Tulane Economics Working Paper Series

GMM Efficiency and IPW Estimation for Nonsmooth Functions

Otávio Bartalotti
Department of Economics
Tulane University
New Orleans, LA
bartalot@tulane.edu

Working Paper 1301
January 2013

Abstract

In a GMM setting this paper analyzes the problem in which we have two sets of moment conditions, where two sets of parameters enter into one set of moment conditions, while only one set of parameters enters into the other, extending Prokhorov and Schmidt's (2009) redundancy results to nonsmooth objective functions, and obtains relatively efficient estimates of interesting parameters in the presence of nuisance parameters. One-step GMM estimation for both set of parameters is asymptotically more efficient than two-step procedures. These results are applied to Wooldridge's (2007) inverse probability weighted estimator (IPW), generalizing the framework to deal with missing data in this context. Two-step estimation of β_o is more efficient than using known probabilities of selection, but this is dominated by one-step joint estimation. Examples for missing data quantile regression and instrumental variable quantile regression are provided.

Keywords: generalized method of moments, nonsmooth objective functions, inverse probability weighting, missing data, quantile regression
JEL: C13

GMM Efficiency and IPW Estimation for Nonsmooth Functions *

Otávio Bartalotti†

Tulane University

First Draft: December, 2009

This Draft: April, 2012

Abstract

In a GMM setting this paper analyzes the problem in which we have two sets of moment conditions, where two sets of parameters enter into one set of moment conditions, while only one set of parameters enters into the other, extending Prokhorov and Schmidt's (2009) redundancy results to nonsmooth objective functions, and obtains relatively efficient estimates of interesting parameters in the presence of nuisance parameters. One-step GMM estimation for both set of parameters is asymptotically more efficient than two-step procedures. These results are applied to Wooldridge's (2007) inverse probability weighted estimator (IPW), generalizing the framework to deal with missing data in this context. Two-step estimation of β_o is more efficient than using known probabilities of selection, but this is dominated by one-step joint estimation. Examples for missing data quantile regression and instrumental variable quantile regression are provided.

1 Introduction

This paper extends Prokhorov and Schmidt (2009) analysis to the estimation of a general GMM problem with nonsmooth objective functions in which nuisance parameters are present. The framework developed encompasses several interesting problems in econometrics such as, missing data, censored or truncated data, treatment effects, instrumental variables etc. More importantly, by allowing nonsmooth objective functions,

*JEL: C13; Keywords: Generalized method of moments, Nonsmooth objective functions, Inverse Probability Weighting, Missing data, Quantile regression.

†Department of Economics, Tulane University. E-mail: bartalot@tulane.edu. Address: 206 Tilton Hall, Tulane University, New Orleans, LA, 70118, USA. I thank Jeff Wooldridge, Tim Vogelsang, Gary Solon and Peter Schmidt for helpful discussions and suggestions. Also, I am indebted to Quentin Brummet, Steve Dieterle, Thomas Fujiwara, Ilya Rahkovsky and participants at the 25^{th} Annual Congress of the European Economic Association, the 2010 Midwest Economics Association Annual Meeting and the 36th Eastern Economics Association Annual Conference. Any remainder errors and omissions are mine.

the analysis extends to models that have gained additional importance in recent years, e.g., LAD, quantile regression, censored LAD, quantile treatment effects and IVQR.

The results rely on Newey and McFadden (1994) to obtain the asymptotic variance of the GMM estimator under less restrictive assumptions on the smoothness of the objective functions. For that consider two sets of moment conditions, the first includes both the parameters of interest (β_o) and certain nuisance parameters (γ_o) while the second subset includes only the nuisance parameters. By defining four competing estimators based on different assumptions regarding the information available about these nuisance parameters and the moment conditions utilized, results about the relative efficiency of each proposed estimator are derived. These results provide guidance to applied work in the presence nuisance parameters.

As discussed by Prokhorov and Schmidt (2009), joint estimation of nuisance parameters and parameters of interest is more efficient than a two-step procedure or knowing the true nuisance parameters and disregarding the second set of moment conditions. This fact is due to the information contained in correlation between both sets of moment conditions which is useful, even when γ_o is known. Using only the first set of moment conditions and known values of γ_o in the estimation procedure does not use the additional information embedded in the second set of moment conditions. These results are shown to hold when the objective functions are nonsmooth.

The general results are directly applicable to missing data problems and encompass Wooldridge's (2002b, 2007) Inverse Probability Weighting (IPW) estimators, extending its use for nonsmooth objective functions under the usual assumptions about the selection process, typically referred to as "ignorability". The general estimation results described confirm the validity of the result described by Wooldridge (2007), i.e., that it is better (in an efficiency sense) to estimate the selection probabilities, even if the latter are known. In other terms, we obtain more efficient estimates for β_o if we estimate γ_o than if we use the true γ_o. This result is "puzzling" because knowledge of γ_o, if properly exploited, cannot be harmful. Previous works discussed this result, such as Wooldridge (2002b, 2007) in the context of IPW. Hirano et al. (2003), Hitomi et al. (2008) and Prokhorov and Schmidt (2009) addressed the problem for the smooth objective function case. Even though this issue have been considered by Chen, Hong and Tarozzi (2008) in a semiparametric context with nonsmooth objective functions, the parametric approach proposed here provide, as a novelty, the conditions under which this result is valid and, furthermore, shows that the two-step estimator is usually dominated by a one-step joint estimation procedure that uses both the weighted moment conditions and the conditions associated with the selection model.

There have been several papers devoted to general theories of estimation in settings where nonsmooth objective functions are allowed, following Daniels (1961) and Huber (1967). Studies that allow for estimation of models based on nonsmooth objective functions include, among others, Pollard (1985), Pakes and Pollard (1989), Newey and McFadden (1994, section 7). Recent studies have approached the problem of

nonsmoothness with focus on semiparametric models, see Chen, Linton and Van Keilegom (2003) for a general estimation approach; Chen, Hong and Tarozzi (2008) for an approach for missing data problems with nonparametric first stage; and Cattaneo (2010) for an approach on the estimation of multi-valued treatment effects on a semiparametric framework.

The remainder of the paper is organized as follows. Section 2 sets up the general GMM framework used in the analysis and presents results regarding efficiency and redundancy of the estimators proposed, as well as estimators for the asymptotic variances of the parameters estimated. Section 3 studies the IPW approach to missing data problems proposed by Wooldridge (2002b, 2007), extending its scope to nonsmooth objective functions. Section 4 provides examples of the uses of the framework proposed here by, first, considering a model for the conditional quantile in a context with missing data; secondly I consider a simplified IVQR model as proposed by Chernozhukov and Hansen (2005, 2006). Section 5 concludes.

2 General Estimation Problem

Let $\omega^* \in Q^* \subset R^{\dim(\omega^*)}$ be a random vector; $\theta \in \mathbf{\Theta} \subset \mathbb{R}^P$ be a parameter vector, $\mathbf{\Theta}$ is a compact set, and the population condition

$$g_o(\theta_o) = E[g(\omega^*, \theta_o)] = 0 \tag{1}$$

where $g : Q^* \times \mathbf{\Theta} \to \mathbb{R}^m$ is a vector of known real-valued moment functions.

Newey and McFadden (1994) have shown consistency (Theorem 2.6) and asymptotic normality (Theorem 7.2) of the Generalized Method of Moments (GMM) estimator that uses the population moment condition above. These theorems cover the case in which the moment functions, $g(\cdot)$, are allowed to be nonsmooth. The GMM estimator minimizes the objective function

$$g_n(\theta)' \widehat{W} g_n(\theta) \tag{2}$$

where $\widehat{W}$ converges in probability to W, the appropriate positive semidefinite weighting matrix and $\omega_i^*, i = 1, ..., n$, are i.i.d. Both Theorem 2.6 and 7.2 from Newey and McFadden (1994) will be used to derive the asymptotic variance of the estimators. The first regards the consistency of the GMM estimator, relies on relatively weak conditions, and allow for discontinuities in the objective function. The second theorem demonstrates the asymptotic normality of the GMM estimator under a certain form of nonsmoothness of the objective function. As shown by Pollard (1985) the differentiability of the objective function $g(\omega_i^*, \theta)$ can be replaced by the differentiability of $g_o(\theta)$ for the purpose of obtaining the asymptotic normality of these estimators. The key condition to allow for nonsmooth objective functions is a "stochastic equicontinuity" assumption that guarantees uniform convergence in probability of the linear approximation of $g_o(\theta)$ by $g(\omega_i^*, \theta)$ in a shrinking neighborhood of θ_o. This is similar to the stochastic differentiability condition in

Pollard (1985) and primitive conditions are available in Pollard (1985), Andrews (1994) and Chen, Linton and Van Keilegom (2003).

Suppose that θ can be partitioned into subsets of parameters $(\beta', \gamma')' \in \mathbf{B} \times \mathbf{\Gamma} \subset \mathbb{R}^{p_1} \times \mathbb{R}^{p_2}$ and that $g(\cdot)$ can be partitioned into subsets of functions $(g_1(\cdot)', g_2(\cdot)')'$ as defined below. For notational convenience, ω^* is suppressed in the following discussion, then

$$E[g_1(\beta_o, \gamma_o)] = 0 \tag{3}$$

$$E[g_2(\gamma_o)] = 0 \tag{4}$$

where $\beta \in \mathbf{B}$, $\gamma \in \mathbf{\Gamma}$, $g_1(\cdot)$ and $g_2(\cdot)$ are m_1 and m_2 vectors of known functions, respectively ($m = m_1 + m_2$). Note that the second set of moment conditions does not depend on β while the first set of moment conditions depend on the full parameter set θ. Let $g_{n1}(\theta) = n^{-1} \sum_{i=1}^{n} g_1(\omega_i^*, \theta)$ and $g_{n2}(\gamma) = n^{-1} \sum_{i=1}^{n} g_2(\omega_i^*, \gamma)$,the sample analogues of the population moments. The framework developed here is valid for the general case of overidentification, *i.e.*, $m_1 \geqslant p_1$ and $m_2 \geqslant p_2$. This, and the appropriate rank conditions guarantees that two step estimation is possible. Let the asymptotic covariance matrix for the moment functions, Σ, be defined as

$$\Sigma = V\left[g(\theta_o)\right] \equiv \begin{bmatrix} C_{11} & C_{12} \\ C_{21} & C_{22} \end{bmatrix}$$

where we assume Σ is finite and nonsingular so its inverse exists:

$$\Sigma^{-1} \equiv \begin{bmatrix} C^{11} & C^{12} \\ C^{21} & C^{22} \end{bmatrix} = \begin{bmatrix} C_{11}^{-1}(I + C_{12}E^{-1}C_{21}C_{11}^{-1}) & -C_{11}^{-1}C_{12}E^{-1} \\ -E^{-1}C_{21}C_{11}^{-1} & E^{-1} \end{bmatrix}$$

since Σ (and Σ^{-1}) is symmetric $C_{12} = C_{21}'$ and the second equality holds (see White (1984), p.80) for $E \equiv C_{22} - C_{21}C_{11}^{-1}C_{12}$.

