1 Asymptotic properties of the GMM estimator

Hereinafter consistency and asymptotic normality of the proposed GMM estimator under correct parametric specification are derived, and explicit expressions for the influence functions are given. The derivation of the asymptotic properties draws on the results of Newey and McFadden (1994), Heckman, Ichimura, and Todd (1998) and Hoeffding (1948) and Serfling (1980). Recall the moment vector of length $K + VL$

$$g_n(\theta, \hat{\mathbf{m}}_{VL}, \hat{p}) = \frac{1}{n}\sum_i \begin{pmatrix} A(X_i) \cdot (Y_i - \varphi(X_i, \theta)) \cdot D_i \\ (\Lambda(X_i) \otimes \varphi(X_i, \theta) - \hat{\mathbf{m}}_{VL}(\hat{p}(X_i))) \cdot (1 - D_i) \cdot 1(X_i \in \hat{S}) \end{pmatrix}, \tag{1}$$

where $\hat{S} = \{x : \hat{p}(x) > 0\}$ is an estimate of the support S of X among the respondents ($D = 1$), and

$$\hat{\mathbf{m}}_{VL}(\hat{p}(X_i)) = \left(\hat{m}_1'\left(\hat{p}\left(X_i\right)\right) \cdot \Lambda_1\left(X_i\right), .., \hat{m}_l'\left(\hat{p}\left(X_i\right)\right) \cdot \Lambda_l\left(X_i\right), .., \hat{m}_L'\left(\hat{p}\left(X_i\right)\right) \cdot \Lambda_L\left(X_i\right)\right)',$$

is the vector of nonparametric estimates for all subpopulations $l = 1, .., L$, multiplied by the population indicator function Λ. Within each subpopulation $\hat{m}_l(\cdot) = (\hat{m}_{1l}(\cdot), .., \hat{m}_{vl}(\cdot), .., \hat{m}_{Vl}(\cdot))'$ is the *vector-valued* estimator of the conditional expected outcome vector $E[Y|p(X) = \rho, \Lambda_l(X) = 1, X \in S]$, and $\hat{m}_{vl}(\rho)$ is the estimator of the *v-th component* of the conditional expectation of the outcome vector $E[Y_v|p(X) = \rho, \Lambda_l(X) = 1, X \in S]$.

Theorem 1 (Consistency) *If*
(i) the parametric function $\varphi(x, \theta)$ is continuous in θ,
(ii) has a unique solution $\theta_0 \in \Theta$, with Θ compact, such that $\varphi(x, \theta) = E[Y|X = x]$ $\forall x$ iff $\theta = \theta_0$,
(iii) the moments $E\sup_{\theta\in\Theta} \|A(X) \cdot (Y - \varphi(X, \theta))\|$ and
$E\sup_{\theta\in\Theta} \left\|\varphi(X, \theta) \cdot \Lambda(X)(1 - D)1(X \in \hat{S}) - E\left[Y \cdot \Lambda\left(X\right)(1 - D)1\left(X \in S\right)\right]\right\|$ *are finite, in all subpopulations defined by $\Lambda(x)$,*
(iv) the number of subpopulations L is finite,
(v) $\hat{m}_{vl}(\hat{p}(x))$ is a consistent estimator of $E[Y_v|p(X) = p(x), \Lambda_l(X) = 1, X \in S]$,
(vi) the weighting matrix $\hat{W}$ converges in probability to a positive semidefinite matrix,
then the GMM estimator $\hat{\theta}_n = \arg\min_\theta g_n(\theta, \hat{\mathbf{m}}_{VL}, \hat{p})'\hat{W}g_n(\theta, \hat{\mathbf{m}}_{VL}, \hat{p})$ with moment vector (1) is consistent.

Remark 1 *Assumptions (iii) and (iv) could be relaxed to the form given in Corollary 2 (f). If the parametric specification φ is bounded, condition (iii) is automatically satisfied.*

The proof proceeds in three steps, similar to Newey and McFadden (1994). In Corollary 1 sufficient conditions for the consistency of an extremum estimator $\hat{\theta}(\hat{\mu}) = \arg\min \hat{Q}_n(\theta, \hat{\mu})$ are laid down, where $\hat{\mu}$ is a nonparametric preliminary estimate of some object μ_0. In Corollary 2 these sufficient conditions are specified more precisely for a generic GMM estimator. Finally, it is shown that the specific GMM estimator with moment function (1) satisfies these conditions.

Consider a generic extremum estimator

$$\hat{\theta}(\hat{\mu}) = \arg\min_\theta \hat{Q}_n(\theta, \hat{\mu}),$$

where $\hat{\mu}$ is a nonparametric estimate of μ_0 and $\hat{Q}_n$ is a stochastic objective function. Let Q_0 denote the nonstochastic limit function of $\hat{Q}_n$ and let θ_0 (the true value) be the minimizer of Q_0. Suppose that the estimator $\hat{\mu}$ converges in probability to μ_0. Define $B(\mu_0)$ as an arbitrarily small ball around μ_0. Consistency of $\hat{\mu}$ means that w.p.a.1 (with probability approaching one) $\hat{\mu}$ lies in the ball $B(\mu_0)$:

$$\lim_{n\to\infty} P(\hat{\mu} \in B(\mu_0)) = 1. \tag{2}$$

Corollary 1 *If*
(i) $Q_0(\theta, \mu)$ is uniquely minimized at (θ_0, μ_0),

(ii) $\theta_0 \in \Theta$*, with* Θ *the compact parameter space,*
(iii) $Q_0(\theta, \mu)$ *is continuous,*
(iv) $\hat{Q}_n(\theta, \mu)$ *converges uniformly in* Θ *to* $Q_0(\theta, \mu)$ *for all* $\mu \in B(\mu_0)$*:*

$$\lim_{n\to\infty} P\left(\sup_{\theta\in\Theta}\left|\hat{Q}_n(\theta,\mu) - Q_0(\theta,\mu)\right| < \varepsilon_1\right) = 1 \qquad \forall \mu \in B(\mu_0) \qquad \text{with } \varepsilon_1 > 0, \tag{3}$$

(v) $plim\, \hat{\mu} = \mu_0$*,*
then the estimator $\hat{\theta}(\hat{\mu}) = \arg\min_{\theta\in\Theta} \hat{Q}_n(\theta, \hat{\mu})$ *converges in probability to* θ_0*.*

Proof. With $\hat{\mu}$ consistent it follows by the Slutzky theorem that also the nonstochastic function $Q_0(\theta_0, \hat{\mu})$ is convergent:

$$\lim_{n\to\infty} P\left(|Q_0(\theta_0,\hat{\mu}) - Q_0(\theta_0,\mu_0)| < \varepsilon_2\right) = 1 \qquad \text{with } \varepsilon_2 > 0 \tag{4}$$

First it is shown that $Q_0(\hat{\theta}, \hat{\mu})$ converges to $Q_0(\theta_0, \mu_0)$ from above. Write $Q_0(\hat{\theta}, \hat{\mu}) - Q_0(\theta_0, \mu_0)$ as

$$\left(Q_0(\hat{\theta},\hat{\mu}) - \hat{Q}_n(\hat{\theta},\hat{\mu})\right) + \left(\hat{Q}_n(\hat{\theta},\hat{\mu}) - \hat{Q}_n(\theta_0,\hat{\mu})\right) + \left(\hat{Q}_n(\theta_0,\hat{\mu}) - Q_0(\theta_0,\hat{\mu})\right) + (Q_0(\theta_0,\hat{\mu}) - Q_0(\theta_0,\mu_0)).$$

From the uniform convergence assumption (3) together with (2) it follows that w.p.a.1 $\left|\hat{Q}_n(\hat{\theta},\hat{\mu}) - Q_0(\hat{\theta},\hat{\mu})\right| < \varepsilon_1$ and w.p.a.1 $\left|\hat{Q}_n(\theta_0,\hat{\mu}) - Q_0(\theta_0,\hat{\mu})\right| < \varepsilon_1$. From (4) it follows that w.p.a.1 $|Q_0(\theta_0,\hat{\mu}) - Q_0(\theta_0,\mu_0)| < \varepsilon_2$. The term $\hat{Q}_n(\hat{\theta},\hat{\mu}) - \hat{Q}_n(\theta_0,\hat{\mu})$ is negative by the definition of the estimator with $\hat{Q}_n(\hat{\theta},\hat{\mu}) = \min_{\theta\in\Theta}\hat{Q}_n(\theta,\hat{\mu})$. Thus the first, third and fourth terms are w.p.a.1 smaller than an arbitrarily small number and the second term is smaller than zero. Accordingly it follows with $\varepsilon \equiv \max(\varepsilon_1, \varepsilon_2)$

$$Q_0(\hat{\theta},\hat{\mu}) < Q_0(\theta_0,\mu_0) + 3\varepsilon \qquad \text{w.p.a.1.} \tag{5}$$

The following reasoning is similar in spirit to that of Theorem 2.1 in Newey and McFadden (1994). Let $\mathcal{N}$ be any open subset of Θ with $\theta_0 \in \mathcal{N}$ and let $\mathcal{N}^c = \Theta\backslash\mathcal{N}$ be its complement. From $\mathcal{N}^c$ compact and $Q_0(\theta, \mu)$ continuous it follows that $\inf_{\theta\in\mathcal{N}^c} Q_0(\theta,\mu_0) > Q_0(\theta_0,\mu_0)$, since θ_0 uniquely minimizes Q_0. Choosing $3\varepsilon = \inf_{\theta\in\mathcal{N}^c} Q_0(\theta,\mu_0) - Q_0(\theta_0,\mu_0)$ it follows w.p.a.1 that $Q_0(\hat{\theta},\hat{\mu}) < \inf_{\theta\in\mathcal{N}^c} Q_0(\theta,\mu_0)$. This means that w.p.a.1 $\hat{\theta}$ cannot be element of $\mathcal{N}^c$ and thus $\hat{\theta} \in \mathcal{N}$ must hold. Hence for sufficiently small ε all open subsets of Θ which contain θ_0 also w.p.a.1 contain $\hat{\theta}$, and all subsets of Θ which do not contain θ_0 also w.p.a.1 do not contain $\hat{\theta}$. Thus $\hat{\theta}$ converges in probability to θ_0. ■

Now the sufficient conditions of Corollary 1 are specified for a generic GMM estimator.

