Extremum Estimators

Renato Molina[*]

May 23, 2019

Contents

[*] Most of these notes follow the volume IV of the “Handbook of Econometrics” by W. Newey & D. McFadden, and the second edition of “Econometric Analysis of Cross Section and Panel Data” by J. Woolridge. All errors are my own.

1 Motivation

When we deal with non linear models, it is not longer valid to use the Ordinary Least Squares (OLS) approach. For instance, consider the Poisson regression model:

$$y_i = e^{\boldsymbol{x}_i'\boldsymbol{\theta}_0} + \epsilon_i \tag{1.1}$$

The problem with this model is that we no longer have a closed form solution for nonlinear estimators, so we have to estimate $\boldsymbol{\theta}_0$ using numerical optimization along with several properties we know for extremum estimators. Below I cover several methods to deal with this problem and study their behavior at the limit.

2 Non Linear Least Squares

2.1 Specification

Suppose we are interested in specifying and estimating a model for predicting the conditional mean $\mathbb{E}[y_i|\boldsymbol{x}_i]$. We let $m(\boldsymbol{x}_i, \boldsymbol{\theta}_0)$ denote a finite parametric model for $\mathbb{E}[y_i|\boldsymbol{x}_i]$, which can or cannot be a linear function of $\boldsymbol{\theta}_0$. The functional form of $m(\boldsymbol{x}_i, \boldsymbol{\theta}_0)$ can come from economic theory, assumptions on the conditional distribution, $f(y_i|\boldsymbol{x}_i)$, or just from intuition so as to approximate the unknown functional form of $\mathbb{E}[y_i|\boldsymbol{x}_i]$.

Accordingly, we can specify some common estimators as Non Linear Squares (NLLS):

- Poisson regression model: $m(\boldsymbol{x}_i, \boldsymbol{\theta}_0) = e^{\boldsymbol{x}_i'\boldsymbol{\theta}_0}$
- Probit model: $m(\boldsymbol{x}_i, \boldsymbol{\theta}_0) = \Phi(\boldsymbol{x}_i, \boldsymbol{\theta}_0)$[1]
- Logistic model: $m(\boldsymbol{x}_i, \boldsymbol{\theta}_0) = \frac{e^{\boldsymbol{x}_i'\boldsymbol{\theta}_0}}{1+e^{\boldsymbol{x}_i'\boldsymbol{\theta}_0}}$
- Censored y model: $m(\boldsymbol{x}_i, \boldsymbol{\theta}_0) = \Phi(\boldsymbol{x}_i, \boldsymbol{\theta}_0) \times (\boldsymbol{x}_i'\boldsymbol{\theta}_0) + \Phi(\boldsymbol{x}_i, \boldsymbol{\theta}_0)$

[1] $\Phi() =$ cdf of $N(0,1)$

2.2 Estimation

Following our underlying assumption of $\mathbb{E}[y_i|\boldsymbol{x}_i] = m(\boldsymbol{x}_i, \boldsymbol{\theta}_0)$, we can proceed to estimate $\boldsymbol{\theta}_0$ by minimizing the mean squared error (MSE):

$$\boldsymbol{\theta}_0 \equiv \min_{\boldsymbol{\theta}_0 \in \Theta} \mathbb{E}[y_i - m(\boldsymbol{x}_i, \boldsymbol{\theta}_0)]^2 \tag{2.1}$$

This expression should be familiar, as the OLS is based on the same logic, where $m(\boldsymbol{x}_i, \boldsymbol{\theta}_0)$ is linear on $\boldsymbol{\theta}_0$. Now, having a clear grasp on the assumptions about $\mathbb{E}[y_i|\boldsymbol{x}_i] = m(\boldsymbol{x}_i, \boldsymbol{\theta}_0)$, we can talk about extremum estimators which are analogous to this model.

3 Extremum Estimators

Extremum estimators is a class of estimators that maximize an objective function that depends on the data and an unknown parameter over a parameter set. Formally:

$$\hat{\theta} = \operatorname*{argmax}_{\boldsymbol{\theta} \in \Theta} Q_n(\boldsymbol{\theta}), \qquad Q_n(\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^{n} a(\boldsymbol{w}_i, \boldsymbol{\theta}) \tag{3.1}$$

Where $\boldsymbol{w}_i = (y_i, \boldsymbol{x}_i)$ is the vector of data for observation i.

If we recall some common estimators, all of them can be formulated as extremum estimators:

