7 Semiparametric Methods and Partially Linear Regression

7.1 Overview

A model is called semiparametric if it is described by θ and τ where θ is finite-dimensional (e.g. parametric) and τ is infinite-dimensional (nonparametric). All moment condition models are semiparametric in the sense that the distribution of the data (τ) is unspecified and infinite dimensional. But the settings more typically called "semiparametric" are those where there is explicit estimation of τ.

In many contexts the nonparametric part τ is a conditional mean, variance, density or distribution function.

Often θ is the parameter of interest, and τ is a nuisance parameter, but this is not necessarily the case.

In many semiparametric contexts, τ is estimated first, and then $\hat{\theta}$ is a two-step estimator. But in other contexts (θ, τ) are jointly estimated.

7.2 Feasible Nonparametric GLS

A classic semiparametric model, which is not in Li-Racine, is feasible GLS with unknown variance function. The seminal papers are Carroll (1982, Annals of Statistics) and Robinson (1987, Econometrica). The setting is a linear regression

$$\begin{aligned} y_i &= X_i'\theta + e_i \\ \mathrm{E}\left(e_i \mid X_i\right) &= 0 \\ \mathrm{E}\left(e_i^2 \mid X_i\right) &= \sigma^2\left(X_i\right) \end{aligned}$$

where the variance function $\sigma^2(x)$ is unknown but smooth in x, and $x \in \mathbb{R}^q$. (The idea also applies to non-linear but parametric regression functions). In this model, the nonparametric nuisance parameter is $\tau = \sigma^2(\cdot)$

As the model is a regression, the efficiency bound for θ is attained by GLS regression

$$\tilde{\theta} = \left(\sum_{i=1}^{n} \frac{1}{\sigma^2\left(X_i\right)} X_i X_i'\right)^{-1} \left(\sum_{i=1}^{n} \frac{1}{\sigma^2\left(X_i\right)} X_i y_i\right).$$

This of course is infeasible. Carroll and Robinson suggested replacing $\sigma^2\left(X_i\right)$ with $\hat{\sigma}^2\left(X_i\right)$ where $\hat{\sigma}^2(x)$ is a nonparametric estimator. (Carroll used kernel methods; Robinson used nearest neighbor methods.) Specifically, letting $\hat{\sigma}^2(x)$ be the NW estimator of $\sigma^2(x)$, we can define the feasible estimator

$$\hat{\theta} = \left(\sum_{i=1}^{n} \frac{1}{\hat{\sigma}^2\left(X_i\right)} X_i X_i'\right)^{-1} \left(\sum_{i=1}^{n} \frac{1}{\hat{\sigma}^2\left(X_i\right)} X_i y_i\right).$$

This seems sensible. The question is find its asymptotic distribution, and in particular to find

if it is asymptotically equivalent to $\tilde{\theta}$.

7.3 Generated Regressors

The model is

$$\begin{aligned} y_i &= \theta\tau(X_i) + e_i \\ \mathrm{E}\left(e_i \mid X_i\right) &= 0 \end{aligned}$$

where θ is finite dimensional but τ is an unknown function. Suppose τ is identified by another equation so that we have a consistent estimate of $\hat{\tau}(x)$ for $\tau(x)$.(Imagine a non-parametric Heckman estimator).

Then we could estimate θ by least-squares of y_i on $\hat{\tau}(Z_i)$.

This problem is called generated regressors, as the regressor is a (consistent) estimate of a infeasible regressor.

In general, $\hat{\theta}$ is consistent. But what is its distribution?

7.4 Andrews' MINPIN Theory

A useful framework to study the type of problem from the previous section is given in Andrews (Econometrica, 1994). It is reviewed in section 7.3 of Li-Racine but the discussion is incomplete and there is at least one important omission. If you really want to learn this theory I suggest reading Andrews' paper.

The setting is when the estimator $\hat{\theta}$ MINimizes a criterion function which depends on a Preliminary Infinite dimensional Nuisance parameter estimator, hence MINPIN.