Corollary 2 *Suppose*
(a) $\hat{\mu}$ *is a consistent estimator of* μ_0 *and* $B(\mu_0)$ *a ball around* μ_0*, such that* $\lim_{n\to\infty} P(\hat{\mu} \in B(\mu_0)) = 1$*,*
(b) the data Z_i *are iid,* $\hat{W} \xrightarrow{p} W$*, where* W *a positive semidefinite matrix,*
(c) $WE\left[g(Z,\theta,\mu)\right] = 0$ *if and only if* $\theta = \theta_0$ *and* $\mu = \mu_0$*,*
(d) $\theta_0 \in \Theta$*, with* Θ *compact,*
(e) $g(Z, \theta, \mu)$ *is continuous in* θ *and* μ
(f) $E\left(\sup_{\mu\in B(\mu_0)}\sup_{\theta\in\Theta}\|g(Z,\theta,\mu)\|\right) < \infty$*,*
then the GMM estimator of $\hat{Q}_n(\theta,\hat{\mu}) = \left(\frac{1}{n}\sum g(Z_i,\theta,\hat{\mu})\right)' \hat{W} \left(\frac{1}{n}\sum g(Z_i,\theta,\hat{\mu})\right)$ *with limit function* $Q_0(\theta,\mu) = (Eg(Z,\theta,\mu))' W (Eg(Z,\theta,\mu))$ *satisfies the conditions of Corollary 1 and the GMM estimator is consistent.*

Proof. Showing that the conditions of Corollary 1 are satisfied follows with only minor modifications Lemma 2.4 and Theorem 2.6 of Newey and McFadden (1994) and is here omitted. ■

Proof. [of Theorem 1] It must be shown, that the conditions (a) to (f) of Corollary 2 are satisfied. Define

$$\hat{\mu} = n^{-1}\sum_i \hat{\mathbf{m}}_{VL}(\hat{p}(X_i)) \cdot (1 - D_i)1(X_i \in \hat{S}),$$

as the column vector of length VL. Each element $\hat{\mu}_{vl}(\hat{m}_{VL}, \hat{p})$ converges under assumption (v) in probability to

$$\underset{n\to\infty}{plim}\hat{\mu}_{vl} = E\left[Y_v \cdot \Lambda_l\left(X\right) \cdot (1 - D)1\left(X \in S\right)\right] \equiv \mu_{vl,0},$$

hence condition (a) is satisfied. Denote $\mu_0 = plim\,\hat{\mu}$. The moment function can be written as

$$g_n = n^{-1}\sum_i \begin{pmatrix} A(X_i) \cdot (Y_i - \varphi(X_i, \theta))D_i \\ \Lambda(X_i) \otimes \varphi(X_i, \theta)(1 - D_i)1(X_i \in \hat{S}) \end{pmatrix} - \begin{pmatrix} \mathbf{0}_K \\ \hat{\mu} \end{pmatrix},$$

and continuity of g_n depends only on the specification φ. Thus condition (e) is satisfied by assumption (i). Furthermore, conditions (b) and (d) are satisfied by assumptions (vi) and (ii).

The identification condition (c) is implied by assumption (ii): Since the upper part of the moment vector is independent of μ it can only have expectation zero if it represents the true mean function. By assumption (ii) this can only be the case if $\theta = \theta_0$. But in this case the lower part of the moment vector can only be zero if $\mu = \mu_0$.

Condition (f) is implied by assumptions (iii) and (iv):

$$\begin{aligned}
& E \sup_{\mu\in B(\mu_0),\theta\in\Theta} \sup \|g(Z, \theta, \mu)\| \\
= \; & E \sup_{\mu\in B(\mu_0),\theta\in\Theta} \sup \left\| n^{-1}\sum_i \begin{pmatrix} A(X) \cdot (Y - \varphi(X, \theta))D \\ \Lambda(X) \otimes \varphi(X, \theta)(1 - D)1(X \in \hat{S}) - \mu_0 \end{pmatrix} + \begin{pmatrix} \mathbf{0}_K \\ \mu_0 - \mu \end{pmatrix} \right\| \\
\leq \; & E \sup_{\mu\in B(\mu_0),\theta\in\Theta} \sup \left(\left\| \begin{pmatrix} A(X) \cdot (Y - \varphi(X, \theta))D \\ \Lambda(X) \otimes \varphi(X, \theta)(1 - D)1(X \in \hat{S}) - \mu_0 \end{pmatrix} \right\| + \left\| \begin{pmatrix} \mathbf{0}_K \\ \mu_0 - \mu \end{pmatrix} \right\| \right) \\
= \; & E \sup_{\theta\in\Theta} \left\| \begin{pmatrix} A(X) \cdot (Y - \varphi(X, \theta))D \\ \Lambda(X) \otimes \varphi(X, \theta)(1 - D)1(X \in \hat{S}) - \mu_0 \end{pmatrix} \right\| + E \sup_{\mu\in B(\mu_0)} \|\mu_0 - \mu\| \\
\leq \; & E \sup_{\theta\in\Theta} \left(\begin{array}{c} \|A(X) \cdot (Y - \varphi(X, \theta))D\| \\ +\sum_l^L \left|\Lambda_l(X) \cdot \varphi(X, \theta)(1 - D)1(X \in \hat{S}) - \mu_{l,0}\right| \end{array} \right) + E \sup_{\mu\in B(\mu_0)} \|\mu_0 - \mu\| \\
\leq \; & E \sup_{\theta\in\Theta} \left(\|A(X) \cdot (Y - \varphi(X, \theta))\|\right) \\
& + L \cdot E \max_l \sup_{\theta\in\Theta} \left|\Lambda_l(X) \cdot \varphi(X, \theta)(1 - D)1(X \in \hat{S}) - \mu_{l,0}\right| \\
& + E \sup_{\mu\in B(\mu_0)} \|\mu_0 - \mu\|
\end{aligned}$$

if all these terms have finite expectations. The last term is finite, since the size of the ball $B(\mu_0)$ around μ_0 becomes arbitrarily small. Since the expectations in the first two terms are bounded in each subpopulation by assumption (iii) and since the number of subpopulations L is finite by (iv), the whole expression is finite and condition (f) is satisfied. ■

To establish asymptotic normality I draw on the results of Heckman, Ichimura, and Todd (1998) who defined the class of *asymptotically linear with trimming* estimators. Let $\hat{p}(x)$ be an estimator of the probability $p(x) = P(D = 1|X = x)$ and $\hat{S}$ be an estimator of the support S of X in the $D = 1$ population, i.e. $S = \{x : p(x) > 0\}$. Furthermore, let $\hat{m}_{vl}(\rho)$ be an estimator of the v-th component of the conditional expectation of the outcome vector: $m_{vl}(\rho) = E[Y_v|p(X) = \rho, \Lambda_l(X) = 1, X \in S]$. The

estimator $\hat{m}_{vl}(\hat{p}(x))$ of $m_{vl}(p(x))$ is called asymptotically linear with trimming if it can be written as

$$[\hat{m}_{vl}(\hat{p}(x)) - m_{vl}(p(x))] \cdot \Lambda_l(x) 1(x \in \hat{S}) = \\ n_{l,1}^{-1} \sum_j \Psi_{vl,m}(Y_j, D_j, X_j; x) + n^{-1} \sum_j \Psi_{vl,p}(Y_j, D_j, X_j; x) + \hat{b}_{vl}(x) + \hat{R}_{vl}(x),$$

with $E[\Psi_{vl,p}(Y_j, D_j, X_j; X)|X = x] = 0$, $E[\Psi_{vl,m}(Y_j, D_j, X_j; X)|X = x] = 0$. Furthermore, *plim* $n_{l,1}^{-\frac{1}{2}} \sum \hat{b}_{vl}(X_j) = b_{vl} < \infty$ and $n_{l,1}^{-\frac{1}{2}} \sum \hat{R}_{vl}(X_j) = o_p(1)$, where $n_{l,1}$ is the number of $D = 1$ observations who belong to the l-th subpopulation. $\Psi_{vl,p}$ and $\Psi_{vl,m}$ are the influence functions stemming from the estimation of the probability $p(\cdot)$ and the regression curve $m_{vl}(\cdot)$, respectively, and are mean-zero. These local influence functions are allowed to depend on the sample size, e.g. through bandwidth parameters that converge to zero with increasing sample size. $\hat{b}_{vl}(x)$ is a local bias term converging to zero and with degenerate limit distribution of its sample average multiplied by $\sqrt{n}$. By choosing a bandwidth value that shrinks to zero sufficiently fast, the local bias term can be made to be of order $o_p(1)$. Lower order terms are summarized in the residual term $\hat{R}_{vl}(x)$. Heckman, Ichimura, and Todd (1998) give conditions on the density of X, the regression curve m and the estimators $\hat{m}_{vl}$ and $\hat{p}$ under which local polynomial regression estimators, e.g. kernel or local linear regression, are asymptotically linear with trimming (see the corollaries 3 and 4 below).

