Nonlinear Parametric Smoothing

Robert A. Miller

Structural Econometrics

November 2026

Limitations of the Linear Model

Motivation and agenda

- The RDD specification described last lecture shows the linearity assumption accommodates nonlinear explanatory variables
- Nevertheless the linearity assumption is very restrictive because:
  1. Nonlinearity arises from selection on unobserved variables.
  2. The structure itself might be nonlinear in the parameters.
  3. We might be reluctant to place any parametric structure.
- Roughly speaking this lecture divides into four parts:
  1. giving examples that elaborate on the limitations listed above.
  2. defining estimators for nonlinear parametric models based on nonlinear instrumental variables, maximum likelihood, nonlinear least squares, generalized methods of moments, and order statistics.
  3. applying these estimators to continuous time point processes.
  4. combining these estimators in sequential estimation, for example as minimum distance estimators.

Limitations of the Linear Model

Selection on unobserved variables

- Consider the following static model of labor supply:

$$l_n = \begin{cases} \beta_0 + \beta_1 w_n + \epsilon_{1n} & \text{if } \beta_0 + \beta_1 w_n + \epsilon_{1n} > \epsilon_{2n} \\ 0 & \text{if } \beta_0 + \beta_1 w_n + \epsilon_{1n} \leq \epsilon_{2n} \end{cases}$$

  where $E[w_n \epsilon_{jn}] = E[\epsilon_{jn}] = 0$ for $j \in \{1, 2\}$ and:

  - l_n denotes the labour supply of individual n.
  - w_n denotes her wage offer.
  - ϵ_{1n} and ϵ_{2n} are unobserved amenity values associated with market work and staying home respectively.

- Let:

$$d_n \equiv \begin{cases} 1 \text{ if } l_n > 0 \\ 0 \text{ if } l_n = 0 \end{cases}$$

- Suppose the data comprise N observations on $(l_n, d_n w_n)$.

Limitations of the Linear Model

Nonlinearity induced by selection

- Conditioning on the wage:

$$E\left[l_n \left| d_n = 1, w_n \right.\right] = \beta_0 + \beta_1 w_n + E\left[\epsilon_{1n} \left| d_n = 1, w_n \right.\right]$$

- But:

$$\begin{aligned} \Pr\left\{d_n = 1 \left| w_n \right.\right\} &= \int\int_{\beta_0+\beta_1 w_n > \epsilon_{2n}-\epsilon_{1n}} dF\left(\epsilon_{1n}, \epsilon_{2n}\right) \\ &= \int_{\epsilon_{2n}=-\infty}^{\epsilon_{2n}=\infty} \int_{\epsilon_{1n}=\epsilon_{2n}-\beta_0-\beta_1 w_n}^{\epsilon_{1n}=\infty} dF\left(\epsilon_{1n}, \epsilon_{2n}\right) \end{aligned}$$

- Hence $E\left[\epsilon_{1n} \left| d_n = 1, w_n \right.\right] =$

$$\Pr\left\{d_n = 1 \left| w_n \right.\right\}^{-1} \int_{\epsilon_{2n}=-\infty}^{\epsilon_{2n}=\infty} \int_{\epsilon_{1n}=\epsilon_{2n}-\beta_0-\beta_1 w_n}^{\epsilon_{1n}=\infty} \epsilon_{1n} dF\left(\epsilon_{1n}, \epsilon_{2n}\right) > 0$$

so the OLS and IV estimators previously defined are biased.

Limitations of the Linear Model

Cobb-Douglas production function

- Let:
  - y_n measure output by the n^{th} plant, industry or economy.
  - x_{kn} measure the k^{th} input to n.
  - ϵ_n measure technological progress (or regression).
- The Cobb-Douglas production function is:

$$y_n = \exp(\beta_0 + \epsilon_n) \prod_{k=1}^{K} x_{kn}^{\beta_k}$$

  and has the log linear form:

$$\log y_n = \beta_0 + \sum_{k=1}^{K} \beta_k \log x_{kn} + \epsilon_n$$

- Unbiased estimates of $(\beta_0, \ldots, \beta_K)$ can be obtained with data on $\{(y_n, x_{1n}, \ldots, x_{Kn})\}_{n=1}^{N}$ if $E[\epsilon_n] = 0$ and $E\left[\epsilon_n \middle| x^{(N)}\right] = 0$ for all $k \in \{1, \ldots, K\}$.