- <u>OLS</u>: $Q_n(\boldsymbol{\theta}) = -\frac{1}{n}\sum\limits_{i=1}^{n}(y_i - \boldsymbol{x}_i'\boldsymbol{\theta})^2$, which implies $a(\boldsymbol{w}_i, \boldsymbol{\theta}) = -(y_i - \boldsymbol{x}_i'\boldsymbol{\theta})^2$
- <u>NLLS</u>: $Q_n(\boldsymbol{\theta}) = -\frac{1}{n}\sum\limits_{i=1}^{n}(y_i - m(\boldsymbol{x}_i, \boldsymbol{\theta}))^2$, where $m(\boldsymbol{x}_i, \boldsymbol{\theta}_0) \equiv \mathbb{E}[y_i|\boldsymbol{x}_i]$
- <u>MLE</u>: $Q_n(\boldsymbol{\theta}) = -\frac{1}{n}\sum\limits_{i=1}^{n} \ln[f(y_i|\boldsymbol{x}_i; \boldsymbol{\theta})]$, where $f(y_i|\boldsymbol{x}_i; \boldsymbol{\theta})$ is the density function $y_i|\boldsymbol{x}_i$
- <u>GMM</u>: $Q_n(\boldsymbol{\theta}) = -\left(\frac{1}{n}\sum\limits_{i=1}^{n} g(\boldsymbol{w}_i, \boldsymbol{\theta})\right)' W_n \left(\frac{1}{n}\sum\limits_{i=1}^{n} g(\boldsymbol{w}_i, \boldsymbol{\theta})\right)$

3.1 Consistency

To evaluate the consistency of extremum estimators we need to define:

$$Q_0(\boldsymbol{\theta}) \equiv \operatorname{plim} Qn(\boldsymbol{\theta}) \tag{3.2}$$

and

$$\begin{aligned} \boldsymbol{\theta}_0 &\equiv \operatorname*{argmax}_{\boldsymbol{\theta} \in \Theta} Q_0(\boldsymbol{\theta}) && (3.3) \\ &\equiv \operatorname*{argmax}_{\boldsymbol{\theta} \in \Theta}[\operatorname{plim} Qn(\boldsymbol{\theta})] && (3.4) \end{aligned}$$

Recall from (3.1):

$$\begin{aligned} \hat{\boldsymbol{\theta}} &= \operatorname*{argmax}_{\boldsymbol{\theta} \in \Theta} Q_n(\boldsymbol{\theta}) \\ \implies \operatorname{plim}(\hat{\boldsymbol{\theta}}) &= \operatorname{plim}\left(\operatorname*{argmax}_{\boldsymbol{\theta} \in \Theta} Q_n(\boldsymbol{\theta})\right) && (3.5) \end{aligned}$$

So, to ensure consistency, we need to show that $\text{plim}(\hat{\boldsymbol{\theta}}) = \boldsymbol{\theta}_0$, formally that:

$$\text{plim}(\hat{\boldsymbol{\theta}}) \equiv \text{plim}\left(\underset{\boldsymbol{\theta}\in\Theta}{\text{argmax}}\, Q_n(\boldsymbol{\theta})\right) = \underset{\boldsymbol{\theta}\in\Theta}{\text{argmax}}\left[\text{plim}\left(Q_n(\boldsymbol{\theta})\right)\right] \equiv \boldsymbol{\theta}_0 \tag{3.6}$$

It turns out, that if our objective function satisfies some regularity conditions we can ensure the consistency of our estimator. However, we need one definition before doing that:

Definition 3.1. Pointwise Convergence in Probability. In a repeated sampling experiment, $Q_n(\boldsymbol{\theta})$ is a sequence of random functions. We say that $Q_n(\boldsymbol{\theta})$ converges pointless in probability to a non-random function $Q_0(\boldsymbol{\theta})$ if:

$$\text{plim}\left(Q_n(\boldsymbol{\theta})\right) = Q_0(\boldsymbol{\theta}),\ \forall\ \boldsymbol{\theta} \in \Theta \tag{3.7}$$

Although it might seem confusing, pointwise convergence is the same as convergence in probability, but it applies to random functions. We can use a weak Law of Large Numbers (LLN) to show pointwise convergence in probability. Sometimes, however, even if $Q_n(\boldsymbol{\theta})$ is continuous in $\boldsymbol{\theta}$, the limit function $Q_0(\boldsymbol{\theta})$ may not be. In this case we use the uniform LLN (Appendix A.1).

Now that we have all the tools we need to establish the conditions for consistency. We refer to Newey and McFadden (1994), or N&M, to deal with the proper formalities of establish consistency. For a general objective function we know that if the conditions for Theorem 2.1 are satisfied (Appendix A.2), the estimator is consistent. Moreover, if we have a concave objective function, if the conditions for Theorem 2.7 (Appendix A.3) are satisfied, the estimator is consistent.

3.2 Asymptotic Normality of Extremum Estimators

In the case of OLS, we used its asymptotic properties to establish its behavior at the limit. We can do the same with extremum estimators, where given certain conditions (Appendix A.8), we can show their asymptotic normality. First, we need to note that:

$$\begin{aligned}\hat{\boldsymbol{\theta}} &= \underset{\boldsymbol{\theta}\in\Theta}{\text{argmax}}\, Q_n(\boldsymbol{\theta}) \\ \implies \nabla_\theta Q_n(\hat{\boldsymbol{\theta}}) &= 0 \end{aligned} \tag{3.8}$$