Theorem 2 (Asymptotic Normality) *If the conditions of Theorem 1 hold and*
(i) the estimator $\hat{m}_{vl}(\hat{p}(x))$ is asymptotically linear with trimming in all subpopulations l and with respect to all outcome variables v, with bias terms $\hat{b}_{vl}(x)$ of order $o_p(1)$; in addition the estimator of the support $\hat{S}$ fulfills the conditions in Heckman, Ichimura, and Todd (1998) (see the corollaries 3 and 4 below),
(ii) $VL \cdot Var\left(\Psi_{vl,m}(Y_j, D_j, X_j; X_i)\right)$ and $VL \cdot Var\left(\Psi_{vl,p}(Y_j, D_j, X_j; X_i)\right)$ are of order $o(n)$ for each outcome variable v and each subpopulation l,
(iii) $\varphi(x, \theta)$ is continuously differentiable with bounded derivative in a neighbourhood of θ_0,
$E\left[\|A(X)(Y - \varphi(X, \theta))\|^2\right] < \infty$, *and GWG nonsingular, where G is the expected gradient of the moment vector,*
(iv) $\lim_{n\to\infty} \frac{n_{l,1}}{n} = \lambda_l$ with $0 < \lambda_l < \infty$, for each subpopulation $l = 1, .., L$,
then the GMM estimator $\hat{\theta} = \arg\min_\theta g_n(\theta, \hat{\mathbf{m}}_{VL}, \hat{S})' \hat{W} g_n(\theta, \hat{\mathbf{m}}_{VL}, \hat{S})$ with moment vector (1) is asymptotically normal distributed

$$n^{\frac{1}{2}}(\hat{\theta} - \theta_0) \xrightarrow{d} N\left(0, (G'WG)^{-1} G'W \cdot E[JJ'] \cdot WG(G'WG)^{-1}\right) \tag{6}$$

with

$$J = g(Y, D, X, \theta_0, \mathbf{m}_{VL}) \\ - \begin{pmatrix} \mathbf{0}_K \\ \lambda_1^{-1} \cdot E[\Psi_{11,m}(Y, D, X; X_2)(1 - D_2)|Y, D, X] + E[\Psi_{11,p}(Y, D, X; X_2)(1 - D_2)|Y, D, X] \\ \vdots \\ \lambda_L^{-1} \cdot E[\Psi_{VL,m}(Y, D, X; X_2)(1 - D_2)|Y, D, X] + E[\Psi_{VL,p}(Y, D, X; X_2)(1 - D_2)|Y, D, X] \end{pmatrix},$$

where the expectation operator is with respect to X_2 and D_2.

Remark 2 *The matrix $E[JJ']$ can be estimated by the sample average $n^{-1} \sum J_i J_i'$. Its inverse $\Omega = [EJJ']^{-1}$ can be used as the weighting matrix in a second step GMM estimator. Further $n \cdot g_n' \hat{\Omega} g_n$ is asymptotically $\chi^2_{(VL)}$ distributed with number of freedoms equal to the number of overidentifying restrictions VL, which can be used to test whether the parametric specification is correct. Evaluation of J_i however requires expected values of the influence functions $\Psi_{vl,p}$ and $\Psi_{vl,m}$, which themselves can be estimated by sample averages. The influence functions depend on the employed estimators for the probabilities p and the regression curves m.*

Proof. The GMM estimator $\hat{\theta}_n = \arg\min_{\theta} g_n(\theta, \hat{\mathbf{m}}_{VL}, \hat{p})' \hat{W} g_n(\theta, \hat{\mathbf{m}}_{VL}, \hat{p})$ with moment vector (1) can be expressed by its first order condition as

$$G_n(\hat{\theta}, \hat{p})' \hat{W} \cdot g_n(\theta, \hat{\mathbf{m}}_{VL}, \hat{p}) = 0, \tag{7}$$

where $G_n = \frac{\partial g_n(\cdot)}{\partial \theta'}$ is the gradient of g_n with respect to θ and does not depend on $\hat{\mathbf{m}}_{VL}$. Applying the mean value theorem to $g_n(\theta, \hat{\mathbf{m}}_{VL}, \hat{p})$ about the true coefficient vector θ_0 yields, with $\bar{\theta}$ on the line between $\hat{\theta}$ and θ_0,

$$G_n'(\hat{\theta}, \hat{p}) \hat{W} \cdot \Big[g_n(\theta_0, \hat{\mathbf{m}}_{VL}, \hat{p}) + G_n(\bar{\theta}, \hat{p}) \cdot (\hat{\theta} - \theta_0) \Big] = 0. \tag{8}$$

Solving for $\hat{\theta} - \theta_0$ gives

$$n^{\frac{1}{2}} (\hat{\theta} - \theta_0) = - \Big(G_n(\hat{\theta}, \hat{p})' \hat{W} G_n(\bar{\theta}, \hat{p}) \Big)^{-1} G_n(\hat{\theta}, \hat{p})' \hat{W} \cdot n^{\frac{1}{2}} g_n(\theta_0, \hat{\mathbf{m}}_{VL}, \hat{p}). \tag{9}$$

Turning first to the term $n^{\frac{1}{2}} g_n(\theta_0, \hat{\mathbf{m}}_{VL}, \hat{p})$. Inserting (1) gives

$$\begin{aligned} n^{\frac{1}{2}} g_n(\theta_0, \hat{\mathbf{m}}_{VL}, \hat{p}) &= n^{-\frac{1}{2}} \sum g(Z_i, \theta_0, \mathbf{m}_{VL}, p) \\ &+ n^{-\frac{1}{2}} \sum \begin{pmatrix} \mathbf{0}_K \\ (\Lambda(X_i) \otimes \varphi(X_i, \theta_0) - \mathbf{m}_{VL}(p(X_i))) \cdot (1 - D_i) \cdot \Big[1(X_i \in \hat{S}) - 1(X_i \in S) \Big] \end{pmatrix} \\ &- n^{-\frac{1}{2}} \sum \begin{pmatrix} \mathbf{0}_K \\ (\hat{\mathbf{m}}_{VL}(\hat{p}(X_i)) - \mathbf{m}_{VL}(p(X_i))) \cdot (1 - D_i) 1(X_i \in \hat{S}) \end{pmatrix}, \end{aligned} \tag{10}$$

where $Z_i = (Y_i, D_i, X_i)$ and $\hat{S} = \{x : \hat{p}(x) > 0\}$ is an estimator of the support of X in the $D = 0$ population. Furthermore, $\mathbf{m}_{VL}$, p and S denote the nonstochastic limit functions of $\hat{\mathbf{m}}_{VL}$, $\hat{p}$ and $\hat{S}$. If the means $\mathbf{m}_{VL}$ and the probability p were known, only the first term would remain and the usual GMM properties would apply, see e.g. Newey and McFadden (1994). With $\mathbf{m}_{VL}$ and p estimated the second term accounts for the estimation of the support and the third term takes account of the estimation of $\mathbf{m}_{VL}$. The second term is $o_p(1)$ as shown in Heckman, Ichimura, and Todd (1998, p. 291). By inserting the definition of $\hat{\mathbf{m}}_{VL}$, this expression can be written as

$$\begin{aligned} n^{\frac{1}{2}} g_n(\theta_0, \hat{\mathbf{m}}_{VL}, \hat{p}) &= n^{-\frac{1}{2}} \sum g(Z_i, \theta_0, \mathbf{m}_{VL}, p) + o_p(1) \\ -n^{-\frac{1}{2}} \sum_i & \begin{pmatrix} \mathbf{0}_K \\ (\hat{m}_{11}(\hat{p}(X_i)) - m_{11}(p(X_i))) \cdot \Lambda_1(X_i) \cdot (1 - D_i) 1(X_i \in \hat{S}) \\ \vdots \\ (\hat{m}_{VL}(\hat{p}(X_i)) - m_{VL}(p(X_i))) \cdot \Lambda_L(X_i) \cdot (1 - D_i) 1(X_i \in \hat{S}) \end{pmatrix}, \end{aligned} \tag{11}$$

and with $\hat{m}_{vl}(\hat{p}(\cdot))$ asymptotically linear with trimming this equals

$$\begin{aligned} &= n^{-\frac{1}{2}} \sum g(Z_i, \theta_0, \mathbf{m}_{VL}, p) + o_p(1) \\ - n^{-\frac{1}{2}} \sum_i & \begin{pmatrix} \mathbf{0}_K \\ \Big[n_{1,1}^{-1} \sum_j \Psi_{11,m}(Y_j, D_j, X_j; X_i) + n^{-1} \sum_j \Psi_{11,p}(Y_j, D_j, X_j; X_i) + \hat{b}_{11}(X_i) + \hat{R}_{11}(X_i) \Big] \cdot (1 - D_i) \\ \vdots \\ \Big[n_{L,1}^{-1} \sum_j \Psi_{VL,m}(Y_j, D_j, X_j; X_i) + n^{-1} \sum_j \Psi_{VL,p}(Y_j, D_j, X_j; X_i) + \hat{b}_{VL}(X_i) + \hat{R}_{VL}(X_i) \Big] \cdot (1 - D_i) \end{pmatrix}. \end{aligned} \tag{12}$$