Define the matrix of derivatives as

$$\begin{aligned} G &\equiv \nabla_\theta g_o(\theta_o) = \nabla_\theta E[g(\theta_o)] \equiv \begin{bmatrix} G_{11} & G_{12} \\ 0 & G_{22} \end{bmatrix} \\ G_{11} &\equiv \nabla_\beta E[g_1(\beta_o, \gamma_o)] \\ G_{12} &\equiv \nabla_\gamma E[g_1(\beta_o, \gamma_o)] \\ G_{22} &\equiv \nabla_\gamma E[g_2(\gamma_o)] \end{aligned}$$

where the lower off-diagonal matrix equals zero since the second set of moment conditions does not depend on β.

Following Prokhorov and Schmidt (2009), define four different possible GMM estimators that differ in which moment conditions are used and/or whether γ is treated as known.

Definition 1 *Call the estimator of θ_o that minimizes*

$$g_n(\theta)'\Sigma^{-1}g_n(\theta) \tag{5}$$

the ONE-STEP estimator.

This is the usual GMM estimator that uses all the available orthogonality conditions jointly to estimate $\beta_{\mathbf{o}}$ and γ_o.

Definition 2 *Call the estimator of β_o that minimizes*

$$g_{n1}(\beta,\gamma_o)'C_{11}^{-1}g_{n1}(\beta,\gamma_o) \tag{6}$$

and γ_o is treated as known the KNOW-γ estimator.

This estimator ignores the second set of orthogonality conditions 4, treating γ_o as a known vector of parameters and estimating β_o using only the information available in the first set of moment assumptions.

Definition 3 *Call the estimator of β_o that minimizes*

$$g_n(\beta,\gamma_o)'\Sigma^{-1}g_n(\beta,\gamma_o) \tag{7}$$

and γ_o is treated as known the KNOW-γ-JOINT estimator.

This is the GMM estimator for β_o in the form considered by Qian and Schmidt (1999). In this case, one has information about the true values of γ_o but still uses both set of moments conditions in obtaining an estimate for β_o.

Definition 4 *Call the estimator of θ_o obtained in the following fashion, the TWO-STEP estimator:*

(i) the estimator $\widehat{\gamma}$ is obtained by minimizing

$$g_{n2}(\gamma)'C_{22}^{-1}g_{n2}(\gamma) \tag{8}$$

(ii) the estimator $\widehat{\beta}$ is obtained by minimizing

$$g_{n1}(\beta,\widehat{\gamma})'C_{11}^{-1}g_{n1}(\beta,\widehat{\gamma}) \tag{9}$$

and $\widehat{\gamma}$ is treated as given.

This is the sequential estimator that uses only the second set of moment conditions 4 to obtain a consistent estimator of the unknown parameter vector γ_o and then uses only the first set of moment conditions 3 to obtain the estimator of β_o. This estimator is widely used in the applied economics literature and encompasses several common applications.

The estimators defined above depend on a known Σ. In practice, Σ is not known and has to be replaced by an initial consistent estimate. Nevertheless, this does not impact the asymptotic variance of the feasible estimators.

The asymptotic variances of these estimators are derived directly from Newey and McFadden (1994) Theorem 7.2.

Theorem 1 *Let $V_{ONE-STEP}$, $V_{KNOW-\gamma}$,$V_{KNOW-\gamma-JOINT}$ and $V_{TWO-STEP}$ denote the asymptotic variance of ONE-STEP, KNOW-γ, KNOW-γ-JOINT and TWO-STEP respectively. Then, under the conditions described in Newey and McFadden (1994) Theorems 2.6 and 7.2,*

$$V_{ONE-STEP} = \left(G'\Sigma^{-1}G\right)^{-1} \tag{10}$$

$$V_{KNOW-\gamma} = \left(G_{11}'C_{11}^{-1}G_{11}\right)^{-1} \tag{11}$$

$$V_{KNOW-\gamma-JOINT} = \left(G_{11}'C^{11}G_{11}\right)^{-1} \tag{12}$$

$$V_{TWO-STEP} = B\Sigma B' \tag{13}$$

where,

$$B = \begin{bmatrix} B_{11} & B_{12} \\ 0 & B_{22} \end{bmatrix}$$

with

$$\begin{aligned} B_{11} &= -\left(G_{11}'C_{11}^{-1}G_{11}\right)^{-1} G_{11}'C_{11}^{-1} \\ B_{12} &= \left(G_{11}'C_{11}^{-1}G_{11}\right)^{-1} G_{11}'C_{11}^{-1}G_{12}\left(G_{22}'C_{22}^{-1}G_{22}\right)^{-1} G_{22}'C_{22}^{-1} \\ B_{22} &= -\left(G_{22}'C_{22}^{-1}G_{22}\right)^{-1} G_{22}'C_{22}^{-1} \end{aligned}$$

Proof. All proofs are provided in the appendix. ■

Since the structure of the variances presented in 1 is identical to the structure found by Prokhorov and Schmidt (2009) for the case in the objective functions are smooth, it is possible to analyze the relative asymptotic efficiency of these estimators by applying Theorem 2.2 on Prokhorov and Schmidt (2009).[1] Thus extending their result to allow nonsmooth objective functions

Corollary 1 *For the estimators defined above as the ONE-STEP, KNOW-γ, KNOW-γ-JOINT and TWO-STEP with asymptotic variances given by 10, 11, 12 and 13, respectively, the following statements hold:*

1. *KNOW-γ-JOINT is no less efficient than ONE-STEP, KNOW-γ and TWO-STEP for β_o.*
2. *If $C_{12} = 0$ then KNOW-γ-JOINT and KNOW-γ are equally efficient for β_o.*

[1] I denote the asymptotic variance of $\widehat{\theta}$ as V meaning that $\sqrt{n}(\widehat{\theta} - \theta_o)$ converges in distribution to $N(0, V)$.

3. If $G_{12} = 0$ then TWO-STEP and KNOW-γ are equally efficient for β_o.

4. If $C_{12} = 0$ and $G_{12} = 0$, then ONE-STEP, KNOW-γ, KNOW-γ-JOINT and TWO-STEP are equally efficient for β_o, and ONE-STEP and TWO-STEP are equally efficient for γ_o.

5. ONE-STEP is no less efficient than TWO-STEP.

6. If $m_1 = p_1$ then the ONE-STEP and TWO-STEP estimates of γ_o are equal.

7. If $m_1 = p_1$ and $m_2 = p_2$ then the ONE-STEP and TWO-STEP estimates are equal for both β_o and γ_o.

8. If $m_1 = p_1$ and $C_{12} = 0$ then the ONE-STEP and TWO-STEP estimates are equally efficient for both β_o and γ_o.

9. If $G_{12} = C_{12}C_{22}^{-1}G_{22}$, then KNOW-$\gamma$-JOINT and ONE-STEP are equally efficient for β_o.

10. If $G_{12} = C_{12}C_{22}^{-1}G_{22}$, then ONE-STEP, KNOW-$\gamma$-JOINT and TWO-STEP are no less efficient for β_o than KNOW-γ.

Statement 1 shows, as expected, that KNOW-γ-JOINT dominates the other estimators. This is an intuitive result since the known value of γ_o is at least as efficient as any estimate of γ_o, and KNOW-γ-JOINT uses the full set of relevant moment conditions.

Statement 2 is the result Qian and Schmidt (1999), where it is shown that using additional moment conditions that include no unknown parameters (as is the case for KNOW-γ-JOINT) improves efficiency except in the special case in which $C_{12} = 0$. In other words, the second set of moments is redundant in the estimation of β_o, Prokhorov and Schmidt (2009) call this M-redundancy.

Statement 3 gives the condition under which the first stage estimation of the nuisance parameter γ_o does not affect the asymptotic behavior of the second stage estimate of β_o. This result is similar to the one shown in Wooldridge (2002a), however in this case we are dealing with a nonsmooth objective function and, therefore, the restriction $G_{12} \equiv \nabla_\gamma E[g_1(\beta_o, \gamma_o)] = 0$ differs from the one proposed by Wooldridge.

Statement 4 provides conditions under which the ONE-STEP, KNOW-γ, KNOW-γ-JOINT and TWO-STEP estimators are equally efficient for β_o, hence the use of the additional moment conditions in 4 by the ONE-STEP, KNOW-γ-JOINT and TWO-STEP estimators does not improve the precision of the estimated parameters of interest as in the previous statement; and the knowledge of γ_o does not help in estimating β_o. This holds if the two sets of moment conditions are asymptotically uncorrelated and γ is not present in the first set of moment conditions.

Statement 5 is the usual result that in general, sequential estimation procedures are less efficient than joint (one step) estimation.

Statement 6, 7 and 8 follow directly from Ahn and Schmidt (1995) and show that the GMM separability holds in the framework that allows non-smooth objective functions. The GMM estimates for γ_o are not improved by the inclusion of an equal number of additional moment conditions and parameters. It can be

shown that if G_{11} is nonsingular, the ONE-STEP estimator for β_o can be written in terms of the ONE-STEP estimator of γ_o using the equation $g_{n1}(\widehat{\beta},\widehat{\gamma}) = C_{12}C_{22}^{-1}g_{n2}(\widehat{\gamma})$ (see appendix for details). Thus, as described by Prokhorov and Schmidt (2009) the ONE-STEP and TWO-STEP estimators for β_o will be derived from the same equation as long as $g_{n2}(\widehat{\gamma}) = 0$, which will be true under exact identification of γ_o, and asymptotically equally efficient if $C_{12} = 0$, since the moment conditions will be asymptotically uncorrelated, not adding to the information set exploited by ONE-STEP relatively to TWO-STEP.

Statement 9 and 10 are direct extensions of Prokhorov and Schmidt (2009). Statement 9 says that KNOW-γ-JOINT and ONE-STEP are equally efficient for the estimation of β_o, which means that knowledge of γ_o is not useful in terms of the efficiency of the estimates for β_o if we are using the full set of moment conditions and $G_{12} = C_{12}C_{22}^{-1}G_{22}$.

Statement 10 shows that under the same condition about G_{12}, KNOW-γ is dominated by ONE-STEP, KNOW-γ-JOINT and TWO-STEP. Knowledge of γ_o is not useful in the estimation of β_o in this case, and the KNOW-γ estimator does not use the information in the second set of moment conditions, which is useful unless $C_{12} = 0$.

The statements presented in Corollary 1 show that the results for GMM redundancy presented by Prokhorov and Schmidt (2009) extend to GMM estimation procedures based on nonsmooth objective functions.

Under the conditions of parts 9 and 10 of Corollary 1, the following corollary can be obtained.