Limitations of the Linear Model

Nonlinear parametric production technologies

- Supposing the mapping takes the nonlinear CES like form:

$$y_n = \exp(\beta_0 + \epsilon_n) \left[\sum_{k=1}^{K} \beta_k (x_{kn})^{\rho_k} \right]^{\rho_0}$$

and $E[\epsilon_n] = E\left[\epsilon_n \middle| x^{(N)}\right] = 0$, take logs to obtain:

$$\epsilon_n = \ln y_n - \beta_0 - \rho_0 \ln \left[\sum_{k=1}^{K} \beta_k (x_{kn})^{\rho_k} \right] \tag{1}$$

- Now multiply (1) by the instrument vector:

$$x_n \equiv \left(1, x_{1n}, x_{1n}^2, \ldots, x_{Kn}, x_{Kn}^2, x_{1n}^3\right)'$$

to obtain $2(K+1)$ nonlinear equations in the $2(K+1)$ parameters $(\rho_0, \beta_0, \ldots, \rho_K, \beta_K)$:

$$\begin{aligned} \frac{1}{N} \sum_{n=1}^{N} x_n \ln y_n &= \beta_0^{(N)} \frac{1}{N} \sum_{n=1}^{N} x_n \\ &\quad + \rho_0^{(N)} \frac{1}{N} \sum_{n=1}^{N} x_n \ln \left[\sum_{k=1}^{K} \beta_k^{(N)} (x_{kn})^{\rho_k^{(N)}} \right] \end{aligned}$$

Limitations of the Linear Model

Nonlinear nonparametric production technologies

- Nonlinearity complicates the problem of deriving the finite sample properties of the estimator for $(\rho_0, \beta_0, \ldots, \rho_k, \beta_K)$.
- Now suppose we do not impose any parametric assumptions:

$$y_n \equiv E\left[y_n \left| x_{1n}, \ldots, x_{Kn}\right.\right] + \epsilon_n \equiv \phi\left(x_{1n}, \ldots, x_{Kn}\right) + \epsilon_n$$

- By definition $E\left[\epsilon_n\right] = 0$, and $E\left[\epsilon_n \left| x_{kn}\right.\right] = 0$ for all $k \in \{1, \ldots, K\}$.
- If $\phi\left(x_{1n}, \ldots, x_{Kn}\right)$ is analytic it has the representation:

$$\phi\left(x_{1n}, \ldots, x_{Kn}\right) = \sum_{l_1=0}^{\infty} \cdots \sum_{l_K=0}^{\infty} \beta_{l_1 \ldots l_K} \left(x_{1n}^{l_1} x_{2n}^{l_2} \ldots x_{Kn}^{l_K}\right)$$

- In principle we could form an instrument vector x_n from $x_{1n}^{l_1} \ldots x_{Kn}^{l_K}$ for any K dimensional integer $(l_1, \ldots, l_K)$.
- But since there are an infinite number of coefficients to estimate, how would we compute, let alone invert $x_n x_n'$?
- Stepping back, how can we estimate an infinite number of parameters with only a finite number of observations?

Methods of Moments Estimation

Linear IV as a sample analogue to a function of population moments

- These limitations motivate the development of estimators for parametric and nonparametric nonlinear systems.
- For a sample $\{y_n, x_n\}_{n=1}^N$ recall the linear model is:

$$\epsilon_n = y_n - x_n'\beta$$

where the k dimensional vector β is unknown.