Now, examine the vl-th element in the last term of (12) in more detail:

$$n^{-\frac{1}{2}} \sum_i \left[n_{l,1}^{-1} \sum_j \Psi_{vl,m}(Y_j, D_j, X_j; X_i) + n^{-1} \sum_j \Psi_{vl,p}(Y_j, D_j, X_j; X_i) + \hat{b}_{vl}(X_i) + \hat{R}_{vl}(X_i) \right] \cdot (1 - D_i), \tag{13}$$

which can be reformulated as

$$
\begin{aligned}
= \quad & \frac{n^{\frac{3}{2}}}{2n_{l,1}} \left\{ n^{-2} \sum_i \sum_j \left(\Psi^{j,i}_{vl,m}(1-D_i) + \Psi^{i,j}_{vl,m}(1-D_j) \right) \right\} + \frac{n^{\frac{1}{2}}}{2} \left\{ n^{-2} \sum_i \sum_j \left(\Psi^{j,i}_{vl,p}(1-D_i) + \Psi^{i,j}_{vl,p}(1-D_j) \right) \right\} \\
& + n^{-\frac{1}{2}} \sum_i \hat{b}_{vl}(X_i)(1-D_i) + n^{-\frac{1}{2}} \sum_i \hat{R}_{vl}(X_i)(1-D_i)
\end{aligned}
$$

where $\Psi_{vl,m}(Y_j, D_j, X_j; X_i)$ and $\Psi_{vl,p}(Y_j, D_j, X_j; X_i)$ are abbreviated as $\Psi^{j,i}_{vl,m}$ and $\Psi^{j,i}_{vl,p}$, respectively. The last term converges to zero, as does the third term (by condition (i) of Theorem 2). Hence, the asymptotic distribution is driven by the first two terms in curly brackets. These terms are von Mises statistics (e.g. see Serfling 1980) and are asymptotically equivalent to the projections of the corresponding U-statistics if

$$
E \left\| (\Psi^{j,i}_{vl,m} + \Psi^{j,i}_{vl,p})(1-D_i) + (\Psi^{i,j}_{vl,m} + \Psi^{i,j}_{vl,p})(1-D_j) \right\|^2 = o(n) \tag{14}
$$

holds, see corollary 5 below. Because all influence functions are mean-zero, the term on the left in 14 equals

$$
\begin{aligned}
& = Var \left((\Psi^{j,i}_{vl,m} + \Psi^{j,i}_{vl,p})(1-D_i) + (\Psi^{i,j}_{vl,m} + \Psi^{i,j}_{vl,p})(1-D_j) \right) \\
& \leq 16 \cdot \max \left(Var \left(\Psi^{j,i}_{vl,m}(1-D_i) \right), Var \left(\Psi^{j,i}_{vl,p}(1-D_i) \right) \right) \\
& \leq 16 \cdot \max \left(Var \left(\Psi^{j,i}_{vl,m} \right), Var \left(\Psi^{j,i}_{vl,p} \right) \right) = o(n),
\end{aligned}
$$

as implied by assumption (ii). Thus, the U-statistics projection theorem (Corollary 5) can be applied and the term (13) is equivalent to

$$
\begin{aligned}
= \quad & \frac{n^{\frac{1}{2}}}{n_{l,1}} \sum_i E\left[\Psi_{vl,m}(Y_i, D_i, X_i; X_j)(1-D_j) | Y_i, D_i, X_i \right] \\
& + n^{-\frac{1}{2}} \sum_i E\left[\Psi_{vl,p}(Y_i, D_i, X_i; X_j)(1-D_j) | Y_i, D_i, X_i \right] \\
& + n^{-\frac{1}{2}} \sum_i \hat{b}_{vl}(X_i)(1-D_j) + n^{-\frac{1}{2}} \sum_i \hat{R}_{vl}(X_i)(1-D_j) + o_p\left(\frac{n}{n_{l,1}} \right) + o_p(1),
\end{aligned} \tag{15}
$$

where $n_{l,1}/n$ converges to λ_l by assumption (iv). The asymptotic distribution of this expression is driven by the first two terms, which determine the asymptotic variance. The other terms are all of order $o_p(1)$.

Applying the U-statistic projection theorem simultaneously to all outcome variables v and all subpopulations l in expression (12) requires that the full influence function vector with respect to all outcome variables and all subpopulations has expected squared norm of order $o(n)$, which is satisfied by assumption (ii) because the squared norm of a vector is smaller or equal than the sum of the squared norms of all vector elements, which are $o(n)$ as shown above.

Hence a central limit theorem can be applied to (12). Define J_i as

$$
\begin{aligned}
J_i = & \, g(Z_i, \theta_0, \mathbf{m}_{VL}, p) \\
& - \begin{pmatrix}
\mathbf{0}_K \\
\lambda_1^{-1} \cdot E[\Psi_{11,m}(Y_i, D_i, X_i; X_j)(1-D_j) | Y_i, D_i, X_i] + E[\Psi_{11,p}(Y_i, D_i, X_i; X_j)(1-D_j) | Y_i, D_i, X_i] \\
\vdots \\
\lambda_L^{-1} \cdot E[\Psi_{VL,m}(Y_i, D_i, X_i; X_j)(1-D_j) | Y_i, D_i, X_i] + E[\Psi_{VL,p}(Y_i, D_i, X_i; X_j)(1-D_j) | Y_i, D_i, X_i]
\end{pmatrix}.
\end{aligned}
$$

It follows by the multivariate Lindeberg-Feller central limit theorem (Greene 1997, Theorem 4.14) under the regularity conditions that $EJ_iJ_i' < \infty \; \forall i$, that all mixed third moments of the multivariate distribution are finite, that $EJJ' = \lim n^{-1} \sum EJ_iJ_i'$ is a finite and positive definite matrix, and that $\lim_{n\to\infty} \left(\sum_{i=1}^{n} EJ_iJ_i' \right)^{-1} EJ_iJ_i' = 0 \; \forall i$, that the moment function (1) is asymptotically normal distributed:

$$
n^{\frac{1}{2}} g_n(\theta_0, \hat{\mathbf{m}}_{VL}, \hat{p}) \xrightarrow{d} N(\, 0 \,, EJJ') \,. \tag{16}
$$

It remains to show that G_n converges in probability to the nonstochastic expected gradient G. The gradient of the moment vector (1) is

$$G_n(\hat{\theta},\hat{p}) = \frac{1}{n}\sum_i \begin{pmatrix} -A(X_i)\cdot\frac{\partial\varphi(X_i,\theta)}{\partial\theta'}D_i \\ \Lambda(X_i)\otimes\frac{\partial\varphi(X_i,\theta)}{\partial\theta'}(1-D_i)1(X_i\in\hat{S}) \end{pmatrix} \tag{17}$$

$$= \frac{1}{n}\sum_i \begin{pmatrix} -A(X_i)\cdot\frac{\partial\varphi(X_i,\theta)}{\partial\theta'}D_i) \\ \Lambda(X_i)\otimes\frac{\partial\varphi(X_i,\theta)}{\partial\theta'}(1-D_i)1(X_i\in S) \end{pmatrix} + \frac{1}{n}\sum_i \begin{pmatrix} \mathbf{0}_K \\ \Lambda(X_i)\otimes\frac{\partial\varphi(X_i,\theta)}{\partial\theta'}(1-D_i)\cdot\Big[1(X_i\in\hat{S})-1(X_i\in S)\Big] \end{pmatrix}$$

The latter term converges to zero since the first derivative of φ is bounded by assumption (iii) and $1(X_i\in\hat{S})$ converges to $1(X_i\in S)$. The first term converges to the expected gradient G by a law of large numbers. Hence the GMM estimates $\hat{\theta}$ are asymptotically normal

$$n^{\frac{1}{2}}(\hat{\theta}-\theta_0) \xrightarrow{d} N\left(\,0\,,(G'WG)^{-1}G'W\cdot E[JJ']\cdot WG(G'WG)^{-1}\right).$$

■

1.1 Influence functions for particular estimators

For estimating $E[JJ']$ the influence functions need to be known, which are derived below for Maximum Likelihood estimation of the probabilities p and kernel or local linear estimation of m.

Maximum Likelihood estimators of the coefficients β of a parametric regression model $\zeta(x,\beta)$ with likelihood function $l(Y,X;\beta)$ can be written in asymptotically linear form as

$$n^{\frac{1}{2}}(\hat{\beta}-\beta_0) = n^{-\frac{1}{2}}\sum_{j=1}^{n}\left[-E\frac{\partial^2\ln l(Y,X;\beta_0)}{\partial\beta\partial\beta'}\right]^{-1}\frac{\partial\ln l(Y_j,X_j;\beta_0)}{\partial\beta}+o_p(1), \tag{18}$$

see Newey and McFadden (1994, p.2141 ff). This expression represents the 'global' influence function for the coefficients β. It is global in the sense that it affects the estimated conditional mean function $\zeta(x,\hat{\beta})$ at all points x. To obtain an expression for the estimated conditional mean $\zeta(x,\hat{\beta})$ at a particular point x, expand $\zeta(\cdot)$ about β_0 to obtain the 'local' asymptotically linear representation

$$n^{-\frac{1}{2}}\Big(\zeta(x,\hat{\beta})-\zeta(x,\beta_0)\Big) = -n^{-\frac{1}{2}}\frac{\partial\zeta(x,\beta_0)}{\partial\beta'}\,[EH]^{-1}\sum_{j=1}^{n}\frac{\partial\ln l(Y_j,X_j;\beta_0)}{\partial\beta}+o_p(1), \tag{19}$$

where $EH = E\frac{\partial^2\ln l(Y,X;\beta_0)}{\partial\beta\partial\beta'}$ is the expected Hessian at β_0, as in (18). Hence, the parametric Maximum Likelihood estimate is asymptotically linear with trimming with zero local bias.