Let $\theta \in \Theta$ be the parameter of interest, let $\tau \in T$ denote the infinite-dimensional nuisance parameter. Let θ_0 and τ_0 denote the true values.

Let $\hat{\tau}$ be a first-step estimate of τ, and assume that it is consistent: $\hat{\tau} \to_p \tau_0$

Now suppose that the criterion function for estimation of θ depends on the first-step estimate $\hat{\tau}$. Let the criterion be $Q_n\left(\theta, \tau\right)$ and suppose that

$$\hat{\theta} = \underset{\theta}{\operatorname{argmin}}\, Q_n\left(\theta, \hat{\tau}\right)$$

Thus $\hat{\theta}$ is a two-step estimator.

Assume that

$$\begin{aligned} \frac{\partial}{\partial\theta} Q_n\left(\theta, \tau\right) &= \bar{m}_n\left(\theta, \tau\right) + o_p\left(\frac{1}{\sqrt{n}}\right) \\ \bar{m}_n\left(\theta, \tau\right) &= \frac{1}{n}\sum_{i=1}^{n} m_i\left(\theta, \tau\right) \end{aligned}$$

where $m_i(\theta, \tau)$ is a function of the i'th observation. In just-identified models, there is no $o_p\left(\frac{1}{\sqrt{n}}\right)$ error (and we now ignore the presence of this error).

In the FGLS example, $\tau(\cdot) = \sigma^2(\cdot)$, $\hat{\tau}(\cdot) = \hat{\sigma}^2(\cdot)$, and

$$m_i(\theta, \tau) = \frac{1}{\tau(X_i)} X_i \left(y_i - X_i'\theta\right).$$

In the generated regressor problem,

$$m_i(\theta, \tau) = \tau(X_i)\left(y_i - \theta\tau(X_i)\right).$$

In general, the first-order condition (FOC) for $\hat{\theta}$ is

$$0 = \bar{m}_n\left(\hat{\theta}, \hat{\tau}\right)$$

Assume $\hat{\theta} \to_p \theta_0$. Implicit to obtain consistency is the requirement that the population expecation of the FOC is zero, namely

$$\mathrm{E} m_i(\theta_0, \tau_0) = 0$$

and we assume that this is the case.

We expand the FOC in the first argument

$$\begin{aligned} 0 &= \sqrt{n}\bar{m}_n\left(\hat{\theta}, \hat{\tau}\right) \\ &= \sqrt{n}\bar{m}_n(\theta_0, \hat{\tau}) + M_n(\theta_0, \hat{\tau})\sqrt{n}\left(\hat{\theta} - \theta_0\right) + o_p(1) \end{aligned}$$

where

$$M_n(\theta, \tau) = \frac{\partial}{\partial \theta'}\bar{m}_n(\theta, \tau)$$

It follows that

$$\sqrt{n}\left(\hat{\theta} - \theta_0\right) \simeq -M_n(\theta_0, \hat{\tau})^{-1}\sqrt{n}\bar{m}_n(\theta_0, \hat{\tau}).$$

If $M_n(\theta, \tau)$ converges to its expectation

$$M(\theta, \tau) = \mathrm{E}\frac{\partial}{\partial \theta'} m_i(\theta, \tau)$$

uniformly in its arguments, then $M_n(\theta_0, \hat{\tau}) \to_p M(\theta_0, \tau_0) = M$, say. Then

$$\sqrt{n}\left(\hat{\theta} - \theta_0\right) \simeq -M^{-1}\sqrt{n}\bar{m}_n(\theta_0, \hat{\tau}).$$

We cannot take a Taylor expansion in τ because it is infinite dimensional. Instead, Andrews uses a stochastic equicontinuity argument. Define the population version of $\bar{m}_n(\theta, \tau)$

$$m(\theta, \tau) = \mathrm{E} m_i(\theta, \tau).$$

Note that at the true values $m(\theta_0, \tau_0) = 0$ (as discussed above) but the function is non-zero for generic values.