- Suppose there is an $l \times 1$ instrument vector w_n with $l \geq k$, and $E[w_n\epsilon_n] = 0$, so for any $k \times l$ matrix A:

$$0 = E[Aw_n\epsilon_n] = E\left[Aw_n\left(y_n - x_n'\beta_0\right)\right] \Rightarrow E[Aw_n y_n] = E\left[Aw_n x_n'\right]\beta_0$$

- If in addition $E[Aw_n x_n']$ is invertible, it now follows that:

$$\beta_0 = E\left[Aw_n x_n'\right]^{-1} E[Aw_n y_n]$$

- The method of moments approach approximates $E[Aw_n x_n']$ with $N^{-1}\sum_{n=1}^N Aw_n x_n'$, and $E[Aw_n y_n]$ with $N^{-1}\sum_{n=1}^N Aw_n y_n$.

Methods of Moments Estimation

Generalizing the approach

- This approach extends to nonlinear settings.
- Suppose $n \in \{1, 2, \ldots\}$. x_n is an $h \times 1$ vector of observed variables, and β_0 is a $k \times 1$ unknown parameter of interest.
- Also let $g(x_n, \beta)$ denote an $l \times 1$ vector function $g : \mathbb{R}^h \times \mathbb{R}^k \to \mathbb{R}^l$ with expectation function $E[g(x_n, \beta)] : \mathbb{R}^k \to \mathbb{R}^l$.
- To facilitate identification, assume $E[g(x_n, \beta_0)] = 0$ and that $E[g(x_n, \beta)] \neq 0$ for all $\beta \neq \beta_0$.
- Define a methods of moments estimator by setting:

$$A\left[\frac{1}{N}\sum_{n=1}^{N} g\left(x_n, \beta_{MM}^{(N)}\right)\right] = 0 \tag{2}$$

where A is a $k \times l$ weight matrix that weights the importance of each component in the vector function $g(x_n, \beta_0)$.

- Note $\beta_{MM}^{(N)}$ solves k equations in k unknowns, and might be close to the unique root to $AE[g(x_n, \beta)]$, that is β_0.

Maximum Likelihood

The score function

- Maximum Likelihood (ML) estimators are often based on the FOC, which in effect are the defining equations.
- Suppose x_n is an independent and identically distributed (iid) random variable defined on a continuous support with probability distribution function density $f(x_n, \beta)$.
- In this case the ML estimator is defined as:

$$\beta_{ML}^{(N)} = \underset{\beta}{\operatorname{argmax}} \left[\prod_{n=1}^{N} f(x_n, \beta) \right] = \underset{\beta}{\operatorname{argmax}} \left[\sum_{n=1}^{N} \ln f(x_n, \beta) \right]$$

- Assuming $f(x_n, \beta)$ is differentiable in β the FOC is:

$$0 = \sum_{n=1}^{N} \frac{\partial f\left(x_n, \beta_{ML}^{(N)}\right) / \partial \beta}{f\left(x_n, \beta_{ML}^{(N)}\right)} \equiv \frac{1}{N} \sum_{n=1}^{N} g\left(x_n, \beta_{ML}^{(N)}\right) \tag{3}$$

- If the likelihood is globally concave, then its maximum is uniquely defined by the FOC.

Maximum Likelihood

The population analogue to the score

- The average derivative of the log likelihood (3) is called the score.
- It is the sample analogue to an expectation in the population, which we typically assume has a unique root at β_0.
- To see this point, note that:

$$\begin{aligned} 1 &= \int f(x_n, \beta_0)\, \mathrm{d}x \\ \Rightarrow 0 &= \int \frac{\partial f}{\partial \beta}(x_n, \beta_0)\, \mathrm{d}x \\ &= \int \frac{\partial f(x_n, \beta_0)/\partial \beta}{f(x_n, \beta_0)} f(x_n, \beta_0)\, \mathrm{d}x \\ &= E\left[\frac{\partial f(x_n, \beta_0)/\partial \beta}{f(x_n, \beta_0)}\right] \end{aligned}$$

