}\sum_j K[H^{-1}(z_i - z_j)]z_i g(\alpha x_j + \beta z_i) - E[z_i g(\alpha x + \beta z_i)|z_i]|\mathbf{z}] = \quad (25)\\ &= (nh)^{-1}\sum_j K[H^{-1}(z_i - z_j)]z_i \int g(\alpha x + \beta z_i)(f_{x|z}(x, z_j) - f_{x|z}(x, z_i))dx. \end{aligned}$$

It is interesting that it is only the conditional distribution that has the discrepancy. Taking now expectation w.r.t. z_j as well,

$$\begin{aligned} &E[(nh)^{-1}\sum_j K[H^{-1}(z_i - z_j)]z_i g(\alpha x_j + \beta z_i) - E[z_i g(\alpha x + \beta z_i)|z_i]|z_i] = \quad (26)\\ &= h^{-1}\int K[H^{-1}(z_i - z)]z_i \int g(\alpha x + \beta z_i)(f_{x|z}(x, z) - f_{x|z}(x, z_i))f(z)dxdz = \\ &= \int K[\Delta z](z_i \int g(\alpha x + \beta z_i) Df_{x|z}(x, z_i)dx\Delta z(f(z_i) + Df(\bar{z})\cdot\Delta z\cdot H)d\Delta z + \\ &+ \int \Delta z' H D^2 f_{x|z}(x, \bar{\bar{z}}) H\Delta z dx \end{aligned}$$

after taking a second-order Taylor expansion in $f(z)$ and estimating $f_{x|z}(x, z) - f_{x|z}(x, z_i)$ similarly, finally, substituting $\Delta z = H^{-1}(z_i - z)$ for integration. By our boundedness assumptions, this is going to be bounded uniformly over z_i.

Given that we have second order kernel, we collect the terms and take integrals

(everything has the same rate uniformly over z_i)

$$E[(nh)^{-1}\sum_j K[H^{-1}(z_i - z_j)]z_i g(\alpha x_j + \beta z_i) - E[z_i g(\alpha x + \beta z_i)|z_i]] = \quad (27)$$
$$= O(h_{max}^2)$$

(rewrite this), but checked

7.3 Proposition 2

Before we would start, we prove that the infeasible optimal weighting matrix is indeed optimal.

Now we prove Proposition 2.