Provided that it is continuously differentiable, the first order conditions (FOCs) will admit a mean-value expansion around $\boldsymbol{\theta}_0$:

$$\nabla_\theta Q_n(\hat{\boldsymbol{\theta}}) = \nabla_\theta Q_n(\boldsymbol{\theta}_0) + \nabla_{\theta\theta} Q_n(\bar{\boldsymbol{\theta}})(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) = 0 \tag{3.9}$$

with $\bar{\boldsymbol{\theta}}$ as a mean value between $\hat{\boldsymbol{\theta}}$ and $\boldsymbol{\theta}_0$. We can therefore rearrange as follows:

$$\begin{aligned}\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) &= -\left(\nabla_{\theta\theta} Q_n(\bar{\boldsymbol{\theta}})\right)^{-1} \sqrt{n}\nabla_\theta Q_n(\boldsymbol{\theta}_0) \\ &= -\left(\frac{1}{n}\sum_{i=1}^n \nabla_{\theta\theta} a(\boldsymbol{w}_i, \bar{\boldsymbol{\theta}})\right)^{-1} \sqrt{n}\frac{1}{n}\sum_{i=1}^n \nabla_\theta a(\boldsymbol{w}_i, \boldsymbol{\theta}_0)\end{aligned} \tag{3.10}$$

Notice that the identification assumption for the consistency theorems implies:

$$\mathbb{E}[\nabla_\theta a(\boldsymbol{w}_i, \boldsymbol{\theta}_0)] = 0$$

So, this situation already looks like a Central Limit Theorem (CLT). If we recall the conditions necessary, provided that $J \equiv \mathbb{E}[\nabla_\theta a(\boldsymbol{w}_i, \boldsymbol{\theta}_0)\nabla_\theta a(\boldsymbol{w}_i, \boldsymbol{\theta}_0)'] < \infty$, and given that we have data that is independently and identically distributed (iid):

$$\sqrt{n}\frac{1}{n}\sum_{i=1}^{n} \nabla_\theta a(\boldsymbol{w}_i, \boldsymbol{\theta}_0) \xrightarrow{d} N(0, J) \tag{3.11}$$

For the remaining component we can use the uniform LLN:

$$\frac{1}{n}\sum_{i=1}^{n} \nabla_{\theta\theta} a(\boldsymbol{w}_i, \bar{\boldsymbol{\theta}}) \xrightarrow{p} \mathbb{E}[\nabla_{\theta\theta} a(\boldsymbol{w}_i, \boldsymbol{\theta}_0)] \equiv H \tag{3.12}$$

Note that since we defined $\bar{\boldsymbol{\theta}}$ as some value between $\hat{\boldsymbol{\theta}}$ and $\boldsymbol{\theta}_0$, and given the consistency of $\hat{\boldsymbol{\theta}}$, it must converge to $\boldsymbol{\theta}_0$.

Finally, and using the Continuous Mapping Theorem we have:

$$\left(\frac{1}{n}\sum_{i=1}^{n} \nabla_{\theta\theta} a(\boldsymbol{w}_i, \bar{\boldsymbol{\theta}})\right)^{-1} \xrightarrow{p} \mathbb{E}[\nabla_{\theta\theta} a(\boldsymbol{w}_i, \boldsymbol{\theta}_0)]^{-1} = H^{-1} \tag{3.13}$$

Provided that H is invertible, by equations (3.12) and (3.14) we use the scaling properties of normally distributed variables and get the asymptotic normality of extremum estimators:

$$\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \xrightarrow{d} N(0, H^{-1}JH^{-1}) \tag{3.14}$$

4 Large Sample Properties of the Maximum Likelihood Estimator

As we noted before, the Maximum Likelihood Estimator (MLE) can be expressed as an extremum estimator. Recall that given a probability distribution (pdf) for which we have an iid sample $\boldsymbol{w}_i$, we estimate the parameter value $\boldsymbol{\theta}_0$. Take note that this an untestable assumption and it completely drives the results.

The MLE of $\boldsymbol{\theta}_0$ is the value of $\boldsymbol{\theta}$ that maximizes the likelihood function. That is, the probability of observing the sample $\boldsymbol{w}_i$ we have drawn, given the distributional assumptions we make on the population.

4.1 Conditional and Unconditional Likelihood

Suppose that we draw the sample $\boldsymbol{w}_i = (y_i, \boldsymbol{x}_i) : i = 1, ..., n$ from a well defined population. Then the probability density function of each draw $\boldsymbol{w}_i$ is $f(\boldsymbol{w}_i; \Psi_0)$ or $f(\boldsymbol{y}_i, \boldsymbol{x}_i; \Psi_0)$. The parameter Ψ_0 fully characterizes the pdf, so with knowledge of f and Ψ_0 we can generate the sample (i.e., normal distribution $\Psi = (\mu, \sigma^2)$). Because the sample is iid, the joint density of the sample is given by the product of the marginals:

$$f(\boldsymbol{w}_1, ..., \boldsymbol{w}_n; \Psi_0) = \prod_{i=1}^{n} f(\boldsymbol{w}_i; \Psi_0) \tag{4.1}$$

The likelihood function is obtained by replacing the true parameter value, Ψ_0, by an hypothetical value, Ψ, and interpreting the joint pdf as a function of Ψ where the sample is fixed:

$$L(\Psi; \boldsymbol{w}_1, ..., \boldsymbol{w}_n) = L(\Psi) = \prod_{i=1}^{n} f(\boldsymbol{w}_i, \Psi) \tag{4.2}$$