Continuous Time Models with Point Processes

Failure times and the hazard rate

- Let $T \in \mathbb{R}$ denote the *failure time* of a random event.
- Denote by $f(t)$ its pdf, $F(t)$ its cdf
- We call $1 - F(t)$ the *survivor function*.
- Define the *hazard rate* for T at t as:

$$h(t) \equiv \lim_{\Delta t \downarrow 0} \frac{\Pr\{T \in [t, t + \Delta t] | T \geq t\}}{\Delta t} = \frac{f(t)}{1 - F(t)}$$

- Integrating both sides:

$$-\int_0^t h(s)\, ds = \ln[1 - F(t)] \Rightarrow F(t) = 1 - \exp\left(-\int_0^t h(s)\, ds\right)$$

- Let $H(t) \equiv \int_0^t h(s)\, ds$ denote the *integrated hazard*:
  - Then $H(t) = -\log[1 - F(t)]$
  - If $F(\infty) = 1$, then failure is certain, and $H(\infty) = \infty$.

Continuous Time Models with Point Processes

Estimating a hazard rate

- Since there is a one-to-one mapping between $h(t)$ and $f(t)$, given by:

$$f(t) = h(t) \exp\left(- \int_0^t h(s)\, ds \right) \equiv h(t) \exp\left[-H(t) \right]$$

  we could specify the model in terms of $h(t)$ rather than $f(t)$.
- The log-likelihood for $\{t_n\}_{n=1}^N$, a sample of N iid failure times, is:

$$\sum_{n=1}^N \ln f(t_n) = \sum_{n=1}^N \left[\ln h(t_n) - H(t_n) \right]$$

- Some common examples of hazard rate models include:
  - constant: $h(t) = a$ (where $a > 0$) $\Rightarrow f(t) = a \exp(-at)$
  - linear: $h(t) = a + bt$ (where $a > 0$ and $b > 0$)
    $\Rightarrow f(t) = (a + bt) \exp(-at - bt/2)$
  - power: $h(t) = ct^{c-1}$ (where $c > 0$) $\Rightarrow f(t) = ct^{c-1} \exp(t^c)$
  - exponential: $h(t) = e^t \Rightarrow f(t) = e^t \exp\{-e^t + 1\}$
- The ML estimator for almost all these models is nonlinear.

Continuous Time Models with Point Processes

Models of competing risk

- Extending the model, suppose the system fails in J possible ways:
  - T_j denotes the *latent failure time* for a type $j \in \{1, \ldots, J\}$ failure.
  - $T \equiv \min\{T_1, \ldots, T_J\}$ denotes the failure time of the system.
  - The marginal hazard for the latent event T_j by $h_j(t)$.
- There is zero probability of two or more failures occurring simultaneously so:

$$h(t) = \sum_{j=1}^{J} h_j(t) \Rightarrow f(t) = h(t) e^{-H(t)} = \sum_{j=1}^{J} h_j(t) e^{-H(t)}$$

- Let $d_{jn} \in \{0, 1\}$ where $d_{jn} = 1$ indicates the n^{th} failure time, at t_n, is attributable to a type j failure.
- Note $\sum_{j=1}^{J} d_{jn} = 1$ and write $d_n \equiv (d_{1n}, \ldots, d_{Jn})$.
- The log likelihood for an iid sample of $\{t_n, d_n\}_{n=1}^{N}$ is then:

$$\sum_{n=1}^{N} \ln \sum_{j=1}^{J} d_{jn} h_j(t_n) e^{-H(t_n)} = \sum_{n=1}^{N} \sum_{j=1}^{J} d_{jn} \left[\ln h_j(t) - H(t_n)\right]$$

- Finally suppose when the system fails:
  - event $j \in \{0, \ldots, J\}$ occurs with probability $\theta_j(x)$
  - the hazard to the next failure time is given by $h(t; x)$
- For example in a economic scenario, imagine:
  - at sporadic intervals $\{t_1, t_2, \ldots\}$ agent(s) in the model moves (move)
  - j is the outcome of an individual solving an optimization problem
  - or j is a simultaneous move of several agents of a noncooperative game
  - $\theta_j(x)$ is an probability determined in equilibrium by other parameters characterizing preferences and technology.
- The data
  - comprise N observations of the form $\{t_n, x_n, j_n\}_{n=1}^{N}$,
  - t_n measures the time elapse since $\sum_{m=1}^{n-1} t_m$
  - x_n gives the state of the system
  - are truncated, never reporting $j = 0$ type events.