Corollary 2 *If $G_{12} = C_{12}C_{22}^{-1}G_{22}$ and G_{22} is invertible, then*

$$V(\widehat{\beta}_{TWO-STEP}) = \left(G_{11}'C_{11}^{-1}G_{11}\right)^{-1}G_{11}'C_{11}^{-1}D_oC_{11}^{-1\prime}G_{11}\left(G_{11}'C_{11}^{-1}G_{11}\right)^{-1} \tag{14}$$

Additionally, if G_{11} is invertible, then

$$V(\widehat{\beta}_{TWO-STEP}) = G_{11}^{-1}D_oG_{11}^{-1\prime} \tag{15}$$

where

$$\begin{aligned} D_o &= E\left[e_ie_i'\right] \\ e_i &= \left[g_1(\omega_i^*,\theta) - C_{12}C_{22}^{-1}g_2(\omega_i^*,\gamma)\right] \end{aligned}$$

Note that e_i is the residual of the linear projection of the first set of moments conditions on the second set of moment conditions. This result is useful in the estimation of the asymptotic variance of the estimators, as I discuss below. Unfortunately, this applies only if the second set of moment conditions is exactly identified for formula 14 and if both sets of moment conditions are exactly identified for formula 15.

An arresting issue is to obtain estimates of the variance matrices described in theorem 1. The nonsmoothness of the objective function creates some obstacles to the usual estimations procedures. As described by Lee (2006) the fact that the estimates for the variances depend on the derivative of the expectation of the estimating function in the nonsmooth case warrants a more careful approach in estimating the variances used for inference.

A general approach that work in most cases is offered in Newey and McFadden (1994), and consists on obtaining consistent estimators for the separate components of the variance matrix. For estimating Σ or its relevant components a standard estimator is available. This procedure can be used in a first-step to obtain consistent estimates of the appropriate weighting matrix for the estimation procedure.

$$\widehat{\Sigma} = n^{-1}\sum_{i=1}^{n} g(\omega_i^*,\hat{\theta})g(\omega_i^*,\hat{\theta}) \qquad \widehat{C}_{11} = n^{-1}\sum_{i=1}^{n} g_1(\omega_i^*,\hat{\theta})g_1(\omega_i^*,\hat{\theta})$$
$$\widehat{C}_{12} = n^{-1}\sum_{i=1}^{n} g_1(\omega_i^*,\hat{\theta})g_2(\omega_i^*,\hat{\gamma}) \quad \widehat{C}_{22} = n^{-1}\sum_{i=1}^{n} g_2(\omega_i^*,\hat{\gamma})g_2(\omega_i^*,\hat{\gamma})$$

To be able to plug this estimates on the equations derived in Theorem 1 we need to obtain estimates of G, which can be difficult to obtain due to the nonsmoothness of the objective function. In this approach an estimate of G is obtained by numerical derivatives. Following Newey and McFadden (1994) let e_i denote the i^{th} unit vector, ϵ_n denote a small positive constant that depends on the sample size. Define the estimators for G and its components as

$$\begin{aligned}
\widehat{G_j} &= \frac{1}{2\epsilon_n}\left[n^{-1}\sum_{i=1}^{n} g(\omega_i^*,\hat{\theta}+e_j\epsilon_n) - g(\omega_i^*,\hat{\theta}-e_j\epsilon_n)\right] \\
\widehat{G}_{11_j} &= \frac{1}{2\epsilon_n}\left[n^{-1}\sum_{i=1}^{n} g_1(\omega_i^*,\hat{\beta}+e_j\epsilon_n,\hat{\gamma}) - g_1(\omega_i^*,\hat{\beta}-e_j\epsilon_n,\hat{\gamma})\right] \\
\widehat{G}_{12_j} &= \frac{1}{2\epsilon_n}\left[n^{-1}\sum_{i=1}^{n} g_1(\omega_i^*,\hat{\beta},\hat{\gamma}+e_j\epsilon_n) - g_1(\omega_i^*,\hat{\beta},\hat{\gamma}-e_j\epsilon_n)\right] \\
\widehat{G}_{22_j} &= \frac{1}{2\epsilon_n}\left[n^{-1}\sum_{i=1}^{n} g_2(\omega_i^*,\hat{\gamma}+e_j\epsilon_n) - g_2(\omega_i^*,\hat{\gamma}-e_j\epsilon_n)\right]
\end{aligned}$$

Where the subscript j denotes the j^{th} column of the matrix being estimated. Newey and McFadden(1994) Theorem 7.4 shows that if ϵ_n converges to zero and $\sqrt{n}\epsilon_n$ converges to infinity as n gets larger, these estimators will be consistent for the terms of the variances presented in theorem 1.

However, these estimators are cumbersome and not practical. As emphasized by Newey and McFadden, the choice of ϵ_n is a difficult problem and the formulation described above, using a unique value for ϵ_n would be good only if the estimated parameters had been scaled to have similar magnitudes. If that is not done, we would have to pick different ϵ_n for different components.

On specific cases, other estimators are available. As discussed in Newey and McFadden (1994) if $g(\omega^*, \hat{\theta})$ is differentiable with probability one, with $\nabla_\theta g(\omega^*, \hat{\theta})$ that is continuous at θ_o with probability one and dominated by an integrable function in a neighborhood of θ_o, then $\widehat{G} = n^{-1} \sum_{i=1}^{n} \nabla_\theta g(\omega_i^*, \hat{\theta})$ is a consistent estimator for G. Hence, the more standard estimator is available and would be easier to implement.

Clearly, alternatives could be available for specific moment conditions. Section 4 provides the example for the leading case of IPW for linear quantile regression.

Even in this case, the calculation of the matrix B that is present in the asymptotic variance of the TWO-STEP estimator could be cumbersome. For the cases in which the conditions from part 9 and 10 of Corollary 1 hold, namely $G_{12} = C_{12}C_{22}^{-1}G_{22}$, Corollary 2 offers a different approach to the problem of estimating the asymptotic variance in those cases (even though we still need to resort to one of the estimators above to obtain $\widehat{G}_{11}$). We can obtain an estimate of the matrix $E\left[e_i e_i'\right]$ by regressing the first set of moment conditions on the second set of moment conditions in the sample to obtain the residuals $\widehat{e}_i = g_1(\omega_i^*, \hat{\beta}, \hat{\gamma}) - \left[n^{-1} \sum_{i=1}^{n} g_2(\omega_i^*, \hat{\gamma}) g_1(\omega_i^*, \hat{\beta}, \hat{\gamma})\right] \left[n^{-1} \sum_{i=1}^{n} g_2(\omega_i^*, \hat{\gamma}) g_2(\omega_i^*, \hat{\gamma})\right]^{-1} g_2(\omega_i^*, \hat{\gamma})$, and calculating the sample analogue of the matrix $\widehat{D} = n^{-1} \sum_{i=1}^{n} \widehat{e}_i \widehat{e}_i'$. Unfortunately, this simple procedure is valid only for the asymptotic variance of the TWO-STEP estimator under the condition above and under exact identification of at least the second set of moment conditions.

For most of the relevant problems, we could use a bootstrap procedure to obtain consistent estimates of the variance of $\hat{\theta}$ directly, but these could be computationally demanding for models in which the solution of the optimization problem for both sets of moment conditions require numerical optimization of the objective function.

3 Estimation with missing data

This section specializes the results of the section 2 to a model in which missing data is allowed in a framework that expands that proposed by Wooldridge (2002b, 2007) to allow nonsmooth objective functions.

Consider $\omega \in Q \subset R^{\dim(\omega)}$ a random vector with density $f(\omega)$; $\beta \in \mathbf{B} \subset \mathbb{R}^{p_1}$ a parameter vector, where $\mathbf{B}$ is a compact set. Suppose there is the population moment equation

$$g_o(\beta_o) = E[g(\omega, \beta_o)] = 0 \tag{16}$$

where $g : Q \times \mathbf{B} \to \mathbb{R}^{m_1}$ is a vector of known real-valued moment functions with $m_1 \geq p_1$, so β_o could be overidentified. Assume β_o is the unique solution to 16. I am interested in estimating β_o.

Note that the moment conditions presented above hold in the unselected population. Assume nonrandom sampling occurs and it is characterized by a selection indicator, $s \in \{0, 1\}$, such that ω_i is observed if and only if $s_i = 1$. All or part of ω_i is not observed when $s_i = 0$.

The GMM estimator based on 16 using the selected sample, in effect makes the empirical moments $n^{-1}\sum_{i=1}^{n} s_i g(\omega_i, \beta)$ close to zero. These empirical moments are the sample analogues of the population moments of the form

$$E\left[sg(\omega, \beta)\right] = 0 \tag{17}$$

which are referred to as the unweighted selected population moments (Prokhorov and Schmidt 2009 and Wooldridge 2002b). The name emphasizes that they are evaluated at the selected rather than the full population of interest and differentiates them from the weighted selected population moments defined below. The selectivity problem occurs exactly because 17 may not hold; in other words, the value β_o that solves 16 may not also solve 17 (Prokhorov and Schmidt 2009). If that happens, the estimate for β_o obtained through this procedure is not generally consistent. In fact, its consistency and potential solutions for the data selection problem will depend on the relationship between the selection process and both the dependent and independent variables.

3.1 Data Selection under Ignorability

A straightforward solution is to solve the nonrandom sampling problem using inverse probability weighting (IPW) as shown by Wooldridge (2002b, 2007). To be able to use IPW we need some variables that are reasonable predictors of selection as described in Wooldridge (2007). This is formally stated as an "ignorability" of selection assumption.

Assumption 1 (Wooldridge, 2007, Assumption 3.1) *(i) ω_i is observed whenever $s_i = 1$;*

(ii) For a random vector z_i such that $P(s_i = 1 \mid \omega_i, z_i) = P(s_i = 1 \mid z_i) \equiv p(z_i)$;

(iii) For all $z \in Z \subset \mathbb{R}^J$, $p(z) > 0$;

(iv) z_i is observed whenever $s_i = 1$.

Item (ii) in this assumption requires that $s \perp \omega \mid z$. In other words, the selection has to be independent of the y and x conditional on z. As discussed at length by Wooldridge (2007), assumption 1 encompasses a variety of selection schemes common in the missing data literature, including "missing at random", "variable probability sampling", "selection on observables" etc. This allows, for example, that the probability of observing ω_i to depend on the stratum in which ω_i falls into; or that z_i is observed only along with ω_i; or that partial information is known about the incompletely observed data. Assumption 1 does not apply to the "selection on unobservables"[2] case as generally used in econometrics. I will not explore these possibilities directly here, referring the reader to Wooldridge (2007).

Assume that a conditional density determining selection is correctly specified and that a maximum likelihood estimator of the selection model is available.