We can proceed further and take logs to get the log-likelihood of the joint distribution:

$$\ln[L(\Psi; \boldsymbol{w}_1, ..., \boldsymbol{w}_n)] = \ln[L(\Psi)] = \sum_{i=1}^{n} \ln[f(\boldsymbol{w}_i, \Psi)] \tag{4.3}$$

Now, recall that the definition of a conditional probability is $P(A|B) = P(A,B)/P(B)$, so we can apply this property to the probability density function:

$$f(y_i, \boldsymbol{x}_i; \Psi_0) = f(y_i|\boldsymbol{x}_i; \boldsymbol{\theta}_0) \times f(\boldsymbol{x}_i; \gamma_0) \tag{4.4}$$

where $\boldsymbol{\theta}_0$ is the parameter of interest, and γ_0 is a nuisance parameter of the marginal distribution of $\boldsymbol{x}_i$.

As the name suggest, we want to estimate $\boldsymbol{\theta}_0$ such as to maximize the previous maximum likelihood function. The issue is that we just added another probability density function, $f(\boldsymbol{x}_i; \gamma_0)$, to the problem. Nonetheless, when we take logs we can get rid of the last term and obtain the following expression:

$$\ln[L(\Psi)] = \sum_{i=1}^{n} \ln[f(\boldsymbol{w}_i, \Psi)] = \sum_{i=1}^{n} \ln[f(y_i|\boldsymbol{x}_i; \boldsymbol{\theta}_0)] + \sum_{i=1}^{n} \ln[f(\boldsymbol{x}_i; \gamma_0)] \tag{4.5}$$

Note that when we take partial derivatives to maximize (4.5) as a function of $\boldsymbol{\theta}_0$, we drop the last term. Now, we can formulate the MLE as an extremum estimator as:

$$\hat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta} \in \Theta}{\operatorname{argmax}} \frac{1}{n} \sum_{i=1}^{n} \ln[f(y_i|\boldsymbol{x}_i; \boldsymbol{\theta})] \tag{4.6}$$

To see how the MLE operates for different distributions we can take a look at some examples:

1. Normal regression model: Let $y_i|\boldsymbol{x}_i \sim N(\boldsymbol{x}_i'\boldsymbol{\beta}_0, \sigma^2)$. The conditional pdf is given by:

$$f(y_i|\boldsymbol{x}_i; \boldsymbol{\theta}_0) = \frac{1}{\sqrt{2\pi}\sigma_0} e^{-\frac{1}{2}\frac{(y_i - \boldsymbol{x}_i'\boldsymbol{\beta}_0)^2}{\sigma_0^2}}, \quad \boldsymbol{\theta}_0 = (\boldsymbol{\beta}_0, \sigma^2) \tag{4.7}$$

   Then, the MLE is given by:

$$\hat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta} \in \Theta}{\operatorname{argmax}} \frac{1}{n} \sum_{i=1}^{n} \left(-\ln[\sqrt{2\pi}] - \ln[\sigma] - \frac{1}{2}\frac{(y_i - \boldsymbol{x}_i'\boldsymbol{\beta})^2}{\sigma^2} \right) \tag{4.8}$$

2. Poisson Regression Model: Let Let $y_i|\boldsymbol{x}_i \sim \mathcal{P}(\lambda_0)$ with $y_i = 0, 1, .., n$ and $\mathbb{E}[y_i|\boldsymbol{x}_i] = \lambda_0$ and $Var(y_i|\boldsymbol{x}_i) = \lambda_0$. Where is often common to let $\lambda_0 = e^{\boldsymbol{x}_i'\boldsymbol{\theta}_0}$. The conditional pdf is given by:

$$f(y_i|\boldsymbol{x}_i; \boldsymbol{\theta}_0) = \frac{e^{y_i \boldsymbol{x}_i'\boldsymbol{\theta}_0} e^{-e^{\boldsymbol{x}_i'\boldsymbol{\theta}_0}}}{y_i!} \tag{4.9}$$

   It follows that the 1951 estimator is:

$$\hat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta} \in \Theta}{\operatorname{argmax}} \frac{1}{n} \sum_{i=1}^{n} \left(-e^{\boldsymbol{x}_i'\boldsymbol{\theta}_0} + y_i \boldsymbol{x}_i'\boldsymbol{\theta}_0 - \ln[y_i!] \right) \tag{4.10}$$

4.2 Consistency and Asymptotic Normality of MLE

If we have $\boldsymbol{w}_i$, a set of iid data drawn from the conditional pdf $f(y_i|\boldsymbol{x}_i;\boldsymbol{\theta}_0)$, where $f(y_i|\boldsymbol{x}_i;\boldsymbol{\theta}_0) > 0$ and $\int f(y_i|\boldsymbol{x}_i;\boldsymbol{\theta}_0)dy = 1 \ \forall \ \boldsymbol{\theta} \in \Theta$. If, in addition, the regularity conditions from N&M are met (Appendix A.5) for our MLE $\hat{\boldsymbol{\theta}}$. Then we have that the estimator is consistent, $\hat{\boldsymbol{\theta}} \xrightarrow{p} \boldsymbol{\theta}_0$.