The 'local' influence function for *kernel* and *local linear regression* at interior points in the one-dimensional regression setting is (Heckman, Ichimura, and Todd 1998)

$$\begin{aligned}\psi_m(Y_j,p(X_j);p(x)) &= (Y_j-E[Y_j|p(X_j),D_j=1])\,\frac{K\left(\frac{p(X_j)-p}{h_n}\right)}{\underset{X_j|D=1}{E}K\left(\frac{p(X_j)-p}{h_n}\right)}D_j\cdot 1(p(x)>0)\\ &= (Y_j-m(p(X_j)))\,\frac{1}{h_n}\frac{K\left(\frac{p(X_j)-p}{h_n}\right)}{E\hat{f}_{p|D=1}(p(x))}D_j\cdot 1(p(x)>0),\end{aligned}$$

since $h^{-1}EK\left(\frac{p(X_j)-p(x)}{h}\right) = \int h^{-1}K\left(\frac{p(u)-p(x)}{h}\right)\cdot f_{p|D=1}(p(u))du$ where $f_{p|D=1}$ is the density of p in the

responding population ($D = 1$) and where $\hat{f}_{p|D=1}(\cdot)$ denotes a kernel density estimate using the same bandwidth h_n. Noting that by continuity of the density $\int h^{-1}K\left(\frac{p(u)-p(x)}{h}\right) \cdot f_{p|D=1}(p(u))du$ converges to $f_{p|D=1}(p(x)) \cdot \int K(u)\,du$ (Pagan and Ullah (1999), p. 362, 364 or Parzen (1962)) and because the kernel function is supposed to integrate to one, the influence function converges to:

$$\psi_m(Y_j, p(X_j); p(x)) \longrightarrow (Y_j - m(p(X_j))) \frac{1}{h_n} \frac{K\left(\frac{p(X_j)-p}{h_n}\right)}{f_{p|D=1}(p(x))} D_j \cdot 1(p(x) > 0). \tag{20}$$

1.2 Combination of both estimators

If the probability p is estimated by Maximum Likelihood and the regression function $m_{vl}(\hat{p})$ is estimated by kernel or local linear regression, the combined influence function in the asymptotically linear with trimming representation (see corollary 4) is

$$[\hat{m}_{vl}(\hat{p}(x)) - m_{vl}(p(x))] \cdot \Lambda_l(x)1(x \in \hat{S}) = \\ n_{l,1}^{-1} \sum_j \Psi_{vl,m}(Y_j, D_j, X_j; x) + n^{-1} \sum_j \Psi_{vl,p}(Y_j, D_j, X_j; x) + \hat{b}_{vl}(x) + \hat{R}_{vl}(x),$$

for the outcome variable v and the subpopulation l defined as $\{X|\Lambda_l(X) = 1\}$, where

$$\Psi_{vl,p}(Y_j, D_j, X_j; x) = -\frac{\partial m_{vl}(p(x))}{\partial p} \frac{\partial p(x, \beta_0)}{\partial \beta'} [EH]^{-1} \frac{\partial \ln l(D_j, X_j; \beta_0)}{\partial \beta} \cdot \Lambda_l(x)1(x \in S), \tag{21}$$

with $p(x)$ parametrically specified as $p(x, \beta_0)$ and EH the expected Hessian at β_0. Further

$$\Psi_{vl,m}(Y_j, D_j, X_j; x) = \frac{\Lambda_l(X_j)D_j}{h_{n_{l,1}}} (Y_{v,j} - m_{vl}(p(X_j))) K\left(\frac{p(X_j) - p(x)}{h_{n_{l,1}}}\right) \cdot \frac{\Lambda_l(x)1(x \in S)}{f_{p|D=1,\Lambda_l=1}(p(x))}. \tag{22}$$

$f_{p|D=1,\Lambda_l=1}(p)$ is the density of the probability p in the responding population belonging to subpopulation l, and can be estimated by kernel density estimation using bandwidth $h_{n_{l,1}}$.

If the probability p is estimated by Probit, i.e. $p(x, \beta_0) = \Phi(x'\beta_0)$, the influence function (21) becomes

$$\Psi_{vl,p}(Y_j, D_j, X_j; x) = \frac{\partial m_{vl}(p(x))}{\partial p} \phi(x'\beta_0)x' \left(E\left[\frac{\phi^2(X'\beta_0)XX'}{\Phi(X_j'\beta_0)\left(1 - \Phi(X_j'\beta_0)\right)} \right] \right)^{-1} \\ \cdot \left[D_j - \Phi(X_j'\beta_0)\right] \frac{\phi(X_j'\beta_0)X_j}{\Phi(X_j'\beta_0)\left(1 - \Phi(X_j'\beta_0)\right)} \cdot \Lambda_l(x)1(x \in S).$$

1.3 Additional corollaries

The following two corollaries are taken from Heckman, Ichimura, and Todd (1998).

Corollary 3 (Asymptotic linearity of $\hat{m}(p)$, Heckman, Ichimura and Todd, 1998) *Assuming that*
(i) sampling of (Y_j, X_j, D_j) is iid with finite variance of Y_j, and $X_j \in \Re^k$
(ii) the regression function $m(p)$ is twice continuously differentiable with second derivative Hölder continuous,
(iii) the stochastic bandwidth sequence a_n satisfies $\underset{n_1 \to \infty}{plim} \frac{a_{n_1}}{h_{n_1}} = a_0 > 0$ for some deterministic sequence $\{h_{n_1}\}$ that satisfies $\frac{n_1 h_{n_1}}{\ln n_1} \to \infty$ and $\lim n_1 h^4 < \infty$,
(iv) the kernel function K is compact and symmetric, $\int K(u)du = 1$, $\int uK(u)du = 0$.
(v) the estimated support $\hat{S} = \{x : \hat{f}_{X|D=1}(x) \geq q_0\}$ is estimated such, that $\sup_{x \in S} \left|\hat{f}_{X|D=1}(x) - f_{X|D=1}(x)\right|$ converges a.s. to zero, where $S = \{x : f_{X|D=1}(x) \geq q_0\}$, $\hat{f}_{X|D=1}$ is a kernel density estimate with kernel moments 1 through k equal to zero, and $f_{X|D=1}$ is $k+1$ times continuously differentiable with $(k+1)$-th

derivative Hölder continuous,
(vi) $m(\cdot)$ is estimated at interior points,
then the local polynomial regression estimator $\hat{m}(p(x))$ of polynomial order ≤ 1[1] is asymptotically linear with trimming:

$$\{\hat{m}(p(x)) - m(p(x))\} \cdot 1(x \in \hat{S}) = n_1^{-1} \sum_j \psi_m(Y_j, p(X_j); p) D_j + \hat{b}_m(p) + \hat{R}_m(p).$$

Remark 3 *By choosing a bandwidth sequence that converges sufficiently fast, such that $\lim n_1 h^4 = 0$ in Condition (iii) above, the local bias term $\hat{b}_m(p)$ will be of order $o_p(1)$. Hence through undersmoothing the bias term can be reduced to be of lower order.*

Corollary 4 (Asymptotic linearity of $\hat{m}(\hat{p})$, Heckman, Ichimura and Todd, 1998) *Suppose that*
(i) the estimator $\hat{p}(x)$ is asymptotically linear with trimming

$$(\hat{p}(x) - p(x))\, 1(x \in \hat{S}) = n^{-1} \sum_j \psi_p(D_j, X_j; x) + \hat{b}_p(x) + \hat{R}_p(x),$$

(ii) $\frac{\partial \hat{m}(p)}{\partial p}$ and $\hat{p}(x)$ are uniformly consistent and converge to $\frac{\partial m(p)}{\partial p}$ and $p(x)$, respectively, with $\frac{\partial m(p)}{\partial p}$ continuous,
(iii) $\underset{n_1 \to \infty}{plim}\, n_1^{-\frac{1}{2}} \sum_j \hat{b}_m(p(X_j)) D_j = b_m$, (iv) $\underset{n \to \infty}{plim}\, n^{-\frac{1}{2}} \sum_j \frac{\partial m(p(X_j))}{\partial p} \cdot \hat{b}_p(p(X_j)) = b_{m_p}$,
(v) $\underset{n \to \infty}{plim}\, n^{-\frac{1}{2}} \sum_j \left[\frac{\partial \hat{m}(\bar{p}(X_j))}{\partial p} - \frac{\partial m(p(X_j))}{\partial p} \right] \cdot \hat{R}_p(X_j) = 0$,
(vi) $\underset{n \to \infty}{plim}\, n^{-\frac{3}{2}} \sum_l \sum_j \left[\frac{\partial \hat{m}(\bar{p}(X_j))}{\partial p} - \frac{\partial m(p(X_j))}{\partial p} \right] \cdot \psi_p(D_l, X_l; X_j) = 0$, where $\bar{p}(x)$ is defined by a Taylor's expansion of $\hat{m}(\hat{p}(x))$ about $p(x)$,[2] then the estimator $\hat{m}(\hat{p}(x))$ of $m(p(x)) = E[Y|p(X) = p(x)]$ is also asymptotically linear with trimming: $[\hat{m}(\hat{p}(x)) - m(p(x))] \cdot 1(x \in \hat{S})$

$$= n_1^{-1} \sum_j \psi_m(Y_j, p(X_j); p) D_j + \frac{\partial m(p(x))}{\partial p} \cdot n^{-1} \sum_j \psi_p(D_j, X_j; x) + \hat{b}(x) + \hat{R}(x) \tag{23}$$

and
$\underset{n \to \infty}{plim}\, n^{-\frac{1}{2}} \sum_j \hat{b}(X_j) = b_m + b_{mp}$

Remark 4 *If the probability p is estimated either nonparametrically by local polynomial regression or parametrically, e.g. by maximum likelihood, then the conditions (i) to (vi) are satisfied (Heckman, Ichimura, and Todd 1998). In the latter case the local bias $\hat{b}_p(x)$ is zero.*

Corollary 5 (Asymptotic equivalence of V-statistic, U statistic and its projection) *Let $H_n(x_1, x_2)$ be a symmetric function and $X_1, .., X_n$ be iid random vectors. A natural estimator of $E[H_n]$ is the one-sample U-statistic*

$$U_n = \binom{n}{2}^{-1} \sum_{1 \leq i < j \leq n} H_n(X_i, X_j).$$

The associated von Mises statistic is

$$V_n = n^{-2} \sum_{i=1}^{n} \sum_{j=1}^{n} H_n(X_i, X_j)$$

[1] The local polynomial regression estimator of order 0 is the Nadaraya-Watson Kernel estimator and the local polynomial regression estimator of order 1 is the local linear estimator.