Continuous Time Models with Point Processes

Estimating models of competing risk with discrete choice

- Drawing upon the results in the previous slides:

$$\begin{aligned} & \sum_{n=1}^{N} \ln f\left(t_n; x_n\right) \\ = & \sum_{n=1}^{N} \ln \left[\sum_{j=1}^{J} d_{jn} \theta_j\left(x_n\right) h\left(\tau_n; x_n\right) e^{\left(-\int_0^{\tau_n} \sum_{j=1}^{J} \theta_j(x_n) h(s; x_n) ds\right)} \right] \\ = & \sum_{n=1}^{N} \left\{ \begin{array}{r} \sum_{j=1}^{J} d_{jn} \left[\ln \theta_j\left(x_n\right) + \ln h\left(t_n; x_n\right)\right] \\ -\left[1-\theta_0\left(x_n\right)\right] \int_0^{t_n} h\left(s; x_n\right) ds \end{array} \right\} \end{aligned}$$

- In the economic scenario, we might sequentially estimate:
  1. the parameters that determine $\theta_1\left(x\right), \ldots, \theta_J\left(x\right)$.
  2. the remaining parameters determining the hazard $h\left(t; x\right)$.
- For example see the market microstructure application of Hollifield, Miller, Sandas and Slive (Journal of Finance, 2006).

Nonlinear Least Squares

Extending OLS to NLS

- Consider the following nonlinear generalization to the linear system:

$$\epsilon_n = y_n - h(x_n, \beta_0)$$

- where:
  - y_n is a 1×1 observed dependent variable.
  - x_n is an $I \times 1$ vector of observed explanatory variables.
  - ϵ_n is a 1×1 unobserved variable.
  - β_0 is the $k \times 1$ parameter to be estimated.
  - $h : \mathbb{R}^I \otimes \mathbb{R}^k \longrightarrow \mathbb{R}$ is a known mapping.
- Parallel to OLS, the nonlinear least squares (NLS) estimator is:

$$\begin{aligned}
\beta_{NLS}^{(N)} &\equiv \underset{\beta}{\operatorname{argmin}} \sum_{n=1}^{N} \left[y_n - h(x_n, \beta) \right]^2 \\
&= \underset{\beta}{\operatorname{argmin}} \sum_{n=1}^{N} \left[-2 y_n h(x_n, \beta) + h(x_n, \beta)^2 \right]
\end{aligned}$$

- After dividing by $2N$ the FOC for NLS becomes:

$$0 = \frac{1}{N}\sum_{n=1}^{N}\left[-y_n\frac{\partial h\left(x_n,\beta_{NLS}^{(N)}\right)}{\partial\beta} + h\left(x_n,\beta_{NLS}^{(N)}\right)\frac{\partial h\left(x_n,\beta_{NLS}^{(N)}\right)}{\partial\beta}\right]$$

- Reverse engineering the approach of replacing sample moments with population moments, it would be nice if β_0 uniquely solved:

$$0 = E\left[\left[y_n - h\left(x_n,\beta\right)\right]\frac{\partial h\left(x_n,\beta\right)}{\partial\beta}\right]$$

in β.