If we maintain the assumptions for consistency, and define some regularity conditions on our estimator (Appendix A.6), we have have everything we need to derive the asymptotic distribution of the MLE. Since data is iid, and given the zero expected score result below, we have that:

$$\sqrt{n}\nabla_\theta Q_n(\boldsymbol{\theta}_0) = \sqrt{n}\frac{1}{n}\sum_{i=1}^{n}\nabla_\theta \ln[f(y_i|\boldsymbol{x}_i;\boldsymbol{\theta}_0)]$$

$$\implies \mathbb{E}[\nabla_\theta \ln[f(y_i|\boldsymbol{x}_i;\boldsymbol{\theta}_0)]] = 0 \qquad \text{(By identification)} \tag{4.11}$$

Provided that $J \equiv \mathbb{E}[(\nabla_\theta \ln[f(y_i|\boldsymbol{x}_i;\boldsymbol{\theta}_0)])(\nabla_\theta \ln[f(y_i|\boldsymbol{x}_i;\boldsymbol{\theta}_0)])'] < \infty$, using the CLT gives the following:

$$\sqrt{n}\frac{1}{n}\sum_{i=1}^{n}\nabla_\theta \ln[f(y_i|\boldsymbol{x}_i;\boldsymbol{\theta}_0)] \xrightarrow{d} N(0, J) \tag{4.12}$$

Continuity and uniform convergence of the expected Hessian, $H(\boldsymbol{\theta}) \equiv \nabla_{\theta\theta}Q_0(\boldsymbol{\theta}) = \mathbb{E}[\ln[f(y_i|\boldsymbol{x}_i;\boldsymbol{\theta}_0)]]$, ensures that:

$$\nabla_{\theta\theta}Q_n(\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^{n}\ln[f(y_i|\boldsymbol{x}_i;\boldsymbol{\theta}_0)] \xrightarrow{u.p.} H(\boldsymbol{\theta}) \tag{4.13}$$

So, provided that the limit Hessian is invertible, $H(\boldsymbol{\theta})^{-1} < \infty$, we finally have:

$$\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \xrightarrow{d} N(0, H^{-1}JH^{-1}) = N(0, J^{-1}) \tag{4.14}$$

The result on the asymptotic variance of the MLE, $J = -H$, is known as the information matrix equality.

4.3 Specific Results to MLE

Likelihood ratio test

Suppose we are interested in testing $H_0 : r(\boldsymbol{\theta}_0) = 0$ against the two sided alternative. One approach is the Wald test based on the unconstrained estimator of $\boldsymbol{\theta}_0$:

$$\hat{W} = nr(\hat{\boldsymbol{\theta}})'[R(\hat{\boldsymbol{\theta}})\hat{J}^{-1}R(\hat{\boldsymbol{\theta}})']^{-1}r(\hat{\boldsymbol{\theta}}) \xrightarrow{d} \chi^2_{(q)} \tag{4.15}$$

On the other hand, we could also use the likelihood ratio test, which consists in comparing the value of the log-likelihood function at the constrained estimate of $\boldsymbol{\theta}_0$ constrained under H_0 and the unconstrained one:

$$\hat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}\in\Theta}{\operatorname{argmax}}\sum_{i=1}^{n}\ln[f(y_i|\boldsymbol{x}_i;\boldsymbol{\theta})] = \underset{\boldsymbol{\theta}\in\Theta}{\operatorname{argmax}}\, L_U(\boldsymbol{\theta}) \tag{4.16}$$

$$\tilde{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}\in\Theta:r(\boldsymbol{\theta})=0}{\operatorname{argmax}}\sum_{i=1}^{n}\ln[f(y_i|\boldsymbol{x}_i;\boldsymbol{\theta})] = \underset{\boldsymbol{\theta}\in\Theta:r(\boldsymbol{\theta})=0}{\operatorname{argmax}}\, L_R(\boldsymbol{\theta}) \tag{4.17}$$

Then, it follows that:

$$LR = 2(L_U(\hat{\boldsymbol{\theta}}) - L_R(\tilde{\boldsymbol{\theta}})) \xrightarrow{d} \chi^2_q \tag{4.18}$$

Zero Expected Score

Under the MLE assumptions, we can show that $J = -H$. First, consider:

$$\begin{aligned} &\int f(y_i|\boldsymbol{x}_i;\boldsymbol{\theta})dy = 1 \\ \implies \nabla_\theta &\int f(y_i|\boldsymbol{x}_i;\boldsymbol{\theta})dy = 0 && \text{(Diff w.r.t } \boldsymbol{\theta}\text{)} \end{aligned} \tag{4.19}$$