Define the function

$$\begin{aligned} \nu_n(\tau) &= \sqrt{n}\left(\bar{m}_n(\theta_0, \tau) - m(\theta_0, \tau)\right) \\ &= \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left(m_i(\theta_0, \tau) - \mathrm{E} m_i(\theta_0, \tau)\right) \end{aligned}$$

Notice that $\nu_n(\tau)$ is a normalized sum of mean-zero random variables. Thus for any fixed τ, $\nu_n(\tau)$ converges to a normal random vector. Viewed as a function of τ, we might expect $\nu_n(\tau)$ to vary smoothly in the argument τ. The stochastic formulation of this is called stochastic equicontinuity. Roughly, as $n \to \infty$, $\nu_n(\tau)$ remains well-behaved as a function of τ. Andrews first key assumption is that $\nu_n(\tau)$ is stochastically equicontinuous.

The important implication of stochastic equicontinuity is that $\hat{\tau} \to_p \tau_0$ implies

$$\nu_n(\hat{\tau}) - \nu_n(\tau_0) \to_p 0.$$

Intuitively, if $g(\tau)$ is continuous, then $g(\hat{\tau}) - g(\tau_0) \to_p 0$. More generally, if $g_n(\tau)$ converges in probability uniformly to a continuous function $g(\tau)$ then $g_n(\hat{\tau}) - g(\tau_0) \to_p 0$. The case of stochastic equicontinuity is the most general, but still has the same implication.

Since $\mathrm{E} m_i(\theta_0, \tau_0) = 0$, when we evaluate the empirical process at the true value τ_0 we have a zero-mean normalized sum, which is asymptotically normal by the CLT:

$$\begin{aligned} \nu_n(\tau_0) &= \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left(m_i(\theta_0, \tau_0) - \mathrm{E} m_i(\theta_0, \tau_0)\right) \\ &= \frac{1}{\sqrt{n}} \sum_{i=1}^{n} m_i(\theta_0, \tau_0) \\ &\to_d N(0, \Omega) \end{aligned}$$

where

$$\Omega = \mathrm{E} m_i(\theta_0, \tau_0)\, m_i(\theta_0, \tau_0)'$$

It follows that

$$\nu_n(\hat{\tau}) = \nu_n(\tau_0) + o_p(1) \to_d N(0, \Omega)$$

and thus

$$\begin{aligned} \sqrt{n}\bar{m}_n(\theta_0, \hat{\tau}) &= \sqrt{n}\left(\bar{m}_n(\theta_0, \hat{\tau}) - m(\theta_0, \hat{\tau})\right) + \sqrt{n} m(\theta_0, \hat{\tau}) \\ &= \nu_n(\hat{\tau}) + \sqrt{n} m(\theta_0, \hat{\tau}) \end{aligned}$$

The final detail is what to do with $\sqrt{n}m(\theta_0, \hat{\tau})$. Andrews directly assumes that

$$\sqrt{n}m(\theta_0, \hat{\tau}) \to_p 0$$

This is the second key assumption, and we discuss it below. Under this assumption,

$$\sqrt{n}\bar{m}_n(\theta_0, \hat{\tau}) \to_d N(0, \Omega)$$

and combining this with our earlier expansion, we obtain:

Andrews MINPIN Theorem. Under Assumptions 1-6 below,

$$\sqrt{n}\left(\hat{\theta} - \theta_0\right) \to_p N(0, V)$$

where

$$\begin{aligned} V &= M^{-1}\Omega M^{-1\prime} \\ M &= \mathrm{E}\frac{\partial}{\partial\theta'}m_i(\theta_0, \tau_0) \\ \Omega &= \mathrm{E}m_i(\theta_0, \tau_0)\, m_i(\theta_0, \tau_0)' \end{aligned}$$

Assumption 1 *As $n \to \infty$, for $m(\theta, \tau) = \mathrm{E}m_i(\theta, \tau)$*