Nonlinear Least Squares

Equivalence of NLS and ML with normal disturbances

- When $\epsilon \sim N\left(0, \sigma^2\right)$ iid the ML estimator is defined as:

$$\begin{aligned}
\beta_{ML}^{(N)} &\equiv \underset{\beta}{\operatorname{argmax}} \left\{ \prod_{n=1}^{N} \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{1}{2\sigma^2} \left(y_n - h\left(x_n, \beta\right) \right)^2 \right] \right\} \\
&= \underset{\beta}{\operatorname{argmax}} \left\{ \sum_{n=1}^{N} -\frac{1}{2\sigma^2} \left(y_n - h\left(x_n, \beta\right) \right)^2 \right\} \\
&= \underset{\beta}{\operatorname{argmin}} \left\{ \sum_{n=1}^{N} \left(y_n - h\left(x_n, \beta\right) \right)^2 \right\} \\
&= \beta_{NLS}^{(N)}
\end{aligned}$$

- In words, the ML and NLS estimators are identical when the unobserved variable is iid normal.

Generalized Methods of Moments Estimators

Definition

- Extending the methods of moments approach a little further suppose:
  - x_n is a $p \times 1$ vector of observed variables for $n \in \{1, 2, \ldots\}$
  - $\theta_0 \in \Theta$ is a $s \times 1$ unknown parameter of interest.
  - $G_1^{(N)}(x) \overset{\text{in some way}}{\longrightarrow} G_1(x) : \mathbb{R}^p \to \mathbb{R}^{q_1}$ is a vector of instruments formed from the observed variables.
  - $G_2(x, \theta) : \mathbb{R}^p \times \Theta \to \mathbb{R}^{q_2}$ is induced by the theoretical model with an assumed functional form.
  - $f^{(N)}(x, \theta) \equiv G_1^N(x) \otimes G_2(x, \theta) : \mathbb{R}^p \otimes \Theta \to \mathbb{R}^q$ where $q \equiv q_1 q_2$.
  - the $q \times q$ matrix. $A_N \overset{\text{in some way}}{\longrightarrow} A_0$ positive definite.
- Then a GMM estimator is defined:

$$\theta_{GMM}^{(N)} \equiv \arg\min_{\theta \in \Theta} \left\{ \left[\frac{1}{N} \sum_{n=1}^{N} f^{(N)}(x_n, \theta) \right]' A_N \left[\frac{1}{N} \sum_{n=1}^{N} f^{(N)}(x_n, \theta) \right] \right\} \tag{4}$$

- Suppose θ_0 uniquely solves $E\left[f\left(x_n, \theta\right)\right] = 0$.
- Again focusing on the FOC suppose θ^N uniquely solves:

$$\begin{aligned} 0 &= \frac{\partial}{\partial \theta}\left\{\left[\frac{1}{N}\sum_{n=1}^{N} f^{(N)}\left(x_n, \theta\right)\right]' A_N \left[\frac{1}{N}\sum_{n=1}^{N} f^{(N)}\left(x_n, \theta\right)\right]\right\} \\ &= \left[\frac{1}{N}\sum_{n=1}^{N} \frac{\partial}{\partial \theta} f^{(N)}\left(x_n, \theta\right)\right]' A_N \left[\frac{1}{N}\sum_{n=1}^{N} f^{(N)}\left(x_n, \theta\right)\right] \end{aligned}$$

- Note $f^{(N)}\left(x, \theta\right) \overset{\text{in some way}}{\longrightarrow} f\left(x, \theta\right) \equiv G_1\left(x\right) \otimes G_2\left(x, \theta\right)$.
- For any $\theta^{(N)} \overset{\text{in some way}}{\longrightarrow} \theta_0$ define A_N^* such that:

$$A_N^* \equiv \left[\frac{1}{N}\sum_{n=1}^{N} \frac{\partial}{\partial \theta} f^{(N)}\left(x_n, \theta^{(N)}\right)\right]' A_N \overset{\text{in some way}}{\longrightarrow} E\left[\frac{\partial}{\partial \theta} f\left(x_n, \theta_0\right)\right]' A_0$$

