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. We assume that h is strictly increasing as 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

$$E\begin{bmatrix} (1-M_i)X_i(Y_i - h(\alpha X_i + Z_i\beta)) \\ (1-M_i)Z_i(Y_i - h(\alpha X_i + Z_i\beta)) \end{bmatrix} = 0 \ a.s. \tag{3}$$

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 first four arguments of 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, while the last argument is 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 sample analogues (for a sample of size n)

$$\begin{aligned} \hat{g}(a,b;\hat{E}[y|z]) &= n^{-1}\sum_{i=1}^{n} g(y_i,z_i,x_i,m_i;a,b;\hat{E}[y|z]) = \qquad (8) \\ &= n^{-1}\sum_{i=1}^{n}\begin{bmatrix} (1-m_i)x_i(y_i-h(ax_i+z_ib)) \\ (1-m_i)z_i(y_i-h(ax_i+z_ib)) \\ m_iz_i(y_i-\hat{E}[y|z](z_i;a,b)) \end{bmatrix}, \\ \hat{E}[y_i|z_i] &= \int h(ax+bz_i)\hat{f}_{x|z}(x,z_i)dx, \qquad (9) \end{aligned}$$

where $\hat{f}_{x|z}$ is an estimator of the conditional pdf $f_{x|z}$. Then the GMM estimator is minimizing

$$\hat{Q}_n(a,b) = \hat{g}(a,b;\hat{E}[y|z])'\hat{W}\hat{g}(a,b;\hat{E}[y|z]) \qquad (10)$$

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

$$\hat{W} \xrightarrow{p} W. \qquad (11)$$

2.2 Asymptotic properties

In the following arguments we closely follow Ichimura and Newey (2015) and Chernozhukov et al. (2018). Below there are further regulatory assumptions of the model. 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 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

$$\begin{aligned} 0 &= \hat{G}'(\hat{\theta}_n; \hat{E}[y|z])\hat{W}\hat{g}(\hat{\theta}_n; \hat{E}[y|z]) = \\ &= \hat{G}'(\hat{\theta}_n; \hat{E}[y|z])\hat{W}\hat{g}(\theta; \hat{E}[y|z]) + \hat{G}'(\hat{\theta}_n; \hat{E}[y|z])\hat{W}\hat{G}(\bar{\theta}_n; \hat{E}[y|z])(\hat{\theta}_n - \theta) = \\ &= \hat{G}'\hat{W}\hat{g}_0 + \hat{G}'\hat{W}\bar{G}(\theta_n - \theta). \end{aligned} \tag{12}$$

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 $\bar{\theta}$.

Given Assumption 2, we have that $\hat{\theta}_n$ consistently estimates θ, as $||\hat{G}'\hat{W}\bar{G}||$ is going to be bounded and $\hat{g}_0 \to g(\theta, E[y|z]) = 0$ with probability approaching 1. This in turn yields

$$\begin{aligned} \hat{\theta}_n - \theta &= -(\hat{G}'\hat{W}\bar{G})^{-1}\hat{G}'\hat{W}\hat{g}_0 \\ &= -(G'WG)^{-1}G'W\hat{g}_0 + o_p(\hat{g}_0). \end{aligned} \tag{13}$$

We are interested in the rate of

$$\hat{g}(\theta, \hat{E}(y|z)) - E[g(z, x, m; \theta; E[y|z])] = \hat{g}(\theta, \hat{E}(y|z)). \tag{14}$$

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

Proposition 1. *If $||E[y|z](z_i, a, b) - \hat{E}[y|z](z_i, a, b)|| = O(n^{-1/4})$ $Z_i - a.s.$ and for every a, b, then the imputation estimator converges with a root-n rate and it is asymptotically normal.*

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

$$\hat{E}(z_i y_i | z_i) = \frac{\sum_j K[H^{-1}(z_i - z_j)] z_i g(\alpha x_j + \beta z_i)}{\sum_j K[H^{-1}(z_i - z_j)]}. \tag{15}$$

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}.

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 are going to do very similar steps as when we analyze the behavior of the Nadaraya-Watson estimator. $\prod h_k = h$

$$\hat{E}(zy|z=z_i) = (nh)^{-1} \frac{\sum_j K[H^{-1}(z_i - z_j)] z_i g(\alpha x_j + \beta z_i)}{(nh)^{-1} \sum_j K[H^{-1}(z_i - z_j)]}, \tag{16}$$

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

$$(nh)^{-1} \sum_j K[H^{-1}(z_i - z_j)] z_i g(\alpha x_j + \beta z_i) - E[zg(\alpha x + \beta z)|z = z_i]. \tag{17}$$