1. $\hat{\theta} \to_p \theta_0$
2. $\hat{\tau} \to_p \tau_0$
3. $\sqrt{n}m(\theta_0, \hat{\tau}) \to_p 0$
4. $m(\theta_0, \tau_0) = 0$
5. $\nu_n(\tau)$ *is stochastically equicontinuous at* τ_0.
6. $\bar{m}_n(\theta, \tau)$ *and* $\dfrac{\partial}{\partial\theta}\bar{m}_n(\theta, \tau)$ *satisfy uniform WLLN's over* $\Theta \times T$. *(They converge in probability to their expectations, uniformly over the parameter space.)*

Discussion of result: The theorem says that the semiparametric estimator $\hat{\theta}$ has the same asymptotic distribution as the idealized estimator obtained by replacing the nonparametric estimate $\hat{\tau}$ with the true function τ_0. Thus the estimator is adaptive. This might seem too good to be true. The key is assumption, which holds in some cases, but not in others.

Discussion of assumptions.

Assumptions 1 and 2 state that the estimators are consistent, which should be separately verified. Assumption 4 states that the FOC identifies θ when evaluated at the true τ_0. Assumptions 5 and 6 are regularity conditions, essentially smoothness of the underlying functions, plus sufficient moments.

7.5 Orthogonality Assumption

The key assumption 3 for Andrews' MINPIN theorey was somehow missed in the write-up in Li-Racine. Assumption 3 is not always true, and is not just a regularity condition. It requires a sort of orthogonality between the estimation of θ and τ.

Suppose that τ is finite-dimensional. Then by a Taylor expansion

$$\begin{aligned}\sqrt{n}m\left(\theta_0,\hat{\tau}\right) &\simeq \sqrt{n}m\left(\theta_0,\tau_0\right)+\frac{\partial}{\partial\tau}m\left(\theta_0,\tau\right)'\sqrt{n}\left(\hat{\tau}-\tau_0\right)\\ &= \frac{\partial}{\partial\tau}m\left(\theta_0,\tau_0\right)'\sqrt{n}\left(\hat{\tau}-\tau_0\right)\end{aligned}$$

since $m\left(\theta_0,\tau_0\right)=0$ by Assumption 4. Since τ is parametric, we should expect $\sqrt{n}\left(\hat{\tau}-\tau_0\right)$ to converge to a normal vector. Thus this expression will converge in probability to zero only if

$$\frac{\partial}{\partial\tau}m\left(\theta_0,\tau_0\right)=0$$

Recall that $m\left(\theta,\tau\right)$ is the expectation of the FOC, which is the derivative of the criterion wrt θ. Thus $\dfrac{\partial}{\partial\tau}m\left(\theta,\tau\right)$ is the cross-derivative of the criterion, and the above statement is that this cross-derivative is zero, which is an orthogonality condition (e.g. block diagonality of the Hessian.)

Now when τ is infinite-dimensional, the above argument does not work, but it lends intuition.

An analog of the derivative condition, which is sufficient for Assumption 3, is that

$$m\left(\theta_0,\tau\right)=0$$

for all τ in a neighborhood of τ_0.

In the FGLS example,

$$m\left(\theta_0,\tau\right)=\mathrm{E}\left(\frac{1}{\tau\left(X_i\right)}X_ie_i\right)=0$$

for all τ, so this is Assumption 3 is satisfied in this example. We have the implication:

Theorem. Under regularity conditions the Feasible nonparametric GLS estimator of the previous section is asymptotically equivalent to infeasible GLS.

In the generated regressor example

$$\begin{aligned}m\left(\theta_0,\tau\right) &= \mathrm{E}\left(\tau(X_i)\left(y_i-\theta_0\tau(X_i)\right)\right)\\ &= \mathrm{E}\left(\tau(X_i)\left(e_i+\theta_0\left(\tau_0(X_i)-\tau(X_i)\right)\right)\right)\\ &= \theta_0\mathrm{E}\left(\tau(X_i)\left(\tau_0(X_i)-\tau(X_i)\right)\right)\\ &= \theta_0\int\tau(x)\left(\tau_0(x)-\tau(x)\right)f(x)dx\end{aligned}$$

Assumption 3 requires $\sqrt{n}m\left(\theta_0,\hat{\tau}\right)\to_p 0$. But note that $\sqrt{n}m\left(\theta_0,\hat{\tau}\right)\simeq\sqrt{n}\left(\tau_0(x)-\hat{\tau}(x)\right)$ which certainly does not converge to zero. Assumption 3 is generically violated when there are generated regressors.