[2] A quantile regression estimator for the case wheen selection is on unobservables is provided by Buchinsky (1998)

Assumption 2 (Wooldridge, 2007, Assumption 3.2) *(i) $G(z,\gamma)$ is a parametric model for $p(z)$, where $\gamma \in \Gamma \subset \mathbb{R}^{p_2}$ and $G(z,\gamma) > 0$ for all $z \in Z$ and $\gamma \in \Gamma$;*

(ii) There exists γ_o in the interior of Γ such that $p(z) = G(z,\gamma_o)$;

(iii) For a random vector v_i such that $D(v_i \mid \omega_i, z_i) = D(v_i \mid z_i)$, the estimator $\widehat{\gamma}$ solves a conditional maximum likelihood problem of the form

$$\max_{\gamma \in \Gamma} \sum_{i=1}^{n} \ln\left[f(v_i \mid z_i, \gamma)\right] \tag{18}$$

where $f(v \mid z, \gamma) > 0$ is a conditional density function known up to the parameters γ_o, and $s_i = h(v_i, z_i)$ for some nonstochastic function $h(\cdot,\cdot)$;

(iv) The solution to 18 has the first-order representation

$$\sqrt{n}(\widehat{\gamma} - \gamma_o) = \left\{E\left[d_i\left(\gamma_o\right) d_i\left(\gamma_o\right)'\right]\right\}^{-1} \left(n^{-\frac{1}{2}} \sum_{i=1}^{n} d_i\left(\gamma_o\right)\right) + o_p(1)$$

with $d_i(\gamma) \equiv \frac{\nabla_\gamma f(v_i|z_i,\gamma)'}{f(v_i|z_i,\gamma)}$, which is the $p_2 \times 1$ score vector for the MLE.

The assumption above requires standard regularity conditions about $G(z,\gamma)$, including smoothness of the parametric model. Even though this restricts the possibilities to model the selection process, it includes the most used probability models used in the literature. By doing so, we concentrate on the impacts of nonsmoothness in the model of interest and provide results about the use of IPW in correcting sample selection for those cases. Assumption 2 covers the cases presented by Wooldridge (2002b) in which the conditional log-likelihood was for a binary response model. The advantage of using this slightly more complicated framework is to allow z_i to be only partially observed and to permit s_i to be a function of another random variable v_i which includes a broader class of selection problems. For a deeper discussion on the extensions allowed by assumption 2, see Wooldridge (2007).

Note that the MLE estimator for γ_o described above can be obtained in a GMM setting as follows.

Let $\widehat{\gamma}$ the Maximum Likelihood Estimator (MLE) of γ_o, that is $\widehat{\gamma}$ solves

$$\max_{\gamma \in \Gamma} \sum_{i=1}^{n} \ln\left[f(v_i \mid z_i, \gamma)\right]$$

Define $g_2(z,\gamma,s) \equiv d(\gamma) = \frac{\nabla_\gamma f(v_i|z_i,\gamma)'}{f(v_i|z_i,\gamma)}$ and $g_{n2}(\gamma) \equiv n^{-1}\sum_{i=1}^{n} g_2(z_i,\gamma,s_i)$. Hence, $g_{n2}(\gamma) \xrightarrow{p} g_{2o}(\gamma) \equiv E\left[g_2(z,\gamma,s)\right]$. Then, the problem above is characterized by the following first order conditions

$$\begin{aligned} n^{-1}\sum_{i=1}^{n} g_2(\mathbf{z}_i,\widehat{\gamma},s_i) &= n^{-1}\sum_{i=1}^{n}\left[\frac{\nabla_\gamma f(v_i \mid z_i,\widehat{\gamma})'}{f(v_i \mid z_i,\widehat{\gamma})}\right] \\ &= n^{-1}\sum_{i=1}^{n} d_i\left(\widehat{\gamma}\right) = o_p(n^{-\frac{1}{2}}) \end{aligned}$$

and,

$$E\left[g_2(\mathbf{z},\gamma_o,s)\right] = E\left[d\left(\gamma_o\right)\right] = 0$$

Under assumption 1, Wooldridge (2002b) lemma 3.1. is valid, then

$$E\left\{\left[\frac{s}{G(z,\gamma_o)}\right] g(\omega,\beta_o)\right\} = E\left[\left(\frac{s}{p(z)}\right) g(\omega,\beta_o)\right] = E\left[g(\omega,\beta_o)\right]$$

Which suggests that we use the sampling probabilities to consistently estimate β_o Consider the weighted selected population moments that weight 17 by the inverse of the selection probability:

$$E\left[\left(\frac{s}{G(z,\gamma_o)}\right) g(\omega,\beta_o)\right] = 0 \tag{19}$$

Given an estimator for γ_o, $\widehat{\gamma}$, we can form $G(z_i,\widehat{\gamma})$ for all i with $s_i = 1$ and we are able to obtain consistent estimates for β_o by using the weighted selected population moments 19 as described in Wooldridge (2007). Note that, by the Law of Large Numbers and Law of Iterated Expectations, assumptions 1, 2 and consistency of $\widehat{\gamma}$ for γ_o (see Wooldridge 2002b, theorem 3.1).

$$\begin{aligned}
& n^{-1}\sum_{i=1}^{n} \frac{s_i}{G(z_i,\widehat{\gamma})} g(\omega_i,\beta) \xrightarrow{p} E\left[\frac{s_i}{p(z_i)} g(\omega_i,\beta)\right] \\
= \;& E\left[E\left[\frac{s_i}{p(z_i)} g(\omega_i,\beta) \mid \omega_i,z_i\right]\right] \\
= \;& E\left[\frac{p(z_i)}{p(z_i)} E\left[g(\omega_i,\beta) \mid \omega_i,z_i\right]\right] \\
= \;& E\left[g(\omega_i,\beta)\right] = g_o(\beta)
\end{aligned}$$

Hence, this provides a set of valid moment conditions that could be used to estimate β_o.

3.1.1 Efficiency Comparisons

The relative efficiency of the estimators for β_o that use IPW to correct a missing data problem under assumption 1 and 2 can be analyzed under the framework developed in section 2. Consider the two sets of valid moment conditions,

$$E\left[g_1(\omega,z,\beta,\gamma,s)\right] = E\left[\frac{s}{G(z,\gamma)} g(\omega,\beta)\right] = 0 \tag{20}$$

$$E\left[g_2(z,\gamma_o,s)\right] = E\left[\frac{\nabla_\gamma f(v \mid z,\gamma)'}{f(v \mid z,\gamma)}\right] = 0 \tag{21}$$

Any of the estimators discussed in section 2 can be used, differing on the set of moment conditions used and the knowledge about the weights.

Under the assumptions on the moment conditions and the selection process discussed in this section, the following lemma holds.

Lemma 1 *If the conditions of Newey and McFadden (1994) Theorems 2.6 and 7.2; Assumptions 1 and 2 hold, and the moment conditions are defined by 20 and 21, then* $G_{12} = C_{12}C_{22}^{-1}G_{22}$.

By using this result, we can see that under these assumptions, the results of Corollary 1 can be directly applied to this specific case.

Theorem 2 *Under the conditions of Lemma 1 , ONE-STEP, KNOW-*γ*-JOINT and TWO-STEP are no less efficient for* β_o *than KNOW-*γ*. Furthermore, ONE-STEP and KNOW-*γ*-JOINT are equally efficient for* β_o.

Hence, unless $C_{12} = 0$ (in which case the four estimators would be equally efficient), using ONE-STEP or TWO-STEP that estimate γ_o through MLE produce more efficient estimates for β_o than using known weights (if we knew them) in the KNOW-γ estimator. The KNOW-γ-JOINT estimator is as efficient as ONE-STEP as well, indicating that the knowledge of γ_o is not useful in terms of the efficiency of the estimates for β_o. The efficiency gains relatively to KNOW-γ are due to the use the information in the second set of moment conditions.

Therefore, the result described in Wooldridge (2002b, 2007) that KNOW-γ is inefficient relative to TWO-STEP, extends to a larger set of estimators in which the original set of unweighted moment conditions is nonsmooth as it was discussed by Chen, Hong and Tarozzi (2008) and Hitomi *et al.* (2008). In these cases we are better off estimating the weights by a conditional MLE than knowing them. Nonetheless, the TWO-STEP estimator is dominated by both ONE-STEP and KNOW-γ-JOINT and those should be used to obtain relatively efficient estimates of β_o.

It is important to note that the framework developed in this paper does not extend directly to semi-parametric cases in which the probability of selection is estimated nonparametrically. That can be a serious inconvenience when we have limited information about the selection process and would benefit from a more flexible estimator to those probabilities. However, as it is shown in the section 3.2 we can obtain consistent estimates for β_o even if using misspecified selection probabilities, as long as the data selection is exogenous.

3.2 Data Selection under Exogeneity of Selection

The literature in sample selection has long established that sample selection does not necessarily cause bias in unweighted estimators. As shown in Wooldridge (2007) if selection is exogenous conditional on the vector of covariates x the estimators of interest using the unweighted moment conditions will be consistent and, in fact, more efficient (Prokhorov and Schmidt, 2009) than their weighted counterparts. Following Wooldridge (2007), the properties of the estimators obtained under exogenous selection but with potential misspecification of the selection model are analyzed. Consider that we have a potentially misspecified model for the probability of selection given by $G(z, \gamma^*)$, which is not necessarily equal to the true $p(z_i)$.

Assume that the estimate $\widehat{\gamma}$ obtained based on that model is consistent to some parameter vector γ^* and $\sqrt{n}(\widehat{\gamma} - \gamma^*) = O_p(1)$.

In this case, the weighted moment condition

$$n^{-1}\sum_{i=1}^{n}\frac{s_i}{G(z,\widehat{\gamma})}g(\omega_i,\beta_o) \xrightarrow{p} E\left[\frac{s}{G(z,\gamma^*)}g(\omega,\beta)\right] \tag{22}$$

instead of $E[g(\omega,\beta)] = 0$, as seen in section 3.1.

Assume that the selection process is exogenous conditional on z.

Assumption 3 (Wooldridge, 2007, Assumption 4.1) *(i) ω_i is observed whenever $s_i = 1$;*

(ii) For a random vector z_i such that $P(s_i = 1 \mid \omega_i, z_i) = P(s_i = 1 \mid z_i) \equiv p(z_i)$;

(iii) z_i is observed whenever $s_i = 1$.

(iv) $\beta_o \in B$ solves the problem

$$E[g(\omega,\beta) \mid z] = 0$$

for all $z \in Z$.

This assumption is the same as in Prokhorov and Schmidt (2009) and as shown by them in Lemma 4.1 and Theorem 4.1 (p.53), which are not altered due to the use of nonsmooth objective functions, it implies

$$E[g(\omega,\beta) \mid z, s] = 0$$

Any function of z and s is uncorrelated with $g(\omega,\beta)$ and both weighted and unweighted moment conditions hold in the selected sample for *any* weight that is a function of z and s. Therefore, the weighted moment condition in equation 22 holds in the selected sample for any misspecified model $G(z,\gamma^*)$, including the unweighted moment conditions, when $G(z,\gamma^*) = 1$.

Hence, under exogeneity of selection, the IPW estimator for β_o is consistent, regardless of the misspecification of the model for probability of selection[3]. This robustness is an important feature of the IPW procedure and adds to its usefulness in applications.