Under the regularity conditions from N&M, which make $f(y_i|\boldsymbol{x}_i;\boldsymbol{\theta})$ smooth enough, differentiable, and with derivative function bounded, we can interchange integration and differentiation. Operationally, this property implies:

$$\begin{aligned} \nabla_\theta \int f(y_i|\boldsymbol{x}_i;\boldsymbol{\theta})dy &= 0 \\ \implies \int \nabla_\theta f(y_i|\boldsymbol{x}_i;\boldsymbol{\theta})dy &= 0 \\ \iff \int \frac{f(y_i|\boldsymbol{x}_i;\boldsymbol{\theta})}{f(y_i|\boldsymbol{x}_i;\boldsymbol{\theta})} \nabla_\theta f(y_i|\boldsymbol{x}_i;\boldsymbol{\theta})dy &= 0 && \text{(Multiply by 1)} \\ \implies \int \nabla_\theta \ln[f(y_i|\boldsymbol{x}_i;\boldsymbol{\theta}_0)] f(y_i|\boldsymbol{x}_i;\boldsymbol{\theta}_0)dy &= 0 && \text{(Evaluate at } \boldsymbol{\theta}_0\text{)} \\ \iff \mathbb{E}[\nabla_\theta \ln[f(y_i|\boldsymbol{x}_i;\boldsymbol{\theta}_0)] &= 0 && \text{(Definition of } \mathbb{E}\text{)} \end{aligned} \tag{4.20}$$

So, we have showed the condition we required for the first part of the asymptotic distribution. Now we can also use the chain rule to get to the final result:

$$\begin{aligned} \nabla_\theta \int \nabla_\theta \ln[f(y_i|\boldsymbol{x}_i;\boldsymbol{\theta})] f(y_i|\boldsymbol{x}_i;\boldsymbol{\theta})dy &= 0 \\ \iff \int \nabla_\theta \left(\nabla_\theta \ln[f(y_i|\boldsymbol{x}_i;\boldsymbol{\theta})] f(y_i|\boldsymbol{x}_i;\boldsymbol{\theta})dy\right) &= 0 \\ \implies \int \nabla_\theta \left(\nabla_\theta \ln[f(y_i|\boldsymbol{x}_i;\boldsymbol{\theta})] f(y_i|\boldsymbol{x}_i;\boldsymbol{\theta})dy\right) &= \int \nabla_{\theta\theta} \ln[f(y_i|\boldsymbol{x}_i;\boldsymbol{\theta})] f(y_i|\boldsymbol{x}_i;\boldsymbol{\theta})dy \\ &\quad + \int (\nabla_\theta \ln[f(y_i|\boldsymbol{x}_i;\boldsymbol{\theta})])(\nabla_\theta f(y_i|\boldsymbol{x}_i;\boldsymbol{\theta}))' dy \end{aligned}$$

For this expression, we can use the same trick as before:

$$\begin{aligned} 0 &= \int \nabla_{\theta\theta} \ln[f(y_i|\boldsymbol{x}_i;\boldsymbol{\theta})] f(y_i|\boldsymbol{x}_i;\boldsymbol{\theta})dy + \int (\nabla_\theta \ln[f(y_i|\boldsymbol{x}_i;\boldsymbol{\theta})]) \left(\nabla_\theta \frac{f(y_i|\boldsymbol{x}_i;\boldsymbol{\theta})}{f(y_i|\boldsymbol{x}_i;\boldsymbol{\theta})} f(y_i|\boldsymbol{x}_i;\boldsymbol{\theta})\right)' dy \\ 0 &= \int \nabla_{\theta\theta} \ln[f(y_i|\boldsymbol{x}_i;\boldsymbol{\theta})] f(y_i|\boldsymbol{x}_i;\boldsymbol{\theta})dy + \int (\nabla_\theta \ln[f(y_i|\boldsymbol{x}_i;\boldsymbol{\theta})]) \left(\nabla_\theta \ln[f(y_i|\boldsymbol{x}_i;\boldsymbol{\theta})]\right)' dy \\ \implies \int \nabla_{\theta\theta} \ln[f(y_i|\boldsymbol{x}_i;\boldsymbol{\theta})] f(y_i|\boldsymbol{x}_i;\boldsymbol{\theta})dy &= -\int (\nabla_\theta \ln[f(y_i|\boldsymbol{x}_i;\boldsymbol{\theta})]) \left(\nabla_\theta \ln[f(y_i|\boldsymbol{x}_i;\boldsymbol{\theta})]\right)' f(y_i|\boldsymbol{x}_i;\boldsymbol{\theta})dy \end{aligned}$$

Now we can plug for $\boldsymbol{\theta}_0$:

$$\begin{aligned} \int \nabla_{\theta\theta} \ln[f(y_i|\boldsymbol{x}_i;\boldsymbol{\theta}_0)] f(y_i|\boldsymbol{x}_i;\boldsymbol{\theta}_0)dy &= -\int (\nabla_\theta \ln[f(y_i|\boldsymbol{x}_i;\boldsymbol{\theta_0})]) \left(\nabla_\theta \ln[f(y_i|\boldsymbol{x}_i;\boldsymbol{\theta_0})]\right)' f(y_i|\boldsymbol{x}_i;\boldsymbol{\theta}_0)dy \\ \mathbb{E}[\nabla_{\theta\theta} \ln[f(y_i|\boldsymbol{x}_i;\boldsymbol{\theta}_0)] &= -\mathbb{E}[(\nabla_\theta \ln[f(y_i|\boldsymbol{x}_i;\boldsymbol{\theta_0})]) \left(\nabla_\theta \ln[f(y_i|\boldsymbol{x}_i;\boldsymbol{\theta_0})]\right)' \\ J &= -H \end{aligned} \tag{4.21}$$

Some comments on MLE

As we just showed, MLE is efficient (smallest asymptotic variance) amongst all estimators that are consistent and asymptotically normal, $Avar = J^{-1}$, instead of $Avar = H^{-1}JH^{-1}$. Nevertheless, we have to keep in mind that MLE is not robust to mistakes on the distribution assumptions; therefore, misspecification can lead to inconsistency. Moreover, for small samples, even the correct distributional assumptions will still render a poor estimation.