- We could now define a GMM estimator as a root to:

$$A_N^* \left[\frac{1}{N}\sum_{n=1}^{N} f^{(N)}\left(x_n, \theta\right)\right]$$

Sequential Estimation

Framework

- Consider the following specialization of the framework where some parameters do not appear in all the orthogonality equations:
  - $\theta_0 \equiv (\theta_{01}, \theta_{02})$ where $\theta_{0i} \in \Theta_i \subseteq \mathbb{R}^{s_i}$ for $i \in \{1, 2\}$.
  - $f_1(x, \theta_1) : \mathbb{R}^p \times \Theta_1 \longrightarrow \mathbb{R}^{q_1}$ with $q_1 \geq s_1$.
  - $f_2(x, \theta_1, \theta_2) : \mathbb{R}^p \times \Theta_1 \times \Theta_2 \longrightarrow \mathbb{R}^{q_2}$ with $q_2 \geq s_2$.
  - $A_1^{(N)} \overset{\text{in some way}}{\longrightarrow} A_1$ and $A_2^{(N)} \overset{\text{in some way}}{\longrightarrow} A_2$.
- To sequentially estimate θ_0, first choose $\theta_1^{(N)} \in \Theta_1$ to minimize:

$$\left[\frac{1}{N}\sum_{n=1}^{N} f_1^{(N)}(x_n, \theta_1)\right]' A_1^{(N)} \left[\frac{1}{N}\sum_{n=1}^{N} f_1^{(N)}(x_n, \theta_1)\right]$$

and then choose $\theta_2^{(N)} \in \Theta_2$ to minimize:

$$\left[\frac{1}{N}\sum_{n=1}^{N} f_2^{(N)}\left(x_n, \theta_1^{(N)}, \theta_2\right)\right]' A_2^{(N)} \left[\frac{1}{N}\sum_{n=1}^{N} f_2^{(N)}\left(x_n, \theta_1^{(N)}, \theta_2\right)\right]$$

Indirect Inference

The basic principle

- As before the primitives of the model are parameterized by $\theta \in \Theta$.
- Suppose the equilibrium or the behavioral equations of the model can be express as $\Psi\left(y_n, x_n, \epsilon_n, \theta\right) = 0$
  - where $\left(y_n, x_n\right)$ are observed variables and no particular significance is attached to the labelling of y_n versus x_n.
  - ϵ_n is an unobserved variable (that prevents the econometrician from solving the model exactly for any given θ)
- The data is generated by $\theta_0 \in \Theta$.
- Run "any" set of auxiliary equations, such as OLS on the data $\left\{y_n, x_n\right\}_{n=1}^{N}$, just once to obtain auxiliary coefficients $\beta_{OLS}^{(N)}$.
- Simulate data from $\Psi\left(y_n, x_n, \epsilon_n, \theta\right)$ for any given $\theta \in \Theta$ and obtain auxiliary coefficients $\beta_{OLS}^{(S)}\left(\theta\right)$.
- Choose $\theta_{II}^{(N)}$ to minimize $\left\| \beta_{OLS}^{(S)}\left(\theta\right) - \beta_{OLS}^{(N)} \right\|$.

Indirect Inference

Method of simulated moments

- The estimator is often used when $f^{(N)}(x_n, \theta)$ is a high dimensional integral that is costly to compute.
- The rationale is that since we are averaging over sample observations, the reason for computing $f^{(N)}(x_n, \theta)$ exactly as opposed to using a random draw for $f^{(N)}(x_n, \theta)$, is less compelling.
- As before the primitives of the model are parameterized by $\theta \in \Theta$.

$$\theta_{MSM}^{(N)} \equiv \arg\min_{\theta \in \Theta} \left\{ \left[\frac{1}{N} \sum_{n=1}^{N} \widehat{f}^{(N)}(x_n, \theta) \right]' A_N \left[\frac{1}{N} \sum_{n=1}^{N} \widehat{f}^{(N)}(x_n, \theta) \right] \right\}$$

where $\widehat{f}^{(N)}(x_n, \theta)$ is simulated, rather than directly computed.

- MSM estimators can be interpreted as examples of II estimators.
- Because MSM estimators are based on the orthogonality conditions of the underlying model their large sample properties are more transparent.