4 Examples

4.1 Quantile Regression under Ignorability of Selection

Quantile regression is one of the main motivations for this research. As an example of the use of the results of this paper, consider I am interested in estimating the conditional quantile function (CQF) of a random variable y conditional on a vector of explanatory variables x. This is defined by,

$$Q_\tau(Y \mid X) = \inf\{y : F_Y(y \mid X) \geq \tau\}$$

[3]This conclusion is equivalent to Theorem 4.1 in Wooldridge (2007), extending it for nonsmooth objective functions.

where $\tau \in (0,1)$ indexes the τ^{th} quantile of the conditional distribution of Y. Suppose that the CQF is a linear model

$$Y = X'\beta_{\tau_o} + \varepsilon$$

and that $Q_\tau(\varepsilon \mid X) = 0$. In the population, β_o solves the following problem

$$\min_{\beta \in \mathbf{B}} E\left[\rho_\tau(Y - X'\beta_\tau)\right]$$

where, $\rho_\tau(u) = (\tau - 1\left[u \leq 0\right])u$

Given a random sample from the population of size n, it is possible to obtain consistent estimates of β_o by a standard quantile regression (QR) estimator.

$$\min_{\beta \in \mathbf{B}} n^{-1} \sum_{i=1}^{n} \rho_\tau(y_i - x_i'\beta_\tau)$$

Note that the minimization problem has the following of the first order conditions and sample analogue (Buchinsky, 1998)

$$\begin{aligned} E\left\{(\tau - 1\left[y - x'\beta_{\tau_o} \leq 0\right])x\right\} &= 0 \\ n^{-1}\sum_{i=1}^{n}\left(\tau - 1\left[y_i - x_i'\widehat{\beta_\tau} \leq 0\right]\right)x_i &= o_p(n^{-\frac{1}{2}}) \end{aligned}$$

Hence, we frame this problem as a GMM estimator that uses as moment conditions the first order conditions of the QR problem that identify β_{τ_o}. However, suppose a random sample of (y, x) is not observed. The selection mechanism is such that the full vector (y_i, x_i) is observed only if a certain binary variable that equals the unity, $s_i = 1$, if $s_i = 0$ at least some part of (y_i, x_i) is not observed. Then, in the selected sample, we can only estimate

$$n^{-1}\sum_{i=1}^{n} s_i\left\{\left(\tau - 1\left[y_i - x_i'\widehat{\beta_\tau} \leq 0\right]\right)x_i\right\} = o_p(n^{-\frac{1}{2}})$$

which is the sample analogue of

$$E\left[s\left[(\tau - 1\left[y - x'\beta_{\tau_o} \leq 0\right])x\right]\right] = 0$$

but the value β_{τ_o} that solves the population moment condition does not necessarily solve the selected population moment condition. Additionally, assume that the probability of selection can be written as a parametric function of some vector of variables (x_i, z_i) and parameters γ_o and that conditional on z_i, the terms of x_i that are not included in z_i and y_i are irrelevant for the probability of selection (Assumption 1).

$$P(s_i = 1 \mid y_i, x_i, z_i) = P(s_i = 1 \mid z_i) \equiv p(z_i, \gamma_o)$$

In this situation, we can estimate consistent and asymptotically normal estimates for β_{τ_o} using the selected sample by IPW. Note that,

$$\begin{aligned} & E\left[\frac{s}{p(z,\gamma_o)}\left[\left(\tau - 1\left[y - x'\beta_{\tau_o} \leq 0\right]\right)x\right]\right] \\ = \; & E\left[E\left[\frac{s}{p(z,\gamma_o)}\left[\left(\tau - 1\left[y - x'\beta_{\tau_o} \leq 0\right]\right)x\right] \mid x,y,z\right]\right] \\ = \; & E\left[\frac{E\left[s \mid z\right]}{p(z,\gamma_o)}\left[\left(\tau - 1\left[y - x'\beta_{\tau_o} \leq 0\right]\right)x\right]\right] \\ = \; & E\left[\left(\tau - 1\left[y - x'\beta_{\tau_o} \leq 0\right]\right)x\right] = 0 \end{aligned}$$

Therefore, we can estimate β_{τ_o} by using those weighted moment conditions. Naturally, we would need to estimate the weights if they are unknown.

Let the true selection model be a standard binary response model for simplicity. Then, estimate the selection of probability by MLE, or more conveniently, a GMM procedure that uses the first order conditions of the MLE for the selection model as moment conditions. The MLE maximization problem and its first order condition are given by, respectively,

$$\max_{\gamma\in\Gamma}\sum_{i=1}^{N}\left\{s_i \ln\left[p(z_i,\gamma)\right] + \left(1-s_i\right)\ln\left[1-p(z_i,\gamma)\right]\right\} \tag{23}$$

$$n^{-1}\sum_{i=1}^{n}\left[\nabla'_{\gamma}p(z_i,\widehat{\gamma})\frac{s_i - p(z_i,\widehat{\gamma})}{p(z_i,\widehat{\gamma})\left(1-p(z_i,\widehat{\gamma})\right)}\right] = o_p(n^{-\frac{1}{2}}) \tag{24}$$

where the estimator for γ_o is defined as the vector $\widehat{\gamma}$. Again, 24 is the sample analogue of the following moment condition,

$$E\left[\nabla'_{\gamma}p(z,\gamma_o)\frac{s - p(z,\gamma_o)}{p(z,\gamma_o)\left(1-p(z,\gamma_o)\right)}\right] = 0$$

Hence, we have two sets of moment conditions that can be used to estimate both the selection model and the conditional median model. The GMM estimator in this case would be given by any of the four estimators proposed in section 2, with

$$\begin{aligned} g_{n1}(\theta) &= n^{-1}\sum_{i=1}^{n}\frac{s_i}{p(z_i,\gamma)}\left\{\left(\tau - 1\left[y_i - x_i'\beta_{\tau} \leq 0\right]\right)x_i\right\} \\ g_{n2}(\gamma) &= n^{-1}\sum_{i=1}^{n}\left[\nabla'_{\gamma}p(z_i,\gamma)\frac{s_i - p(z_i,\gamma)}{p(z_i,\gamma)\left(1-p(z_i,\gamma)\right)}\right] \end{aligned}$$

the variance of the estimates will depend on the choice of estimator as stated by Theorem 1.

To estimate the variance of the estimated parameters we need to obtain valid estimates for the components

of G in the variance of $\hat{\theta}$.Note that, for example,

$$\begin{aligned} G_{11} &\equiv \nabla_\beta E[g_1(\beta_o,\gamma_o)] = \nabla_\beta E\left[\frac{s}{p(z,\gamma_o)}\left[(\tau - 1\left[y - x'\beta_{\tau_o} \leq 0\right])x\right]\right] \\ &= \nabla_\beta E\left[E\left[\frac{s}{p(z,\gamma_o)}\left[(\tau - 1\left[y - x'\beta_{\tau_o} \leq 0\right])x\right] \mid z,x,s\right]\right] \\ &= \nabla_\beta E\left[\frac{s}{p(z,\gamma_o)}\left(\tau - F_{y|z,x,s}(x'\beta_{\tau_o})\right)x\right] \\ &= E\left[\frac{s}{p(z,\gamma_o)} f_{y|z,x,s}(x'\beta_{\tau_o})x'x\right] \end{aligned}$$

hence, consistent estimates can be obtained by the sample analogues,

$$\begin{aligned} \widehat{G}_{11} &= n^{-1}\sum_{i=1}^{n}\frac{s_i}{p(z_i,\hat{\gamma})}\widehat{f}_{y|z,x,s}(x_i'\widehat{\beta}_\tau)x_i'x_i \\ \widehat{G}_{12} &= n^{-1}\sum_{i=1}^{n} -\frac{\nabla_\gamma' p(z_i,\hat{\gamma})}{\left[p(z_i,\hat{\gamma})\right]^2}s_i\left[\left(\tau - 1\left[y_i - x_i'\widehat{\beta}_\tau \leq 0\right]\right)x_i\right] \\ \widehat{G}_{22} &= n^{-1}\sum_{i=1}^{n}\left[\nabla_\gamma' p(z_i,\hat{\gamma})\left(\frac{s_i - p(z_i,\hat{\gamma})}{p(z_i,\hat{\gamma})\left(1-p(z_i,\hat{\gamma})\right)}\right)^2\nabla_\gamma p(z_i,\hat{\gamma})\right] \end{aligned}$$

where the last equality is a direct application of GIME and $\widehat{f}_{y|z,x,s}(\cdot)$ is a suitable estimator of the conditional density of y, commonly by a kernel estimator.

Note that the same asymptotic variance formula for the KNOW-γ estimator for $\widehat{\beta}_\tau$ is obtained by a simple extension of the results for weighted quantile regression presented in Koenker (2005) as shown in claim 1 in the appendix.

Since the conditions in Theorem 2 hold, we will obtain more efficient estimates by estimating the inverse probability weights than using the "true" weights, characterizing the result described in the literature (Wooldridge 2002b, 2007). The relatively more efficient estimate for β_{τ_o} is given by the one-step estimator that jointly estimates both the probability weights and the parameters of interest, β_{τ_o}.

One interesting point to note is that, even this relatively restrictive model for the CQF, which assumes linearity, can be very insightful about the potentially nonlinear true CQF. As discussed in detail by Angrist, Chernozhukov and Fernandez-Val (2006), a linear quantile regression provides the best linear approximation of the true CQF in the sense that it minimizes a weighted mean square error loss function. So even if we have reasons to believe that the true CQF in which we are interested is nonlinear, the use of a linear quantile regression in the example above would provide us with the "best linear approximation" to it in a similar way that a linear OLS model offers the best linear approximation to the conditional mean function. Hence, by using IPW to correct the selection bias caused by missing data we can recover this linear approximation to the CQF of interest, even if we don't know its true specification.

Nevertheless, this framework can be applied to nonlinear conditional quantiles of the form $Q_\tau(Y \mid X) =$

$m\left(X,\beta_{\tau_o}\right)$, with

$$\begin{aligned}
g_{n1}(\theta) &= n^{-1}\sum_{i=1}^{n}\frac{s_i}{p(z_i,\gamma)}\left\{\left(\tau-1\left[y_i-m\left(x_i,\beta_\tau\right)\leq 0\right]\right)\nabla_\beta m\left(x_i,\beta_\tau\right)\right\}\\
\widehat{G}_{11} &= n^{-1}\sum_{i=1}^{n}-\frac{s_i}{p(z_i,\hat{\gamma})}\widehat{f}_{y|z,x,s}\left(m\left(x_i,\widehat{\beta}_\tau\right)\right)\nabla'_\beta m\left(x_i,\widehat{\beta}_\tau\right)\nabla_\beta m\left(x_i,\widehat{\beta}_\tau\right)\\
\widehat{G}_{12} &= n^{-1}\sum_{i=1}^{n}-\frac{\nabla'_\gamma p(z_i,\hat{\gamma})}{\left[p(z_i,\hat{\gamma})\right]^2}s_i\left[\left(\tau-1\left[y_i-m\left(x_i,\widehat{\beta}_\tau\right)\leq 0\right]\right)\nabla_\beta m\left(x_i,\widehat{\beta}_\tau\right)\right]
\end{aligned}$$

and the remaining equations unchanged.