[2] $\hat{m}(\hat{p}(x)) = \hat{m}(p(x)) + \frac{\partial \hat{m}(\bar{p}(x))}{\partial p} (\hat{p}(x) - p(x))$

and the projection of the U-statistic is defined as

$$\hat{U}_n = \frac{n-2}{n} E[H_n] + \frac{2}{n} \sum_{i=1}^{n} E\left[H_n(X_1, X_2) | X_1 = X_i\right].$$

If $E\left\|H_n\right\|^2 = o(n)$ *then* $n^{\frac{1}{2}}(U_n - \hat{U}_n) = o_p(1)$ *and* $n^{\frac{1}{2}}(V_n - \hat{U}_n) = o_p(1)$. *See Hoeffding (1948) and Serfling (1980). Extended by Powell, Stock, and Stoker (1989) to allow* H_n *to depend on sample size.*

2 Power and size of J-test - Additional Monte Carlo results

In the Tables 2.1 to 2.4, the simulated power and size of the test for overidentifying restrictions

$$n \cdot g_n' \hat{\Omega} g_n \xrightarrow{d} \chi^2_{(VL)},$$

is given for the theoretical size of 10%. For the GMM estimators with kernel matching, also a Lagrange Multiplier (LM) test proposed in Imbens, Spady, and Johnson (1998) is examined, which is also asymptotically $\chi^2_{(VL)}$ distributed.[3] Imbens, Spady, and Johnson (1998) analyzed various alternative test statistics of which the LM test

$$LM = \hat{\lambda}' \left(\sum \hat{\pi}_i g_i g_i'\right) \left[\sum \hat{\pi}_i^2 g_i g_i'\right]^{-1} \left(\sum \hat{\pi}_i g_i g_i'\right) \hat{\lambda} \xrightarrow{d} \chi^2_{(VL)} \tag{24}$$

performed best in their Monte Carlo simulations, where the $\hat{\pi}_i$ are estimated empirical likelihood (or exponential tilting) probabilities and $\hat{\lambda}$ are estimated Lagrange multipliers. A convenient alternative to compute the empirical probability weights $\hat{\pi}_i$ provides the estimator of Back and Brown (1993)

$$\hat{\pi}_i = \frac{1}{n} \frac{1 - g_n \hat{\Omega} g_i}{1 - g_n \hat{\Omega} g_n},$$

which comes in a closed form solution and is semiparametrically efficient in estimating the empirical distribution function (Brown and Newey 2002). With this estimator of the empirical probabilities, the lagrange multiplier is

$$\hat{\lambda} = -\hat{\Omega} g_n,$$

where $\hat{\Omega} = [\hat{E}JJ']^{-1}$ is the inverse of the estimated covariance matrix of the moment vector, see also Brown, Newey, and May (2001) or Inkmann (2001, p.98 ff.).

In the Monte Carlo simulations, the J-test and the LM-test (24) are computed for the first and for the second step GMM estimator, with g_i and $\hat{\Omega}$ calculated at the respective estimates $\hat{\theta}_1$ and $\hat{\theta}_2$. The rejection frequencies at the theoretical 10%-critical value of the $\chi^2_{(L)}$ distribution are given in Tables 2.1 to 2.4, with J_1 and LM_1 referring to the statistics computed at the estimates $\hat{\theta}_1$ of the first step GMM estimator and J_2 and LM_2 referring to the second step GMM estimates $\hat{\theta}_2$. Examining the results for the correctly specified models, which are marked in *italics*, both the J-test and the LM-test often depart considerably from their theoretical level of 10%. For large values of L, they strongly tend to over-reject, whereas at L=1 they sometimes over- and sometimes under-reject. Overall, in their current implementation, both tests do not appear to be very reliable for specification testing.

There may be two explanations for this result. First, asymptotic theory might be a poor approximation to the finite-sample properties. Bootstrap versions of these tests, see Hall and Horowitz (1996) and Brown and Newey (2002), might therefore perform better. Alternatively, the tendency to over-reject could also be a result of the bandwidth choice by cross-validation. Centrality of the test statistics requires that bias vanishes faster than variance. This requires some undersmoothing in the bandwidth choice. Cross-validation, on the other hand, chooses a bandwidth value that trades off bias against variance and might thus lead to a non-central χ^2 distribution.

[3] For ridge matching, the middle term of the LM statistic (24) was often not invertible.

Table 2.1: Rejection frequency of GMM test, level 10%, Ridge matching, n=500

		DGP 1				DGP 2				DGP 3			
		φ_0	φ_1	φ_2	φ_3	φ_0	φ_1	φ_2	φ_3	φ_0	φ_1	φ_2	φ_3
L=14	J_1	100.0	*88.5*	100.0	100.0	76.2	97.8	*73.9*	84.3	99.9	100.0	100.0	*92.4*
	J_2	99.8	*83.6*	100.0	99.9	68.5	96.3	*58.2*	65.1	99.5	99.7	99.8	*86.1*
L=10	J_1	100.0	*39.3*	100.0	100.0	58.0	94.8	*58.7*	77.4	99.3	99.1	100.0	*54.5*
	J_2	98.9	*22.5*	99.9	99.3	47.2	91.9	*33.0*	44.6	91.4	95.6	99.0	*21.0*
L=7	J_1	100.0	*29.6*	95.8	100.0	44.7	90.3	*52.0*	71.7	99.3	94.7	100.0	*47.2*
	J_2	96.6	*9.9*	82.0	97.6	31.4	84.5	*20.4*	28.7	90.9	74.6	97.2	*8.1*
L=4	J_1	100.0	*27.9*	96.1	100.0	48.8	92.8	*56.5*	75.6	99.3	95.7	100.0	*48.4*
	J_2	98.4	*5.7*	85.9	99.1	35.5	88.3	*20.6*	33.9	88.4	80.0	98.3	*5.2*
L=1	J_1	100.0	*42.5*	75.0	100.0	45.2	89.4	*29.3*	60.9	99.7	96.1	100.0	*36.7*
	J_2	99.8	*2.5*	57.1	100.0	28.4	80.8	*12.3*	36.2	95.1	88.0	99.5	*2.4*

Note: Rejection frequency at the theoretical 10% critical value of the χ^2 distribution with L degrees of freedom. J_1 is the J-statistic of the first step GMM estimator, J_2 of the second step GMM estimator. The results for correctly specified models are marked in *italics*. L gives the number of subpopulations included.

Table 2.2: Rejection frequency of GMM test, level 10%, Ridge matching, n= 2000

		DGP 1				DGP 2				DGP 3			
		φ_0	φ_1	φ_2	φ_3	φ_0	φ_1	φ_2	φ_3	φ_0	φ_1	φ_2	φ_3
L=14	J_1	100.0	*96.2*	100.0	100.0	96.6	100.0	*86.3*	99.0	100.0	100.0	100.0	*97.3*
	J_2	100.0	*93.1*	100.0	100.0	93.9	100.0	*73.5*	94.8	100.0	100.0	100.0	*94.4*
L=10	J_1	100.0	*41.2*	100.0	100.0	89.6	100.0	*72.1*	98.3	100.0	100.0	100.0	*56.1*
	J_2	100.0	*19.7*	100.0	100.0	84.8	100.0	*44.9*	87.5	100.0	100.0	100.0	*23.4*
L=7	J_1	100.0	*30.8*	100.0	100.0	75.3	100.0	*64.8*	95.3	100.0	100.0	100.0	*53.7*
	J_2	100.0	*7.2*	100.0	100.0	63.5	99.8	*30.3*	74.6	100.0	99.8	100.0	*9.7*
L=4	J_1	100.0	*32.9*	100.0	100.0	77.2	100.0	*61.4*	96.2	100.0	100.0	100.0	*54.2*
	J_2	100.0	*5.5*	100.0	100.0	65.7	100.0	*21.6*	77.0	100.0	99.9	100.0	*5.4*
L=1	J_1	100.0	*45.5*	100.0	100.0	74.9	100.0	*38.6*	89.4	100.0	100.0	100.0	*40.2*
	J_2	100.0	*1.9*	99.9	100.0	55.3	99.4	*15.8*	70.4	100.0	99.9	100.0	*4.9*

Note: See note below Table 2.1.

Table 2.3: Rejection frequency of GMM test, level 10%, kernel matching, n=500

		DGP 1				DGP 2				DGP 3			
		φ_0	φ_1	φ_2	φ_3	φ_0	φ_1	φ_2	φ_3	φ_0	φ_1	φ_2	φ_3
L=14	J_1	100.0	*70.8*	99.9	100.0	59.6	92.4	*67.9*	77.5	99.4	99.3	100.0	*67.9*
	LM_1	94.7	*16.4*	50.9	98.7	6.1	24.2	*9.1*	6.2	57.2	70.1	90.3	*8.4*
	J_2	98.1	*60.2*	98.6	98.6	47.8	86.1	*47.4*	45.4	92.6	95.7	98.1	*43.8*
	LM_2	80.2	*9.1*	35.6	93.1	3.1	18.8	*5.0*	1.8	30.6	58.0	53.7	*3.9*
L=10	J_1	100.0	*60.8*	99.8	100.0	59.1	92.9	*71.5*	82.5	99.4	98.9	100.0	*63.9*
	LM_1	97.6	*22.2*	69.2	99.9	8.8	40.2	*14.1*	9.5	68.1	79.3	96.2	*9.0*
	J_2	97.7	*43.0*	98.2	98.6	47.7	86.8	*47.1*	44.0	91.4	94.8	98.7	*29.9*
	LM_2	81.1	*9.0*	45.4	94.0	4.5	23.4	*7.4*	2.3	28.4	61.3	59.0	*3.1*
L=7	J_1	100.0	*47.5*	91.4	100.0	50.6	87.4	*69.7*	81.2	99.5	94.2	99.9	*57.1*
	LM_1	98.4	*23.2*	59.1	99.9	10.3	51.0	*16.8*	12.7	75.0	79.0	96.7	*5.5*
	J_2	95.3	*22.1*	64.9	96.9	37.0	77.1	*37.4*	33.7	89.9	76.0	96.4	*12.0*
	LM_2	75.8	*5.7*	7.3	94.0	3.6	20.6	*5.9*	1.8	26.1	48.6	49.5	*1.0*
L=4	J_1	100.0	*40.5*	82.6	100.0	30.3	85.0	*62.3*	70.4	99.1	93.3	99.9	*55.5*
	LM_1	99.7	*30.4*	58.1	100.0	10.3	65.9	*38.3*	15.7	91.4	87.7	99.3	*12.5*
	J_2	94.1	*8.5*	49.5	97.3	16.7	68.1	*24.4*	14.2	86.1	72.7	96.6	*4.8*
	LM_2	77.1	*3.3*	7.8	92.9	2.2	15.0	*9.4*	0.8	35.6	49.4	65.9	*1.3*
L=1	J_1	100.0	*56.5*	77.3	100.0	25.4	77.7	*33.0*	36.9	99.5	91.9	99.9	*48.7*
	LM_1	100.0	*55.0*	71.6	100.0	35.2	83.6	*39.7*	44.5	99.4	91.8	99.9	*44.7*
	J_2	42.1	*0.5*	13.7	56.9	5.2	50.3	*7.8*	7.8	32.7	35.5	55.1	*0.4*
	LM_2	10.2	*0.1*	2.4	22.9	2.7	28.4	*5.0*	5.7	7.2	13.5	20.7	*0.1*