First, let us calculate

$$\begin{aligned} &E[(nh)^{-1} \sum_j K[H^{-1}(z_i - z_j)] z_i g(\alpha x_j + \beta z_i)|\mathbf{z}] = \qquad (18)\\ &= (nh)^{-1} \sum_j K[H^{-1}(z_i - z_j)] z_i E[g(\alpha x_j + \beta z_i)|\mathbf{z}] = \\ &= (nh)^{-1} \sum_j K[H^{-1}(z_i - z_j)] z_i \int g(\alpha x + \beta z_i) f_{x|z}(x, z_j) dx, \end{aligned}$$

which gives

$$\begin{aligned}&E[(nh)^{-1}\sum_j K[H^{-1}(z_i - z_j)]z_i g(\alpha x_j + \beta z_i) - E[z_i g(\alpha x + \beta z_i)|z_i]|\mathbf{z}] = \\ &= (nh)^{-1}\sum_j K[H^{-1}(z_i - z_j)]z_i \int g(\alpha x + \beta z_i)(f_{x|z}(x, z_j) - f_{x|z}(x, z_i))dx.\end{aligned}\tag{19}$$

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

$$\begin{aligned}&E[(nh)^{-1}\sum_j K[H^{-1}(z_i - z_j)]z_i g(\alpha x_j + \beta z_i) - E[z_i g(\alpha x + \beta z_i)|z_i]|z_i] = \\ &= h^{-1}\int K[H^{-1}(z_i - z)]z_i \int g(\alpha x + \beta z_i)(f_{x|z}(x, z) - f_{x|z}(x, z_i))f(z)dxdz = \\ &= \int K[\Delta z](z_i \int g(\alpha x + \beta z_i)Df_{x|z}(x, z_i)dx\Delta z(f(z_i) + Df(\bar{z}) \cdot \Delta z \cdot H)d\Delta z + \\ &+ \int \Delta z' H D^2 f_{x|z}(x, \bar{\bar{z}})H\Delta z dx\end{aligned}\tag{20}$$

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)

$$\begin{aligned}&E[(nh)^{-1}\sum_j K[H^{-1}(z_i - z_j)]z_i g(\alpha x_j + \beta z_i) - E[z_i g(\alpha x + \beta z_i)|z_i]] = \\ &= O(h^2_{max})\end{aligned}\tag{21}$$

As long as the bandwidth h_{max} is $o(n^{-1/4})$, we only have to worry about the contribution of the variances. However, we do average over the individual estimates of $E[yz|z = z_i]$, which under random sampling will give that the contribution of the variance is going to be $\sqrt{n}$, under the restriction that the $\hat{E}(.|.)$ converges uniformly to the conditional expectation as a function, which gives the restriction that $nh \to \infty$.

So the rate of the estimator, even if we do not give the weighting matrix the power to eliminate the last k_z moments, will be $\sqrt{n}$. For this we need that the $k_z < 4$ that is involved in the calculations of the conditional expectations. In addition, discrete z_i-s are allowed.

2.3 The role of the weighting matrix and efficiency

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

$$\begin{aligned} diag\left((\hat{\theta}_n-\theta)(\hat{\theta}_n-\theta)'\right) &= \qquad (22)\\ &= ((G'WG)^{-1}G'W\hat{g}_0+o_p(\hat{g}_0))((G'WG)^{-1}G'W\hat{g}_0+o_p(\hat{g}_0))' = \\ &= (G'WG)^{-1}G'W\hat{\Omega}WG(G'WG)^{-1}+o_p(\hat{g}_0\hat{g}_0'). \qquad (23)\end{aligned}$$

Now let us set

$$W^{-1}=\hat{\Omega}=\hat{g}_0\hat{g}_0',$$

then we get that

$$diag\left((\hat{\theta}_n-\theta)(\hat{\theta}_n-\theta)'\right)=diag\left((G'\hat{\Omega}G)^{-1}+o_p(\hat{\Omega}^{-1})\right). \qquad (24)$$

We note that

$$diag\left((G'WG)^{-1}G'W\hat{\Omega}WG(G'WG)^{-1}-(G'\hat{\Omega}G)^{-1}\right)+o_p(\hat{g}_0\hat{g}_0')>0\ ev. \qquad (25)$$

USUAL THEORY - FIX THIS

The relative optimal weight for the qth and rth element of $\hat{g}$ (denoted by superscripts) has the same order as

$$\frac{\hat{g}_0^q-g^q}{\hat{g}_0^r-g^r}.$$

So if a moment does not converge with a $\sqrt{n}$ rate, its relative weight is zero. The optimal weighting matrix selects the moments automatically so that the estimator is always root-n consistent.

The optimal weighting matrix can be estimated by the inverse of $(\hat{g}\hat{g}')$. This is a block-diagonal matrix. The block matrix corresponding to the imputation moments 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.

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)