There is one interesting exception. When $\theta_0 = 0$ then $m(\theta_0, \tau) = 0$ and thus $\sqrt{n}m(\theta_0, \hat{\tau}) = 0$ so Assumption 3 is satisified.

We see that Andrews' MINPIN assumption do not apply in all semiparametric models. Only those which satisfy Assumption 3, which needs to be verified. The other key condition is stochastic equicontinuity, which is difficult to verify but is genearly satisfied for "well-behaved" estimators. The remaining assumptions are smoothness and regularity conditions, and typically are not of concern in applications.

7.6 Partially Linear Regression Model

The semiparametric partially linear regression model is

$$\begin{aligned} y_i &= X_i'\beta + g(Z_i) + e_i \\ \mathrm{E}(e_i \mid X_i, Z_i) &= 0 \\ \mathrm{E}\left(e_i^2 \mid X_i = x, Z_i = z\right) &= \sigma^2(x, z) \end{aligned}$$

That is, the regressors are (X_i, Z_i), and the model specifies the conditional mean as linear in X_i but possibly non-linear in $Z_i \in \mathbb{R}^q$. This is a very useful compromise between fully nonparametric and fully parametric. Often the binary (dummy) variables are put in X_i. Often there is just one nonlinear variable: $q = 1$, to keep things simple.

The goal is to estimate β and g, and to obtain confidence intervals.

The first issue to consider is identification. Since g is unconstrained, the elements of X_i cannot be collinear with any function of Z_i. This means that we must exclude from X_i intercepts and any deterministic function of Z_i. The function g includes these components.

7.7 Robinson's Transformation

Robinson (Econometrica, 1988) is the seminal treatment of the partially linear model. He first contribution is to show that we can concentrate out the unknown g by using a genearlization of residual regression.

Take the equation

$$y_i = X_i'\beta + g(Z_i) + e_i$$

and apply the conditional expectation operator $\mathrm{E}(\cdot \mid Z_i)$. We obtain

$$\begin{aligned} \mathrm{E}(y_i \mid Z_i) &= \mathrm{E}\left(X_i'\beta \mid Z_i\right) + \mathrm{E}(g(Z_i) \mid Z_i) + \mathrm{E}(e_i \mid Z_i) \\ &= \mathrm{E}(X_i \mid Z_i)'\beta + g(Z_i) \end{aligned}$$

(using the law of iterated expectations). Defining the conditional expectations

$$\begin{aligned} g_y(z) &= \mathrm{E}(y_i \mid Z_i = z) \\ g_x(z) &= \mathrm{E}(X_i \mid Z_i = z) \end{aligned}$$

We can write this expression as

$$g_y(z) = g_x(z)'\beta + g(z)$$

Subtracting from the original equation, the function g disappears:

$$y_i - g_y(Z_i) = (X_i - g_x(Z_i))'\beta + e_i$$

We can write this as

$$\begin{aligned} e_{yi} &= e_{xi}'\beta + e_i \\ y_i &= g_y(Z_i) + e_{yi} \\ X_i &= g_x(Z_i) + e_{xi} \end{aligned}$$

That is, β is the coefficient of the regression of e_{yi} on e_{xi}, where these are the conditional expectations errors from the regression of y_i (and X_i) on Z_i only.

This is a conditional expecation generalization of the idea of residual regression.

This transformed equation immediately suggests an infeasible estimator for β, by LS of e_{yi} on e_{xi} :

$$\tilde{\beta} = \left(\sum_{i=1}^{n} e_{xi}e_{xi}'\right)^{-1} \sum_{i=1}^{n} e_{xi}e_{yi}$$

7.8 Robinson's Estimator

Robinson suggested first estimating g_y and g_x by NW regression, using these to obtain residuals $\hat{e}_{xi}$ and $\hat{e}_{xi}$, and replacing these in the formula for $\tilde{\beta}$.