Note: LM_1 and LM_2 are the LM-statistics of the first and second step GMM estimator. See note below Table 2.1.

Table 2.4: Rejection frequency of GMM test, level 10%, kernel matching, n=2000

		DGP 1				DGP 2				DGP 3			
		φ_0	φ_1	φ_2	φ_3	φ_0	φ_1	φ_2	φ_3	φ_0	φ_1	φ_2	φ_3
L=14	J_1	100.0	*85.4*	100.0	100.0	87.9	100.0	*86.2*	99.0	100.0	100.0	100.0	*81.3*
	LM_1	100.0	*52.1*	100.0	100.0	10.1	98.2	*21.8*	30.4	100.0	99.8	100.0	*21.5*
	J_2	100.0	*77.0*	100.0	100.0	80.3	100.0	*67.4*	85.8	100.0	100.0	100.0	*56.7*
	LM_2	100.0	*38.5*	100.0	99.6	1.2	53.2	*2.7*	2.7	86.9	94.6	100.0	*7.9*
L=10	J_1	100.0	*68.2*	100.0	100.0	89.3	100.0	*88.2*	98.8	100.0	100.0	100.0	*71.2*
	LM_1	100.0	*44.0*	100.0	100.0	19.2	98.6	*37.8*	47.1	100.0	99.8	100.0	*13.7*
	J_2	100.0	*51.4*	100.0	100.0	79.0	100.0	*67.6*	86.6	100.0	100.0	100.0	*33.0*
	LM_2	99.8	*25.6*	99.4	99.8	2.8	37.6	*2.9*	4.3	79.3	93.7	100.0	*3.9*
L=7	J_1	100.0	*48.8*	100.0	100.0	85.0	100.0	*88.3*	98.7	100.0	100.0	100.0	*68.2*
	LM_1	100.0	*34.9*	99.5	100.0	29.4	99.4	*49.7*	68.8	100.0	99.8	100.0	*21.5*
	J_2	100.0	*22.0*	99.8	99.8	75.4	99.8	*63.4*	82.0	100.0	99.6	100.0	*19.1*
	LM_2	99.8	*9.3*	72.0	99.8	5.4	45.9	*2.1*	9.7	81.4	88.2	99.7	*1.3*
L=4	J_1	100.0	*43.8*	100.0	100.0	64.0	100.0	*71.3*	95.5	100.0	100.0	100.0	*52.6*
	LM_1	100.0	*40.0*	99.4	100.0	34.1	100.0	*62.0*	86.0	100.0	99.8	100.0	*39.1*
	J_2	100.0	*10.3*	99.2	100.0	39.1	99.8	*29.5*	50.9	100.0	100.0	100.0	*5.2*
	LM_2	99.5	*4.5*	77.6	99.8	0.8	49.7	*2.8*	4.8	87.0	90.2	99.9	*2.1*
L=1	J_1	100.0	*63.2*	100.0	100.0	58.5	100.0	*34.7*	85.0	100.0	97.6	100.0	*48.7*
	LM_1	100.0	*62.2*	100.0	100.0	65.8	100.0	*41.3*	87.6	100.0	97.6	100.0	*47.9*
	J_2	67.5	*0.0*	54.7	85.1	21.7	98.0	*7.8*	50.9	78.5	71.7	89.5	*0.2*
	LM_2	14.5	*0.0*	10.1	27.5	8.5	79.3	*3.0*	35.1	23.6	28.1	46.4	*0.0*

Note: See notes below Tables 2.1 and 2.3.

3 Treatment choice among Swedish rehabilitation programmes

In the following tables the results of a sensitivity analysis are presented, which show that the main findings are robust to changes in the specification. The Tables 3.1 to 3.5 indicate how the optimal allocation changes if different numbers of subpopulations (L=0, 1, 6, 16, 21) are included but the set of explanatory regressors retained unchanged. (The 38 regressors are enumerated in Table B.2.) The Tables 3.6 to 3.8 correspond to an analysis of three specifications where the number of subpopulations is kept at L=11 but the number of explanatory regressors is reduced.

Table 3.1 compares the optimal allocation estimated with L=11 subpopulations (=main specification) to the allocation that is obtained if no subpopulation moments are included (L=0). I.e. it compares the semiparametric estimator with overidentifying restrictions to fully parametric estimation. The columns (r_i^*) refer to the allocation according to the main specification (L=11), whereas the rows (r_i^{**}) refer to the allocation according to the alternative specification (L=0). E.g. the entry 3 in the first row indicates that of all the individuals whose optimal programme is medical rehabilitation, as estimated under the main specification, three would be advised to participate in No rehabilitation under the alternative specification (at level 1-α=0.7). Generally, it is seen that both specifications lead to rather similar predictions, with only few individuals in the off-diagonal classifications (apart from undefined cases). Leaving aside the undefined cases, the fraction of misclassification Δ measures the mismatch between both estimated allocations by the sum of the off-diagonal elements divided by the number of cases with defined optimal treatment (under both specifications). At the 0.7 level, only 1% of optimal programme predictions differ between the two specifications. At the 0.6 level, this increases to 4% and to 14% at the 0.5 level.

Also the implications with respect to the attainable employment rate under optimal allocation are quite similar for both specifications. Under the main specification, the predicted employment rate is 55.7% when all individuals are assigned to their optimal programme if it is defined (at 1-α=0.5) and otherwise randomly to No or workplace rehabilitation. If educational rehabilitation were no longer available, the predicted employment rate is 54.9%. Under the alternative specification, the respective figures are 56.9% and 55.2%.,

Table 3.1: Differences in the estimated optimal allocations, L=0 versus L=11

	r_i^* at 1-α=70%					r_i^* at 1-α=60%					r_i^* at 1-α=50%				
	N	W	E	M	U	N	W	E	M	U	N	W	E	M	U
r_i^{**}=N	516	5	0	3	334	764	17	1	8	330	1062	82	5	31	255
r_i^{**}=W	0	368	0	0	405	0	630	15	10	513	22	1034	86	58	454
r_i^{**}=E	0	0	228	0	290	1	13	429	2	314	11	84	687	22	246
r_i^{**}=M	5	0	0	96	228	17	1	2	195	325	58	28	26	355	329
r_i^{**}=U	97	167	66	81	3398	138	232	105	137	2088	149	158	101	140	804
Δ(%)			1.1					4.1					14.1		

Note: Number of individuals with estimated optimal programme $r_i^* \in \{No,Work,Edu,Med,Undefined\}$ under the main specification *and* estimated optimal programme $r_i^{**} \in \{No,Work,Edu,Med,Undefined\}$ under the alternative specification (L=0), at the level 1-α=0.7 (left), 0.6 (middle) and 0.5 (right). The columns/rows labelled U stand for undefined optimal programme. Δ gives the fraction of misclassification in %, i.e. the number of individuals for whom the optimal programme under the main specification (r_i^*) and under the alternative specification (r_i^{**}) do not coincide (off-diagonal elements) to the total number of individuals with defined optimal programme (under both specifications), leaving aside the undefined cases. The optimal choices r_i^{**} are simulated by 817 bootstrap replications.

Table 3.2 repeats the above analysis for an alternative specification with L=1 population included. Table 3.3 contains the results for an alternative specification with L=6, Table 3.4 for L=16 and Table 3.5 for L=21 populations. Generally, the estimated optimal treatment choices are very similar to those of the main specification. The mismatch is at most 0.1% at the 0.7 level, at most 2.4% at the 0.6 level and at most 11% at the 0.5 level. Furthermore, the maximal attainable employment rate is in all specifications estimated to be about 55-56% (when educational rehabilitation is not available) and about 54-55% (after elimination of educational rehabilitation). This shows that the optimal programme predictions are rather robust to the number of moments included.

Table 3.2: Differences in the estimated optimal allocations, L=1 versus L=11

	r_i^* at 1-α=70%					r_i^* at 1-α=60%					r_i^* at 1-α=50%				
	N	W	E	M	U	N	W	E	M	U	N	W	E	M	U
r_i^{**}=N	589	0	0	0	455	870	10	0	7	448	1212	72	4	43	339
r_i^{**}=W	0	394	0	1	349	0	660	8	7	454	10	1039	59	49	406
r_i^{**}=E	0	0	251	0	351	1	15	470	4	371	12	96	754	23	301
r_i^{**}=M	0	0	0	107	154	1	1	0	203	201	20	20	10	351	235
r_i^{**}=U	29	146	43	72	3346	48	207	74	131	2096	48	159	78	140	807
Δ(%)			0.1					2.4					11.1		

Note: See note below Table 3.1. Optimal choices r_i^{**} under the alternative specification (L=1 subpopulations).