4.2 Instrumental Variable Quantile Regression

Consider a simplified version of the IVQR estimator described in Chernozhukov and Hansen (2006). We focus on the basic linear model that allow for heterogeneous effects given by,

$$Y_d = q(d,x,\tau) = d'\alpha_\tau + x'\beta_\tau$$

where d is a vector of (potentially endogenous) multi-valued treatment variables and x is a vector of covariates. Under the conditions described in Assumption 1 of Chernozhukov and Hansen (2006), the IVQR estimator of the vector of parameters $(\alpha(\tau)',\beta(\tau)')'$ proposed in that paper approximately solves the estimating equation[4]:

$$n^{-1}\sum_{i=1}^{n}(1\left[y_i-d_i'\alpha_\tau-x_i'\beta_\tau\leq 0\right]-\tau)(x_i',\widehat{\Phi}_{i\tau}')' = o_p(n^{-\frac{1}{2}})$$

where $\widehat{\Phi}_{i\tau}\equiv\widehat{\Phi}_\tau(\tau,x_i,z_i)$ is a vector of transformations of the instruments. In a simple model $\widehat{\Phi}_{i\tau}$ can be formed by the least squares projection of d on z and x (and its powers) (Chernozhukov and Hansen, 2006, 2008). In that simple case, we could write the sample analogue of the moment conditions that will identify the parameters of the model as

$$\begin{aligned}
g_{n1}(\theta) &= n^{-\frac{1}{2}}\sum_{i=1}^{n}\left\{(1\left[y_i-d_i'\alpha_\tau-x_i'\beta_\tau\leq 0\right]-\tau)(x_i',(x_i',z_i')\gamma)'\right\}\\
g_{n2}(\gamma) &= n^{-1}\sum_{i=1}^{n}(x_i',z_i')'[d_i-(x_i',z_i')\gamma]
\end{aligned}$$

Where $g_{n2}(\gamma)$Hence, the analysis developed in section 2 can be applied to the IVQR estimator proposed by Chernozhukov and Hansen (2005, 2006, 2008) and the results shown above are valid in its scope. Nevertheless, it is important to note that the framework developed in this paper does not extend directly to semiparametric cases in which the "first stage" is estimated nonparametrically. That can be a serious inconvenience when we have limited information about the form of the transformation on the vector of instruments that would be preferable in estimating IVQR.

[4]For simplicity I'm assuming that the weights $\widehat{V}_{i\tau}$ in Chernozhukov and Hansen (2006) are equal to the unit.

5 Conclusion

This paper (i) extends the GMM efficiency and redundancy results of Prokhorov and Schmidt (2009) to nonsmooth objective functions; (ii) analyzes the extent to which these results could be useful in the context of inverse probability weighting (IPW) as a mechanism to correct missing data issues, thus allowing its use in the LAD and quantile regression framework; (iii) verifies the conditions under which the weighting using known probabilities of selection leads to a less efficient estimate than using estimated probabilities of selection (Wooldridge 2002b, 2007, Prokhorov and Schmidt 2009, Hitomi et al. 2008), is valid under nonsmoothness of the objective functions that characterize the models of interest; and (iv) shows that even in that case the widely used two-step estimator is relatively less efficient than a one-step joint estimator.

Section 2 extends results on redundancy and efficiency due to Prokhorov and Schmidt (2009) that can now be applied to a wide range of contexts in which nonsmooth objective functions can be useful, including LAD, quantile regression, censored LAD and quantile treatment effects. Joint estimation of nuisance parameters and parameters of interest is more efficient than a two-step procedure or knowing the true nuisance parameters in the nonsmooth case. This springs from the information contained in the correlation between both sets of moment conditions which is useful, even when γ_o is known. Using only the first set of moment conditions and known values of γ_o in the estimation procedure does not use the additional information embedded in the second set of moment conditions, being inefficient. Some possible consistent estimators for the variance of both sets of parameters are presented.

Section 3 analyzes the missing data problem described in Wooldridge (2007). The selection model is estimated by a conditional MLE procedure, but the assumptions about the selection model are weak enough to cover most of the common parametric selection processes in the literature, like attrition, variable probability, "missing at random", etc. One important case not covered is "selection on unobservables". If we use both sets of moment conditions, knowledge about the nuisance parameters is not useful for the efficiency of the estimates of the parameters of interest. Additionally, the moment conditions that are associated with the selection model are not redundant, except in special cases. Estimating the parameters of interest using only the first set of moment conditions with known probabilities of selection as weights is inefficient because it ignores information in the second set of moment conditions. This is the type of result referred to in the selectivity literature, specially in the IPW approach to missing data.

In summary, this paper shows that IPW can be used to correct missing data problems when the model of interest is based on nonsmooth objective functions. Furthermore, two-step estimation of β_o is more efficient than using known probabilities of selection. Nonetheless, the two-step estimator is dominated by a one-step joint estimation procedure that uses both the weighted moment conditions and the selection model's conditions. Hence, this paper extends the analysis by Prokhorov and Schmidt (2009) to the relative efficiency of an IPW approach to deal with missing data problems in which the moment conditions of interest

are nonsmooth, encompassing, for example, LAD, quantile regression, Censored LAD and IVQR.

Finally, two illustrative examples of interesting models are provided that are encompassed by the general framework developed in this work. The first is a quantile regression model with missing data and, the second one is a simplified version of the Instrumental Variable Quantile Regression estimator (IVQR) presented by Chernozhukov and Hansen (2006).

6 Appendix

Proof of Theorem 1. For $V_{ONE-STEP}$, $V_{KNOW-\gamma}$ and $V_{KNOW-\gamma-JOINT}$ this result is a direct application of known results in the literature (see, e.g., p. 2186 in Newey and McFadden 1994 or more generally p. 1594 in Chen, Linton and Van Keilegom 2003) and the simplifications that take effect by the use of the appropriate weighting matrix. For $V_{TWO-STEP}$ I rely on the approximations used by Newey and McFadden (1994) in theorem 7.2 and Pakes and Pollard (1989) theorem 3.3 and lemma 3.5. Following Pakes and Pollard (1989), I claim that $g_n(\theta)$ is very well approximated by the linear function

$$\begin{aligned} L_n(\theta) &= \begin{bmatrix} L_{n1}(\theta) \\ L_{n2}(\theta) \end{bmatrix} = g_n(\theta_o) + G(\theta - \theta_o) \\ &= \begin{bmatrix} g_{n1}(\beta_o, \gamma_o) + G_{11}(\beta - \beta_o) + G_{12}(\gamma - \gamma_o) \\ g_{n2}(\gamma_o) + G_{22}(\gamma - \gamma_o) \end{bmatrix} \end{aligned}$$

within a $O_p(n^{-\frac{1}{2}})$ neighborhood of θ_o. More precisely, I need the approximation error to be of order $o_p(n^{-\frac{1}{2}})$ at $\widehat{\theta}$ and at θ^* which minimizes $\|L_n(\theta)\|$ globally. In the case analyzed here,

$$\begin{aligned} \left\| g_n(\widehat{\theta}) - L_n(\widehat{\theta}) \right\| &= \left\| g_n(\widehat{\theta}) - g_n(\theta_o) - G(\widehat{\theta} - \theta_o) \right\| \\ &= \left\| g_n(\widehat{\theta}) - g_n(\theta_o) - G(\widehat{\theta} - \theta_o) - g_o(\widehat{\theta}) + g_o(\widehat{\theta}) \right\| \\ &\leq \left\| g_n(\widehat{\theta}) - g_o(\widehat{\theta}) - g_n(\theta_o) \right\| + \left\| g_o(\widehat{\theta}) - G(\widehat{\theta} - \theta_o) \right\| \\ &\leq o_p(1) n^{-\frac{1}{2}} \left[1 + \sqrt{n} \left\| (\widehat{\theta} - \theta_o) \right\| \right] + o_p(\left\| (\widehat{\theta} - \theta_o) \right\|) \\ &= o_p(n^{-\frac{1}{2}}) \end{aligned}$$

where in the last equality I used the fact that $\left\| (\widehat{\theta} - \theta_o) \right\| \leq O_p(n^{-\frac{1}{2}})$ (see Newey and McFadden 1994, page 2191). To correspond to a minimum of $\|L_n(\theta)\|$, the vector $G(\theta^* - \theta_o)$ must be equal to the linear projection of $-g_n(\theta_o)$ onto the space G. Hence,

$$G(\theta^* - \theta_o) = -G(G'G)^{-1} G' g_n(\theta_o)$$

from this equation, we can obtain

$$\sqrt{n}(\theta^* - \theta_o) = -\sqrt{n}(G'G)^{-1} G' g_n(\theta_o)$$

from Pakes and Pollard (1989) lemma 3.5. the result above holds for the case in which we use the appropriate positive semidefinite weighting matrix $\widehat{W}$ that converges in probability to W , in which case

$$\sqrt{n}(\theta^* - \theta_o) = -\sqrt{n}(G'\widehat{W}G)^{-1}G'\widehat{W}g_n(\theta_o)$$

as shown by Pakes and Pollard (1989) (page 1042) under the conditions listed above θ^* and $\widehat{\theta}$ are close enough in this shrinking neighborhood around θ_o such that we can write

$$\sqrt{n}(\widehat{\theta} - \theta_o) = \sqrt{n}(\theta^* - \theta_o) + o_p(1)$$

Hence, for the first step estimator, the following approximation is valid

$$\sqrt{n}(\widehat{\gamma} - \gamma_o) = -\sqrt{n}\left(G'_{22}C_{22}^{-1}G_{22}\right)^{-1}G'_{22}C_{22}^{-1}g_{n2}(\gamma_o) + o_p(1) \tag{25}$$

Then, for the second step, using the same results, we can approximate

$$\begin{aligned}
\sqrt{n}(\widehat{\beta} - \beta_o) &= -\sqrt{n}\left(G'_{11}C_{11}^{-1}G_{11}\right)^{-1}G'_{11}C_{11}^{-1}g_{n1}(\beta_o, \widehat{\gamma}) + o_p(1) \\
&= -\sqrt{n}\left(G'_{11}C_{11}^{-1}G_{11}\right)^{-1}G'_{11}C_{11}^{-1}\left[g_{n1}(\beta_o, \gamma_o) + G_{12}(\widehat{\gamma} - \gamma_o)\right] + o_p(1) \\
&= -\sqrt{n}\left(G'_{11}C_{11}^{-1}G_{11}\right)^{-1}G'_{11}C_{11}^{-1}g_{n1}(\beta_o, \gamma_o) + \\
&+ \sqrt{n}\left(G'_{11}C_{11}^{-1}G_{11}\right)^{-1}G'_{11}C_{11}^{-1}G_{12}\left(G'_{22}C_{22}^{-1}G_{22}\right)^{-1} \times \\
&\times G'_{22}C_{22}^{-1}g_{n2}(\gamma_o) + o_p(1)
\end{aligned} \tag{26}$$

then, by combining 25 and 26 we can write

$$\sqrt{n}(\widehat{\theta} - \theta_o) = B\sqrt{n}g_n(\theta_o) + o_p(1)$$

where,

$$B = \begin{bmatrix} B_{11} & B_{12} \\ 0 & B_{22} \end{bmatrix}$$

with

$$\begin{aligned}
B_{11} &= -\left(G'_{11}C_{11}^{-1}G_{11}\right)^{-1}G'_{11}C_{11}^{-1} \\
B_{12} &= \left(G'_{11}C_{11}^{-1}G_{11}\right)^{-1}G'_{11}C_{11}^{-1}G_{12}\left(G'_{22}C_{22}^{-1}G_{22}\right)^{-1}G'_{22}C_{22}^{-1} \\
B_{22} &= -\left(G'_{22}C_{22}^{-1}G_{22}\right)^{-1}G'_{22}C_{22}^{-1}
\end{aligned}$$

hence,

$$V_{TWO-STEP} = B\Sigma B'$$

■

Proof of Corollary 1. The proof follows directly from Prokhorov and Schmidt (2009) since theorem 1 has shown that the variance structure of the four estimators considered is the same as in Prokhorov and Schmidt (2009). The proof that the result hold directly for the case in which the objective functions considered are nonsmooth is presented in the technical supplement to this paper, available under request. ■