Specifically, let $\hat{g}_y(z)$ and $\hat{g}_x(z)$ denote the NW estimates of the conditional mean of y_i and X_i given $Z_i = z$. Assuming $q = 1$,

$$\hat{g}_y(z) = \frac{\sum_{i=1}^{n} k\left(\frac{Z_i - z}{h}\right) y_i}{\sum_{i=1}^{n} k\left(\frac{Z_i - z}{h}\right)}$$

$$\hat{g}_x(z) = \frac{\sum_{i=1}^{n} k\left(\frac{Z_i - z}{h}\right) X_i}{\sum_{i=1}^{n} k\left(\frac{Z_i - z}{h}\right)}$$

The estimator $\hat{g}_x(z)$ is a vector of the same dimension as X_i.

Notice that you are regressing each variable (the y_i and the X_i's) separately on the continuous variable Z_i. You should view each of these regressions as a separate NW regression. You should probably use a different bandwidth h for each of these regressions, as the depdendence on Z_i will depend on the variable. (For example, some regressors X_i might be independent of Z_i so you would want to use an infinite bandwidth for those cases.) While Robinson discussed NW regression, this is not essential. You could substitute LL or WNW instead.

Given the regression functions, we obtain the regression residuals

$$\begin{aligned} \hat{e}_{yi} &= y_i - \hat{g}_y(Z_i) \\ \hat{e}_{xi} &= X_i - \hat{g}_x(Z_i) \end{aligned}$$

Our first attempt at an estimator for β is then

$$\hat{\beta} = \left(\sum_{i=1}^{n} \hat{e}_{xi} \hat{e}_{xi}' \right)^{-1} \sum_{i=1}^{n} \hat{e}_{xi} \hat{e}_{yi}$$

7.9 Trimming

The asymptotic theory for semiparametric estimators, typically requires that the first step estimator converges uniformly at some rate. The difficulty is that $\hat{g}_y(z)$ and $\hat{g}_x(z)$ do not converge uniformly over unbounded sets. Equivalently, the problem is due to the estimated density of Z_i in the denominator. Another way of viewing the problem is that these estimates are quite noisy in sparse regions of the sample space, so residuals in such regions are noisy, and this could unduely influence the estimate of β.

suffer a problem that they can be unduly influence by unstable residuals from observations in sparse regions of the sample space. The nonparametric regression estimates depend inversely on the density estimate

$$\hat{f}_z(z) = \frac{1}{nh} \sum_{i=1}^{n} k\left(\frac{Z_i - z}{h} \right).$$

For values of z where $f_z(z)$ is close to zero, $\hat{f}_z(z)$ is not bounded away from zero, so the NW estimates at this point can be poor. Consequently the residuals for observations i such that $f_z(Z_i)$ will be quite unreliable, and can have an undue influence on $\hat{\beta}$.

A standard solution is to introduce "trimming". Let

$$\hat{f}_z(z) = \frac{1}{nh} \sum_{i=1}^{n} k\left(\frac{Z_i - z}{h} \right),$$

let $b > 0$ be a trimming constant and let $1_i(b) = 1\left(\hat{f}_z(Z_i) \geq b\right)$ denote a indicator variable for those observations for which the estimated density of Z_i is above b.

The trimmed version of $\hat{\beta}$ is

$$\hat{\beta} = \left(\sum_{i=1}^{n} \hat{e}_{xi} \hat{e}_{xi}' 1_i(b) \right)^{-1} \sum_{i=1}^{n} \hat{e}_{xi} \hat{e}_{yi} 1_i(b)$$

This is a trimmed LS residual regression.

The asymptotic theory requires that $b = b_n \to 0$ but unfortunately there is not good guidance about how to select b in practice. Often trimming is ignored in applications. One practical suggestion is to estimate β with and without trimming to assess robustness.