5 Large Sample Properties of Generalized Method of Moments Estimators

The method of moments (MM) of the generalized method of moments (GMM) estimates population parameters by a process of matching population moments (which are a function of the parameters of interest) with sample moments. For example, let's take a sample from the uniform distribution: $y_i \sim U[\theta_0]$, where $\theta_0 = b - a$. The density function of this distribution is given by:

$$f(y) = \frac{1}{\theta_0}, \;\; 0 \leq y \leq \theta_0 \tag{5.1}$$

$$\mathbb{E}[y] = \int\limits_0^{\theta_0} y f(y) dy = \int\limits_0^{\theta_0} \frac{y}{\theta_0} dy = \frac{\theta_0}{2} \tag{5.2}$$

Now consider:

$$g_1(y_i, \theta) = y_i - \frac{\theta}{2} \tag{5.3}$$

such that:

$$\begin{aligned}
\mathbb{E}[g_1(y_i, \theta_0)] &= 0 \\
\iff \mathbb{E}\left[y_i - \frac{\theta_0}{2}\right] &= 0 \\
\iff \mathbb{E}[y_i] - \mathbb{E}\left[\frac{\theta_0}{2}\right] &= 0 && \text{(By linearity of } \mathbb{E}\text{)} \\
\implies \theta_0 &= 2\mathbb{E}[y_i] && (5.4)
\end{aligned}$$

Therefore, the MM estimator simply replaces the sample analog for the expected value:

$$\hat{\theta}^{MM,1} = 2\frac{1}{n}\sum_{i=1}^{n} y_i \tag{5.5}$$

With respect to the other moment, consider:

$$\mathbb{E}[y^2] = \int\limits_0^{\theta_0} y^2 f(y) dy = \int\limits_0^{\theta_0} \frac{y^2}{\theta_0} dy = \frac{\theta_0^2}{3} \tag{5.6}$$

so we can define another similar function:

$$g_2(y_i, \theta) = y_i^2 - \frac{\theta^2}{3} \tag{5.7}$$

which we can use to estimate our parameter as follows:

$$\begin{aligned}\mathbb{E}[y_i^2] &= \mathbb{E}\left[\frac{\theta_0^2}{3}\right] \\ \implies \theta_0 &= \sqrt{3\mathbb{E}[y_i^2]} \\ \implies \hat{\theta}^{MM,2} &= \sqrt{3\frac{1}{n}\sum_{i=1}^{n}(y_i^2)}\end{aligned} \tag{5.8}$$

This simple example allows us to say a couple things about method of moments. First, that models can be over-identified, which means that there are more moments than unknown parameters. Second, GMM allows to optimally combine different estimators in over-identified models. For example, in the previous example we could derive $+\infty$ moment conditions. To solve this multiplicity problem, one can always refer to $\mathbb{E}[\epsilon_i|x_i] = 0$, or we one could choose the K best moment conditions for a given model, which brings us to the next section.

5.1 Setup of the GMM Estimator

Let $g(\boldsymbol{w}_i, \boldsymbol{\theta})$ be a $[q, 1]$ vector ($q \geq K$) of moment conditions such that $\mathbb{E}[g(\boldsymbol{w}_i, \boldsymbol{\theta})] = 0$ only when evaluated at $\boldsymbol{\theta} = \boldsymbol{\theta}_0$. Suppose that we have a random sample of iid data $\{\boldsymbol{w}_i = (y_i, \boldsymbol{x}_i) : 1, ..., n\}$ from which we want to estimate a $[K, 1]$ vector of parameters, $\boldsymbol{\theta}_0$. Then, let:

$$g_n(\boldsymbol{w}_i, \boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^{n} g(\boldsymbol{w}_i, \boldsymbol{\theta}) \tag{5.9}$$

be the sample moments corresponding to $\mathbb{E}[g(\boldsymbol{w}_i, \boldsymbol{\theta})]$.