Proof of Corollary 2. Note that the asymptotic variance of $\sqrt{n}(\widehat{\beta}_{TWO-STEP} - \beta_o)$ can be rewritten as (note that $B_{12} = \left(G_{11}'C_{11}^{-1}G_{11}\right)^{-1} G_{11}'C_{11}^{-1}G_{12}\left(G_{22}'C_{22}^{-1}G_{22}\right)^{-1} G_{22}'C_{22}^{-1} = B_{11}G_{12}\left(G_{22}'C_{22}^{-1}G_{22}\right)^{-1} G_{22}'C_{22}^{-1}$)

$$
\begin{aligned}
V(\widehat{\beta}_{TWO-STEP}) &= B_{11}C_{11}B_{11}' + B_{12}C_{21}B_{11}' + B_{11}C_{12}B_{12}' + B_{12}C_{22}B_{12}' \\
&= B_{11}E\left[g_1(\omega_i^*,\theta)g_1(\omega_i^*,\theta)'\right]B_{11}' + B_{12}E\left[g_2(\omega_i^*,\gamma)g_1(\omega_i^*,\theta)'\right]B_{11}' + \\
&\quad + B_{11}E\left[g_1(\omega_i^*,\theta)g_2(\omega_i^*,\gamma)'\right]B_{12}' + B_{12}E\left[g_2(\omega_i^*,\gamma)g_2(\omega_i^*,\gamma)'\right]B_{12}' \\
&= B_{11}E\left[\begin{array}{c}\left(g_1(\omega_i^*,\theta) - G_{12}\left(G_{22}'C_{22}^{-1}G_{22}\right)^{-1}G_{22}'C_{22}^{-1}g_2(\omega_i^*,\gamma)\right)\times \\ \times\left(g_1(\omega_i^*,\theta) - G_{12}\left(G_{22}'C_{22}^{-1}G_{22}\right)^{-1}G_{22}'C_{22}^{-1}g_2(\omega_i^*,\gamma)\right)'\end{array}\right]B_{11}' \\
&\qquad \text{if } G_{12} = C_{12}C_{22}^{-1}G_{22} \\
&= B_{11}E\left[\begin{array}{c}\left(g_1(\omega_i^*,\theta) - C_{12}C_{22}^{-1}G_{22}\left(G_{22}'C_{22}^{-1}G_{22}\right)^{-1}G_{22}'C_{22}^{-1}g_2(\omega_i^*,\gamma)\right)\times \\ \times\left(g_1(\omega_i^*,\theta) - C_{12}C_{22}^{-1}G_{22}\left(G_{22}'C_{22}^{-1}G_{22}\right)^{-1}G_{22}'C_{22}^{-1}g_2(\omega_i^*,\gamma)\right)'\end{array}\right]B_{11}'
\end{aligned}
$$

Since it is assumed that G_{22} is invertible,

$$
\begin{aligned}
&= B_{11}E\left[\begin{array}{c}\left(g_1(\omega_i^*,\theta) - C_{12}C_{22}^{-1}G_{22}\left(G_{22}'C_{22}^{-1}G_{22}\right)^{-1}G_{22}'C_{22}^{-1}G_{22}G_{22}^{-1}g_2(\omega_i^*,\gamma)\right)\times \\ \times\left(g_1(\omega_i^*,\theta) - C_{12}C_{22}^{-1}G_{22}\left(G_{22}'C_{22}^{-1}G_{22}\right)^{-1}G_{22}'C_{22}^{-1}G_{22}G_{22}^{-1}g_2(\omega_i^*,\gamma)\right)'\end{array}\right]B_{11}' \\
&= B_{11}E\left[\left(g_1(\omega_i^*,\theta) - C_{12}C_{22}^{-1}g_2(\omega_i^*,\gamma)\right)\left(g_1(\omega_i^*,\theta) - C_{12}C_{22}^{-1}g_2(\omega_i^*,\gamma)\right)'\right]B_{11}'
\end{aligned}
$$

If we define $e_i = g_1(\omega_i^*,\theta) - C_{12}C_{22}^{-1}g_2(\omega_i^*,\gamma)$, and $D_o = E\left[e_ie_i'\right]$, we can write,

$$
V(\widehat{\beta}_{TWO-STEP}) = \left(G_{11}'C_{11}^{-1}G_{11}\right)^{-1}G_{11}'C_{11}^{-1}D_oC_{11}^{-1\prime}G_{11}\left(G_{11}'C_{11}^{-1}G_{11}\right)^{-1}
$$

In this case, we can write the variance of the two-step estimator for β_o in a quadratic form in which the term in the middle of the matrix is the residual of the linear projection of the first set of moment conditions on the second set of moment conditions.

If, in addition to the conditions above, we assume G_{11} is invertible, the result follows.

$$
V(\widehat{\beta}_{TWO-STEP}) = G_{11}^{-1}D_oG_{11}^{-1\prime}
$$

■

Proof of Lemma 2. First, note that,

$$\begin{aligned}
E\left[sg_2(z,\gamma_o,s)' \mid z\right] &= E\left[s\frac{\nabla_\gamma f(v_i \mid z_i,\gamma)'}{f(v_i \mid z_i,\gamma)} \shortmid z\right] \\
&= \int_{-\infty}^{\infty} s\frac{\nabla_\gamma f(v \mid z,\gamma)'}{f(v \mid z,\gamma)} f(v \mid z,\gamma)dv \\
&= \int_{-\infty}^{\infty} h(v,z)\nabla_\gamma f(v \mid z,\gamma)'dv \\
&= \nabla_\gamma \left[\int_{-\infty}^{\infty} h(v,z)f(v \mid z,\gamma)'dv\right] \\
&= \nabla_\gamma E\left[s \shortmid z\right] \\
&= \nabla_\gamma p(z,\gamma_o)
\end{aligned}$$

this is nonzero in general. Hence,

$$\begin{aligned}
C_{12} &= E\left[g_1(\omega^*,\beta_o,\gamma_o,s)g_2(z,\gamma_o,s)'\right] \\
&= E\left[\frac{s}{p(z,\gamma_o)}g(\omega,\beta_o)g_2(z,\gamma_o,s)'\right] \\
&= E\left[E\left[\frac{s}{p(z,\gamma_o)}g(\omega,\beta_o)g_2(z,\gamma_o,s)' \shortmid z\right]\right] \\
&= E\left[E\left[\frac{1}{p(z,\gamma_o)}g(\omega,\beta_o)sg_2(z,\gamma_o,s)' \shortmid z\right]\right] \\
&= E\left[\frac{1}{p(z,\gamma_o)}E\left[g(\omega,\beta_o) \shortmid z\right]E\left[sg_2(z,\gamma_o,s)' \shortmid z\right]\right], \text{ by ignorability} \\
&= E\left[\frac{g(\omega,\beta_o)}{p(z,\gamma_o)}E\left[sg_2(z,\gamma_o,s) \shortmid z\right]'\right] \\
&= E\left[\frac{g(\omega,\beta_o)}{p(z,\gamma_o)}\nabla_\gamma p(z,\gamma_o)\right]
\end{aligned}$$

which is generally nonzero.

Analyzing G_{12},

$$\begin{aligned}
G_{12} &= \nabla_\gamma E[g_1(\omega^*,\beta_o,\gamma_o,s)] \\
&= \nabla_\gamma E\left[\frac{s}{p(z,\gamma_o)}g(\omega,\beta_o)\right]
\end{aligned}$$

since, $g_1(\omega^*, \beta_o, \gamma_o, s) = \frac{s}{p(\mathbf{z},\gamma)} g(\omega, \beta)$, is smooth in γ,

$$\begin{aligned}
G_{12} &= E\left[\nabla_\gamma \left(\frac{s}{p(\mathbf{z},\gamma_o)}\right) g(\omega,\beta_o)\right] \\
&= E\left[-\frac{s}{\left(p(\mathbf{z},\gamma_o)\right)^2} \nabla_\gamma p(\mathbf{z},\gamma_o) g(\omega,\beta_o)\right] \\
&= E\left[-\frac{s}{p(\mathbf{z},\gamma_o)} g(\omega,\beta_o) \frac{\nabla_\gamma p(\mathbf{z},\gamma_o)}{p(\mathbf{z},\gamma_o)}\right] \\
&= -E\left[E\left[\frac{s}{p(\mathbf{z},\gamma_o)} g(\omega,\beta_o) \frac{\nabla_\gamma p(\mathbf{z},\gamma_o)}{p(\mathbf{z},\gamma_o)}\right] \shortmid z\right], \text{ by LIE} \\
&= -E\left[\frac{E(s \shortmid z)}{p(\mathbf{z},\gamma_o)} E\left[g(\omega,\beta_o) \shortmid z\right] \frac{\nabla_\gamma p(\mathbf{z},\gamma_o)}{p(\mathbf{z},\gamma_o)}\right] \\
&= -E\left[g(\omega,\beta_o) \frac{\nabla_\gamma p(\mathbf{z},\gamma_o)}{p(\mathbf{z},\gamma_o)}\right], \text{ since } E\left[s \shortmid z\right] = p(\mathbf{z},\gamma_o) \\
&= -C_{12}
\end{aligned}$$

Then, to prove the lemma 1 I need that $G_{22} = -C_{22}$, which follows from the Generalized Information Equality (remembering $g_2(\mathbf{z},\gamma_o,s)$ is a smooth function).

$$\begin{aligned}
G_{22} &= \nabla_\gamma E[g_2(\mathbf{z},\gamma_o,s)] \\
&= E[\nabla_\gamma g_2(\mathbf{z},\gamma_o,s)] \\
&= -E\left[g_2(\mathbf{z},\gamma_o,s) g_2(\mathbf{z},\gamma_o,s)'\right] = -C_{22}
\end{aligned}$$

hence, $G_{12} = -C_{12} = -C_{12}(-C_{22}^{-1} G_{22}) = C_{12} C_{22}^{-1} G_{22}$. ■

Proof of Theorem 2. This follows directly from Lemma 1 and statements 9 and 10 in Corollary 1. ■

Claim 1 *Consider the conditional quantile function*

$$Q_\tau(Y \mid X) = X'\beta_{\tau_o}$$

and the weighted linear quantile estimator obtained as

$$\widehat{\beta}_\tau = \arg\min_{b \in \mathbb{R}^p} \sum w_i \rho_\tau(y_i - x_i' b)$$

for some known weight w_i that could be a function of exogenous variables. Under conditions 4 and 5, we have

$$\sqrt{n}\left(\widehat{\beta_\tau} - \beta_\tau\right) \sim N\left(0, \tau(1-\tau) D_1^{-1} D_o D_1^{-1}\right)$$

with, $D_2 = \lim_{n\to\infty} \sum_{i=1}^n w_i f_i(x_i'\beta_{\tau_o}) x_i x_i'$ and $D_o = \lim_{n\to\infty} \sum_{i=1}^n w_i^2 x_i' x_i$

Assumption 4 *For $Y_1, Y_2, \ldots, Y_n$ independent random variables with distribution functions $F_1, F_2, \ldots, F_n$, $\{F_i\}$ are absolutely continuous with continuous densities $f_i(\cdot)$ and weights, w_i, uniformly bounded away from 0 and ∞ at the points $f_i(x_i'\beta_{\tau_o})$ for every i.*

Assumption 5 *There exist positive definite matrices D_o and D_1 such that*

i) $\lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} w_i^2 x_i x_i' = D_o$

ii) $\lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} w_i f_i(x_i'\beta_{\tau_o}) x_i x_i' = D_1$

iii) $\max \frac{\|x_i\|}{\sqrt{n}} \to 0$

Proof of Claim 1. This proof follows the steps presented on Koenker (2005) p. 120.