7.10 Asymptotic Distribution

Robinson (1988), Andrews (1994) and Li (1996) are references. The needed regularity conditions are that the data are iid, Z_i has a density, and the regression functions, density, and conditional variance function are sufficiently smooth with respect to their arguments. Assuming a second-order kernel, and for simplicity writing $h = h_1 = \cdots = h_q$, the important condition on the bandwidths sequence is

$$\sqrt{n} \left(h^4 + \frac{1}{nh^q} \right) \to 0$$

Technically, this is not quite enough, as this ignores the interaction with the trimming parameter b. But since this can be set $b_n \to 0$ at an extremely slow rate, it can be safely ignored. The above condition is similar to the standard convergence rates for nonparametric estimation, multiplied by $\sqrt{n}$. Equivalently, what is essential is that the uniform MSE of the nonparametric estimators converge faster than $\sqrt{n}$, or that the estimators themselves converges faster than $n^{-1/4}$. That is, what we need is

$$n^{-1/4} \sup_z |\hat{g}_y(z) - g_y(z)| \to_p 0$$

$$n^{-1/4} \sup_z |\hat{g}_x(z) - g_x(z)| \to_p 0$$

From the theory for nonparametric regression, these rates hold when bandwidths are picked optimally and $q \leq 3$.

In practice, $q \leq 3$ is probably sufficient. If $q > 3$ is desired, then higher-order kernels can be used to improve the rate of convergence. So long as the rate is faster than $n^{-1/4}$, the following result applies.

Theorem (Robinson). Under regularity conditions, including $q \leq 3$, the trimmed estimator satisfies

$$\sqrt{n} \left(\hat{\beta} - \beta \right) \to_d N(0, V)$$

$$V = \left(E\left(e_{xi} e_{xi}' \right) \right)^{-1} \left(E\left(e_{xi} e_{xi}' \sigma^2(X_i, Z_i) \right) \right)^{-1} \left(E\left(e_{xi} e_{xi}' \right) \right)^{-1}$$

That is, $\hat{\beta}$ is asymptotically equivalent to the infeasible estimator $\tilde{\beta}$.

The variance matrix may be estimated using conventional LS methods.

7.11 Verification of Andrews' MINPIN Condition

This Theorem states that Robinson's two-step estimator for β is asymptotically equivalent to the infeasible one-step estimator. This is an example of the application of Andrews' MINPIN theory. Andrews specifically mentions that the $n^{-1/4}$ convergence rates for $\hat{g}_y(z)$ *and* $\hat{g}_x(z)$ are essential to obtain this result.

To see this, note that the estimator $\hat{\beta}$ solves the FOC

$$\frac{1}{n}\sum_{i=1}^{n}\left(X_i-\hat{g}_x\left(Z_i\right)\right)\left(y_i-\hat{g}_y\left(Z_i\right)-\hat{\beta}'\left(X_i-\hat{g}_x\left(Z_i\right)\right)\right)=0$$

In Andrews MINPIN notation, let $\tau_x=\hat{g}_x$ and $\tau_y=\hat{g}_y$ denote fixed (function) values of the regression estimates, then

$$m_i\left(\theta_0,\tau\right)=\left(X_i-\tau_x\left(Z_i\right)\right)\left(y_i-\tau_y\left(Z_i\right)-\theta_0'\left(X_i-\tau_x\left(Z_i\right)\right)\right)$$

Since

$$y_i=g_y\left(Z_i\right)+\left(X_i-g_x\left(Z_i\right)\right)'\beta+e_i$$

then

$$\mathrm{E}\left(y_i-\tau_y\left(Z_i\right)-\theta_0'\left(X_i-\tau_x\left(Z_i\right)\right)\mid X_i,Z_i\right)=g_y\left(Z_i\right)-\tau_y\left(Z_i\right)-\left(g_x\left(Z_i\right)-\tau_x\left(Z_i\right)+\right)'\theta_0$$