Indirect Inference

An example of the method of simulated moments

- Suppose:

$$y_n = \begin{cases} 1 \text{ if } x_n'\beta_0 + \epsilon_n > 0 \\ 0 \text{ if } x_n'\beta_0 + \epsilon_n \leq 0 \end{cases}$$

where ϵ_n is drawn from a known distribution and (y_n, x_n) are observed for $n \in \{1, \ldots, N\}$.

- Randomly draw $\{\epsilon_n^s\}_{n=1}^N$ and define:

$$\begin{aligned} y(x_n, \epsilon_n^s, \beta) &\equiv \begin{cases} 1 \text{ if } x_n'\beta + \epsilon_n^s > 0 \\ 0 \text{ if } x_n'\beta + \epsilon_n^s \leq 0 \end{cases} \\ f(y_n, x_n, \epsilon_n^s, \beta) &\equiv x_n \left[y_n - y(x_n, \epsilon_n^s, \beta) \right] \end{aligned}$$

- Then $\beta_{MSM}^{(N)}$, a methods of simulated moments estimator for β_0, is obtained by choosing β to minimize:

$$\left[\frac{1}{N} \sum_{n=1}^N f(y_n, x_n, \epsilon_n^s, \beta)' \right] A_N \left[\frac{1}{N} \sum_{n=1}^N f(y_n, x_n, \epsilon_n^s, \beta) \right]$$

Minimum Distance Estimation

An example

- Suppose
$$y_n = \beta_0 + \beta_1 x_{1n} + \beta_2 x_{2n} + \beta_1 \beta_2 x_{3n} + \varepsilon_n$$
- On approach would be to run NLS, choosing $(\beta_0, \beta_1, \beta_2)$ to minimize:
$$\frac{1}{N} \sum_{n=1}^{N} \left(y_n - \beta_0 - \beta_1 x_{1n} - \beta_2 x_{2n} - \beta_1 \beta_2 x_{3n} \right)^2$$
- The Minimum Distance (MD) estimator is a cheaper alternative:
  1. Choose $(\pi_0, \pi_1, \pi_2, \pi_3)$ to minimize the reduced form:
  $$\frac{1}{N} \sum_{n=1}^{N} \left(y_n - \pi_0 - \pi_1 x_{1n} - \pi_2 x_{2n} - \pi_3 x_{3n} \right)^2$$
  and obtain the OLS estimates $\left(\pi_0^{(N)}, \pi_1^{(N)}, \pi_2^{(N)}, \pi_3^{(N)} \right)$
  2. Choose $(\beta_0, \beta_1, \beta_2)$ to minimize:
  $$\gamma_3 \left(\beta_1 \beta_2 - \pi_3^{(N)} \right)^2 + \sum_{k=0}^{2} \gamma_k \left(\beta_k - \pi_k^{(N)} \right)^2$$

Minimum Distance Estimation

Definition

- Define the reduced form as a mapping from

$$\psi(\theta) : \Theta \to \Psi$$

- In the example above $\psi : (\beta_0, \beta_1, \beta_2) = (\beta_0, \beta_1, \beta_2, \beta_1\beta_2)'$.
- First, a GMM estimator for $\psi_0 \equiv \psi_0(\theta_0)$ is found:

$$\psi^{(N)} = \underset{\psi \in \Psi}{\operatorname{argmin}} \left\{ \left(\frac{1}{N} \sum_{n=1}^{N} f^{(N)}(x_n, \psi) \right)' A_N \left(\frac{1}{N} \sum_{n=1}^{N} f^{(N)}(x_n, \psi) \right) \right\}$$

- Second we define

$$\theta_{MD}^{(N)} = \underset{\theta \in \Theta}{\operatorname{argmin}} \left(\left(\psi^{(N)} - \psi(\theta) \right)' B_N \left(\psi^{(N)} - \psi(\theta) \right) \right)$$

for some $B_N \overset{\text{in some way}}{\longrightarrow} B$, an $r \times r$ positive definite matrix.