Consider $u_i = y_i - x_i'\beta_{\tau_o}$, then

$$\begin{aligned}
\widehat{\beta_\tau} &= \arg\min_{b\in\mathbb{R}^p} \sum_{i=1}^{n} w_i \left[(y_i - x_i'b)\left(\tau - 1\left[y_i - x_i'b \le 0\right]\right)\right] \\
&= \arg\min_{b\in\mathbb{R}^p} \sum_{i=1}^{n} w_i \rho_\tau(u_i)
\end{aligned}$$

Now consider the following convex objective function, with unique minimizer at $\sqrt{n}\left(\widehat{\beta_\tau} - \beta_{\tau_o}\right)$,

$$Z_n(\delta) = \sum_{i=1}^{n} w_i \left[\rho_\tau\left(u_i - x_i'\frac{\delta}{\sqrt{n}}\right) - \rho_\tau(u_i)\right]$$

using Knight's identity $\rho_\tau(u-v) - \rho_\tau(u) = -v\Psi_\tau(u) + \int_0^v \left(1\left[u \le S\right] - 1\left[u \le 0\right]\right) dS$, with $\Psi_\tau(u) = \tau - 1\left[u \le 0\right]$

$$\begin{aligned}
Z_n(\delta) &= \sum_{i=1}^{n} w_i \left[-x_i'\frac{\delta}{\sqrt{n}}\Psi_\tau(u_i) + \int_0^{x_i'\frac{\delta}{\sqrt{n}}} \left(1\left[u_i \le S\right] - 1\left[u_i \le 0\right]\right) dS \right] \\
&= Z_{1n}(\delta) + \sum_{i=1}^{n} Z_{2ni}(\delta) = Z_{1n}(\delta) + Z_{2n}(\delta)
\end{aligned}$$

Note that, by the Lindeberg-Feller central limit theorem,

$$\begin{aligned}
Z_{1n}(\delta) &= -\delta'\frac{1}{\sqrt{n}}\sum_{i=1}^{n} w_i x_i' \Psi_\tau(u_i) \\
&= -\delta'\frac{1}{\sqrt{n}}\sum_{i=1}^{n} w_i x_i' \left(\tau - 1\left[u_i \le 0\right]\right) \\
&\sim -\delta' W \\
W &\sim N\left(0, \tau(1-\tau)\lim_{n\to\infty}\sum_{i=1}^{n} w_i^2 x_i x_i'\right)
\end{aligned}$$

Also,

$$\begin{aligned}
Z_{2n}(\delta) &= \sum_{i=1}^{n} Z_{2ni}(\delta) \\
&= \sum_{i=1}^{n} E\left[Z_{2ni}(\delta)\right] + \sum_{i=1}^{n} Z_{2ni}(\delta) - E\left[Z_{2ni}(\delta)\right]
\end{aligned}$$

but,

$$\begin{aligned}\sum_{i=1}^{n} E\left[Z_{2ni}(\delta)\right] &= \sum_{i=1}^{n} w_i \int_0^{x_i'\frac{\delta}{\sqrt{n}}} E\left[1\left[u_i \leq S\right] - 1\left[u_i \leq 0\right]\right] dS \\ &= \sum_{i=1}^{n} w_i \int_0^{x_i'\frac{\delta}{\sqrt{n}}} F_i(x_i'\beta_{\tau_o} + S) - F_i(x_i'\beta_{\tau_o}) dS\end{aligned}$$

let $S = \frac{t}{\sqrt{n}}$, then

$$\begin{aligned}\sum_{i=1}^{n} E\left[Z_{2ni}(\delta)\right] &= \frac{1}{n}\sum_{i=1}^{n} w_i \int_0^{x_i'\delta} \sqrt{n}\left[F_i\left(x_i'\beta_{\tau_o} + \frac{t}{\sqrt{n}}\right) - F_i(x_i'\beta_{\tau_o})\right] dt \\ &= \frac{1}{n}\sum_{i=1}^{n} w_i \int_0^{x_i'\delta} f_i(x_i'\beta_{\tau_o}) t dt + o(1) \\ &= \frac{1}{2n}\sum_{i=1}^{n} w_i f_i(x_i'\beta_{\tau_o})\delta' x_i x_i'\delta + o(1) \\ &\to \frac{1}{2}\delta'\left[\lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n} w_i f_i(x_i'\beta_{\tau_o}) x_i x_i'\right]\delta = \frac{1}{2}\delta' D_1 \delta\end{aligned}$$

Under A2(iii):

$$Z_n(\delta) \sim Z_o(\delta) = -\delta' W + \frac{1}{2}\delta' D_1 \delta$$

then

$$\begin{aligned}\sqrt{n}\left(\widehat{\beta_\tau} - \beta_\tau\right) &= \widehat{\delta_n} = \arg\min Z_n(\delta) \sim \widehat{\delta_o} = \arg\min Z_o(\delta) \\ \widehat{\delta_o} &= D_1^{-1} W\end{aligned}$$

hence,

$$\sqrt{n}\left(\widehat{\beta_\tau} - \beta_\tau\right) \sim N\left(0, \tau(1-\tau) D_1^{-1} D_o D_1^{-1}\right)$$

■

References

[1] Ahn, S. and P. Schmidt, 1995, A Separability Result for GMM Estimation, with Applications to GLS Prediction and Conditional Moment Tests. Econometric Reviews 14: 1, 19-34.

[2] Andrews, D., 1994, Empirical Process Methods in Econometrics, in: R. F. Engle & D. McFadden, (Eds.), Handbook of Econometrics, Vol. 4, pp. 2248-2294.

[3] Angrist, J.; V. Chernozhukov and I. Fernandez-Val, 2006, Quantile Regression Under Misspecification, with an Application to the U.S. Wage Structure. Econometrica 74, 539-563.

[4] Buchinsky, M., 1998, Recent Advances in Quantile Regression Models: A Practical Guideline for Empirical Research, The Journal of Human Resources 33, 88-126.

[5] Cattaneo, M., 2010, Efficient Semiparametric estimation of multi-valued treatment effects under ignorability. Journal of Econometrics 155, 138-154.

[6] Chen, X.; H. Hong and A. Tarozzi, 2008, Semiparametric Efficiency in GMM Models with Auxiliary Data. The Annals of Statistics 36, 808-843.

[7] Chen, X.; O. Linton and I. Van Keilegom, 2003, Estimation of Semiparametric Models When the Criterion Function is Not Smooth. Econometrica 71, 1591-1608.

[8] Chernozhukov, V. and C. Hansen, 2005, An IV Model of Quantile Treatment Effects. Econometrica 73, 245-261.

[9] Chernozhukov, V. and C. Hansen, 2006, Instrumental Quantile Regression Inference for Structural and Treatment Effect Models. Journal of Econometrics 132, 491-525.

[10] Chernozhukov, V. and C. Hansen, 2008, Instrumental Variable Quantile Regression: A Robust Inference Approach. Journal of Econometrics 142, 379-398.

[11] Daniels, H. E., 1961, The Asymptotic Efficiency of a Maximum Likelihood Estimator, in: Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, pp 151-163.

[12] Hirano, K.; G. Imbens and G. Ridder, 2003, Efficient Estimation of Average Treatment Effects Using the Estimated Propensity Score. Econometrica 71, 1161-1189.

[13] Hitomi, K.; Y. Nishiyama and R. Okui, 2008, A Puzzling Phenomenon in Semiparametric Estimation Problems with Infinite-Dimensional Nuisance Parameters. Econometric Theory 24, 1717-1728.

[14] Huber, P. J., 1967, The Behavior of Maximum Likelihood Estimates Under Nonstandard Conditions, in: L. M. LeCam & J. Neyman, (Eds.), Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, pp. 221-233.

[15] Koenker, R., 2005, Quantile Regression, Cambridge University Press.

[16] Lee, W., 2006, Robust Tests of Hypotheses in Models with M-Estimation. Working Paper.

[17] Newey, W. K. and D. McFadden, 1994, Large Sample Estimation and Hypothesis Testing, in: R. F. Engle & D. McFadden, (Eds.), Handbook of Econometrics, Vol. 4, pp. 2111-2245.

[18] Newey, W. K., 1984, A Method of Moments Interpretation of Sequential Estimators. Economic Letters 14, 1349-1382.

[19] Newey, W. K., 1994, The Asymptotic Variance of Semiparametric Estimators. Econometrica 62, 201-206.

[20] Pakes, A. and D. Pollard, 1989, Simulation and the Asymptotics of Optimization Estimators. Econometrica 57, 1027-1057.

[21] Pollard, D., 1985, New Ways to Prove Central Limit Theorems. Econometric Theory 1, 295-314.

[22] Prokhorov, A. and P. Schmidt, 2009, GMM Redundancy Results for General Missing Data Problems. Journal of Econometrics 151, 47-55.

[23] Qian, H. and P. Schmidt, 1999, Improved Instrumental Variables and Generalized Method of Moments Estimators. Journal of Econometrics 91, 145-169.

[24] White, H., 1984, Asymptotic Theory for Econometricians, Academic Press Inc.

[25] Wooldridge, J. M., 2002a, Econometric Analysis of Cross Section and Panel Data, MIT Press, Cambridge, Mass.

[26] Wooldridge, J. M., 2002b, Inverse Probability Weighted M-Estimation for Sample Selection, Attrition, and Stratification. Portuguese Economic Journal 1, 117-139.

[27] Wooldridge, J. M., 2007, Inverse Probability Weighted Estimation for General Missing Data Problems. Journal of Econometrics 141, 1281-1301.