and

$$\begin{aligned}
m\left(\theta_0,\tau\right) &= \mathrm{E}m_i\left(\theta_0,\tau\right)\\
&= \mathrm{EE}\left(m_i\left(\theta_0,\tau\right)\mid X_i,Z_i\right)\\
&= \mathrm{E}\left(\left(X_i-\tau_x\left(Z_i\right)\right)\left(g_y\left(Z_i\right)-\tau_y\left(Z_i\right)-\left(g_x\left(Z_i\right)-\tau_x\left(Z_i\right)\right)'\theta_0\right)\right)\\
&= \mathrm{E}\left(\left(g_x\left(Z_i\right)-\tau_x\left(Z_i\right)\right)\left(g_y\left(Z_i\right)-\tau_y\left(Z_i\right)-\left(g_x\left(Z_i\right)-\tau_x\left(Z_i\right)\right)'\theta_0\right)\right)\\
&= \int\left(\left(g_x\left(z\right)-\tau_x\left(z\right)\right)\left(g_y\left(z\right)-\tau_y\left(z\right)-\left(g_x\left(z\right)-\tau_x\left(z\right)\right)'\theta_0\right)\right)f_z(z)dz
\end{aligned}$$

where the second-to-last line uses conditional expecations given X_i. Then replacing τ_x with $\hat{g}_x$ and τ_y with $\hat{g}_y$

$$\sqrt{n}m\left(\theta_0,\hat{\tau}\right)=\int\left(\left(g_x\left(z\right)-\hat{g}_x\left(z\right)\right)\left(g_y\left(z\right)-\hat{g}_y\left(z\right)-\left(g_x\left(z\right)-\hat{g}_x\left(z\right)\right)'\theta_0\right)\right)f_z(z)dz$$

Taking bounds

$$\left|\sqrt{n}m\left(\theta_0,\hat{\tau}\right)\right|\leq\left(\sup_z\left|g_x\left(z\right)-\hat{g}_x\left(z\right)\right|\sup_z\left|g_y\left(z\right)-\hat{g}_y\left(z\right)\right|+\sup_z\left|g_x\left(z\right)-\hat{g}_x\left(z\right)\right|^2\right)\sup_z\left|f_y\left(z\right)\right|\rightarrow_p 0$$

when the nonparametric regression estimates converge faster than $n^{-1/4}$.

Indeed, we see that the $o(n^{-1/4})$ convergence rates imply the key condition $\sqrt{n}m(\theta_0, \hat{\tau}) \to_p 0$.

7.12 Estimation of Nonparametric Component

Recall that the model is

$$y_i = X_i'\beta + g(Z_i) + e_i$$

and the goal is to estimate β and g. We have described Robinson's estimator for β. We now discuss estimation of g.

Since $\hat{\beta}$ converges at rate $n^{-1/2}$ which is faster than a nonparametric rate, we can simply pretend that β is known, and do nonparametric regression of $y_i - X_i'\hat{\beta}$ on $Z_i - z$

$$\hat{g}(z) = \frac{\sum_{i=1}^n k\left(\frac{Z_i - z}{h}\right)\left(y_i - X_i'\hat{\beta}\right)}{\sum_{i=1}^n k\left(\frac{Z_i - z}{h}\right)}$$

The bandwidth $h = (h_1, ..., h_q)$ is distinct from those for the first-stage regressions.

It is not hard to see that this estimator is asymptotically equivalent to the infeasible regressor when $\hat{\beta}$ is replaced with the true β_0.

Standard errors for $\hat{g}(x)$ may be computed as for standard nonparametric regression.

7.13 Bandwidth Selection

In a semiparametric context, it is important to study the effect a bandwidth has on the performance of the estimator of interest before determining the bandwidth. In many cases, this requires a nonconventional bandwidth rate.

However, this problem does not occur in the partially linear model. The first-step bandwidths h used for $\hat{g}_y(z)$ and $\hat{g}_x(z)$ are inputs for calculation of $\hat{\beta}$. The goal is presumably accurate estimation of β. The bandwith h impacts the theory for $\hat{\beta}$ through the uniform convergence rates for $\hat{g}_y(z)$ and $\hat{g}_x(z)$ – suggesting that we use conventional bandwidth rules, e.g. cross-validation.