Table 3.3: Differences in the estimated optimal allocations, L=6 versus L=11

	r_i^* at 1-α=70%					r_i^* at 1-α=60%					r_i^* at 1-α=50%				
	N	W	E	M	U	N	W	E	M	U	N	W	E	M	U
r_i^{**}=N	601	0	0	0	365	886	0	0	3	374	1240	27	6	20	322
r_i^{**}=W	0	462	0	0	333	0	742	0	6	391	12	1150	31	42	351
r_i^{**}=E	0	0	260	0	271	0	2	473	3	300	3	43	778	18	273
r_i^{**}=M	0	0	0	117	158	1	0	0	216	212	16	24	8	372	234
r_i^{**}=U	17	78	34	63	3528	33	149	79	124	2293	31	142	82	154	908
Δ(%)			0.0					0.6					6.6		

Note: See note below Table 3.1. Optimal choices r_i^{**} under the alternative specification (L=6 subpopulations).

Table 3.4: Differences in the estimated optimal allocations, L=16 versus L=11

	r_i^* at 1-α=70%					r_i^* at 1-α=60%					r_i^* at 1-α=50%				
	N	W	E	M	U	N	W	E	M	U	N	W	E	M	U
r_i^{**}=N	465	0	0	1	221	724	1	0	2	308	1074	36	18	23	291
r_i^{**}=W	0	364	0	0	140	0	655	0	4	220	18	1083	28	24	235
r_i^{**}=E	0	0	125	0	39	0	0	301	0	84	9	18	605	9	98
r_i^{**}=M	0	0	0	74	44	1	1	0	177	111	19	20	11	328	153
r_i^{**}=U	153	176	169	105	4211	195	236	251	169	2847	182	229	243	222	1311
Δ(%)			0.1					0.5					7.0		

Note: See note below Table 3.1. Optimal choices r_i^{**} under the alternative specification (L=16 subpopulations).

Table 3.5: Differences in the estimated optimal allocations, L=21 versus L=11

	r_i^* at 1-α=70%					r_i^* at 1-α=60%					r_i^* at 1-α=50%				
	N	W	E	M	U	N	W	E	M	U	N	W	E	M	U
r_i^{**}=N	456	0	0	1	279	724	5	2	6	368	1041	44	18	39	276
r_i^{**}=W	0	320	0	0	202	3	608	1	2	323	34	994	45	36	295
r_i^{**}=E	0	0	130	0	96	0	8	317	8	216	2	41	589	32	211
r_i^{**}=M	0	0	0	51	37	4	5	0	150	92	16	32	6	277	155
r_i^{**}=U	162	220	164	128	4041	189	267	232	186	2571	209	275	247	222	1151
Δ(%)			0.1					2.4					10.6		

Note: See note below Table 3.1. Optimal choices r_i^{**} under the alternative specification (L=21 subpopulations).

In the Tables 3.6 to 3.8 the optimal allocations according to different sets of explanatory regressors are examined. In Table 3.6 the allocation according to a specification with 30 variables (Specification A) is compared to the main specification. Table 3.7 contains this analysis for a specification with 28 variables (Specification B), and Table 3.8 corresponds to a Specification C with 24 variables. In all specifications the same subpopulations are included (L=11). Specification A differs from the main specification by leaving out the eight variables: *citizenship; white collar; occupation in sciences; previous sick-leave 31*

to 60 days; medical diagnosis: other; medical rehabilitation recommendation: wait&see; medical reasons prevented vocational rehabilitation; and *medical and case worker recommendation: vocational rehabilitation needed*. Specification B differs from the main specification by dropping the ten variables: *gender; low educated blue-collar; educated blue-collar; white collar; occupation in health care; occupation in sciences; sickness registration by private or other; medical diagnosis: psychiatric; medical diagnosis: other; medical reasons prevented vocational rehabilitation*. Hence in both specifications some socioeconomic characteristics are neglected as well as some supplementary information about the rehabilitation examination (in Specification A) or about sickness registration (in Specification B). Nevertheless, the most relevant indicators about the rehabilitation examination are kept. It is seen from Tables 3.6 and 3.7 that the resulting allocations are still quite similar to the main allocation and differ in about 0.5% of the defined cases at the 0.7 level, and about 5% (14.5%) at the 0.6 (0.5) level.

However, leaving out even more variables changes the estimated optimal allocation markedly. Table 3.8 provides the comparison between the allocation according to the main specification and Specification C, where additionally to those variables dropped in Specification A also the variables *income*, *occupation in health care*, *rehabilitation examination not needed* and all the *county* indicators are dropped. With this sparse specification the misclassification rates Δ are 15.8% and 26.4% at the 0.7 and 0.6 level, respectively, and increase to almost 40% at the 0.5 level.

Table 3.6: Differences in the estimated optimal allocations, main specification versus specification A

	r_i^* at 1-α=70%					r_i^* at 1-α=60%					r_i^* at 1-α=50%				
	N	W	E	M	U	N	W	E	M	U	N	W	E	M	U
r_i^{**}=N	346	1	0	0	157	569	14	12	0	209	854	53	40	11	197
r_i^{**}=W	1	339	2	0	394	18	595	34	1	528	115	981	129	13	470
r_i^{**}=E	0	1	112	0	147	2	6	248	0	215	27	65	473	21	230
r_i^{**}=M	0	0	1	119	123	0	4	5	235	180	10	18	29	418	187
r_i^{**}=U	271	199	179	61	3834	331	274	253	116	2438	296	269	234	143	1004
Δ(%)			0.7					5.5					16.3		

Note: See note below Table 3.1. Optimal choices r_i^{**} under the alternative specification A (with 30 variables).

Table 3.7: Differences in the estimated optimal allocations, main specification versus specification B

	r_i^* at 1-α=70%					r_i^* at 1-α=60%					r_i^* at 1-α=50%				
	N	W	E	M	U	N	W	E	M	U	N	W	E	M	U
r_i^{**}=N	437	0	0	0	298	688	17	9	0	392	1011	78	32	9	340
r_i^{**}=W	1	378	4	0	403	16	649	25	3	519	70	1037	118	30	449
r_i^{**}=E	0	0	122	0	86	1	1	263	0	124	14	31	511	5	145
r_i^{**}=M	0	0	0	131	93	2	2	4	261	135	21	14	21	439	170
r_i^{**}=U	180	162	168	49	3775	213	224	251	88	2400	186	226	223	123	984
Δ(%)			0.5					4.1					12.9		

Note: See note below Table 3.1. Optimal choices r_i^{**} under the alternative specification B (with 28 variables).

Table 3.8: Differences in the estimated optimal allocations, main specification versus specification C

	r_i^* at 1-α=70%					r_i^* at 1-α=60%					r_i^* at 1-α=50%				
	N	W	E	M	U	N	W	E	M	U	N	W	E	M	U
r_i^{**}=N	270	6	7	14	329	450	21	20	36	382	658	67	61	89	299
r_i^{**}=W	2	263	10	1	347	27	451	52	9	504	92	787	181	38	433
r_i^{**}=E	4	1	66	3	167	14	13	169	19	267	56	86	369	74	272
r_i^{**}=M	23	25	22	28	465	75	83	46	89	547	186	203	111	199	459
r_i^{**}=U	319	245	189	134	3347	354	325	265	199	1870	310	243	183	206	625
Δ(%)			15.8					26.4					38.2		

Note: See note below Table 3.1. Optimal choices r_i^{**} under the alternative specification C (with 24 variables).

References

Back, K., and D. Brown (1993): "Implied Probabilities in GMM Estimators," *Econometrica*, 61, 971–976.

Brown, B., and W. Newey (2002): "Generalized Method of Moments, Efficient Bootstrapping, and Improved Inference," *Journal of Business and Economic Statistics*, 20, 507–517.

Brown, B., W. Newey, and S. May (2001): "Bootstrapping with Moment Restrictions," mimeo, Rice University and MIT.

Greene, W. (1997): *Econometric Analysis*. Prentice Hall, New Jersey, 3 edn.

Hall, P., and J. Horowitz (1996): "Bootstrap Critical Values for Tests based on Generalized-Method-of-Moments Estimators," *Econometrica*, 64, 891–916.

Heckman, J., H. Ichimura, and P. Todd (1998): "Matching as an Econometric Evaluation Estimator," *Review of Economic Studies*, 65, 261–294.

Hoeffding, W. (1948): "A Class of Statistics with Asymptotically Normal Distribution," *Annals of Mathematical Statistics*, 19, 293–325.

Imbens, G., R. Spady, and P. Johnson (1998): "Information theoretic approaches to inference in moment condition models," *Econometrica*, 66, 333–357.

Inkmann, J. (2001): *Conditional Moment Estimation of Nonlinear Equation Systems*. Springer Verlag, Heidelberg.

Newey, W., and D. McFadden (1994): "Large Sample Estimation and Hypothesis Testing," in *Handbook of Econometrics*, ed. by R. Engle, and D. McFadden. Elsevier, Amsterdam.

Pagan, A., and A. Ullah (1999): *Nonparametric Econometrics*. Cambridge University Press, Cambridge.

Parzen, E. (1962): "On Estimation of a Probability Density and Mode," *Annals of Mathematical Statistics*, 33, 1065–1076.

Powell, J., J. Stock, and T. Stoker (1989): "Semiparametric Estimation of Index Coefficients," *Econometrica*, 57, 1403–1430.

Serfling, R. (1980): *Approximation Theorems of Mathematical Statistics*. Wiley, New York.