The GMM estimator of $\boldsymbol{\theta}_0$ is the value of $\boldsymbol{\theta}$ that sets the sample moment condition as close as possible to $\mathbf{0}$. In particular, there are two possible cases:
Case 1: Exactly-identified models: If the model is exactly identified ($q = K$), or that the parameters we want estimate match the moment conditions, then there is a unique $\hat{\boldsymbol{\theta}}$ that sets the sample moments exactly to $\mathbf{0}$:

$$g_n(\boldsymbol{w}_i, \hat{\boldsymbol{\theta}}) = \frac{1}{n}\sum_{i=1}^{n} g(\boldsymbol{w}_i, \hat{\boldsymbol{\theta}}) = \mathbf{0} \tag{5.10}$$

Case 2: Over-identified models: If the model is over identified ($q > K$), then there are no unique $\hat{\boldsymbol{\theta}}$ that sets the sample moments exactly to $\mathbf{0}$, so multiple solutions exists just as the uniform example we covered earlier. Instead, we are looking for the vector $\hat{\boldsymbol{\theta}}$ that makes the sample moments as close as possible to $\mathbf{0}$ in the following quadratic form:

$$\begin{aligned}\hat{\boldsymbol{\theta}} &= \underset{\theta\in\Theta}{\text{argmax}} - \left(\frac{1}{n}\sum_{i=1}^{n} g(\boldsymbol{w}_i, \boldsymbol{\theta})\right)' W_n \left(\frac{1}{n}\sum_{i=1}^{n} g(\boldsymbol{w}_i, \boldsymbol{\theta})\right) \\ &= \underset{\theta\in\Theta}{\text{argmax}} -g_n(\boldsymbol{w}_i, \boldsymbol{\theta})' W_n g_n(\boldsymbol{w}_i, \boldsymbol{\theta})\end{aligned} \tag{5.11}$$

with W_n as a positive semi-definite weighting matrix that could be stochastic or deterministic.

Whenever W_n is stochastic, it converges in probability to W and does not depend on $\boldsymbol{\theta}$. This probability limit is positive definite, and it assigns the weights given to each moment conditions in estimating $\boldsymbol{\theta}_0$. Whenever the model is just identified, the choice of W_n does not matter.

A natural question that arises after this definition is where these moments conditions come from, and the answer is from three possible places. Consider the following examples:

1. **Assumed orthogonality assumptions**: Recall the simple linear model $y_i = \boldsymbol{x}_i'\boldsymbol{\theta} + \epsilon_i$. If there's no endogeneity, we have:

$$\begin{aligned} &\mathbb{E}[\boldsymbol{x}_i\epsilon_i] = 0 \\ \iff &\mathbb{E}[\boldsymbol{x}_i(y_i - \boldsymbol{x}_i'\boldsymbol{\theta_0})] = 0 \\ &\implies g(\boldsymbol{w}_i, \boldsymbol{\theta}) = \boldsymbol{x}_i(y_i - \boldsymbol{x}_i'\boldsymbol{\theta_0}) \end{aligned} \tag{5.12}$$

If we were to include an instrumental variable $\boldsymbol{z}_i$:

$$\begin{aligned} &\mathbb{E}[\boldsymbol{z}_i\epsilon_i] = 0 \\ \iff &\mathbb{E}[\boldsymbol{z}_i(y_i - \boldsymbol{x}_i'\boldsymbol{\theta_0})] = 0 \\ &\implies g(\boldsymbol{w}_i, \boldsymbol{\theta}) = \boldsymbol{z}_i(y_i - \boldsymbol{x}_i'\boldsymbol{\theta_0}) \end{aligned} \tag{5.13}$$

2. **Distributional assumptions**: Recall that NLLS or MLE require assumptions about the population underlying distribution for their FOCs. For instance, the zero expected score result for the MLE says that:

$$\begin{aligned} &\mathbb{E}[\nabla_\theta \ln(y_i|\boldsymbol{x}_i;\boldsymbol{\theta_0})] = 0 \\ &\implies g(\boldsymbol{w}_i, \boldsymbol{\theta}) = \nabla_\theta \ln(y_i|\boldsymbol{x}_i;\boldsymbol{\theta_0}) \end{aligned} \tag{5.14}$$

3. **Economic theory**: Theoretical derivations also allow us to specify moment conditions. For example, in a consumption-based asset pricing model we know that the Euler Equation is:

$$\begin{aligned} &u'(c_t) = \mathbb{E}_t[(1 + r_{t+1})u'(c_{t+1})|I_t] \\ \implies &0 = \mathbb{E}_t\left[(1 + r_{t+1})\frac{u'(c_{t+1})}{u'(c_t)}|I_t\right] - 1 \end{aligned} \tag{5.15}$$

Consider the power utility function $u(c) = \frac{c^{1-\theta_0-1}}{1-\theta_0} \implies u'(c) = c^{-\theta_0}$. If we want to estimate θ_0, now we have:

$$\begin{aligned} g_t(\theta) &= (1 + r_{t+1})\frac{u'(c_{t+1})}{u'(c_t)} - 1 \\ \mathbb{E}[g_t(\theta)|I_t] &= 0 \end{aligned} \tag{5.16}$$

To implement, we condition on a vector of observed data $\boldsymbol{z}_t \in I_t$. Then, using the law of iterated expectations we get our moment condition:

$$\begin{aligned} &\mathbb{E}[\mathbb{E}[g_t(\theta)|I_t]z_t] = \mathbb{E}[g_t(\theta)z_t] \\ &\implies \mathbb{E}[g_t(\theta)z_t] = 0 \end{aligned} \tag{5.17}$$

5.2 Consistency and Asymptotic Distribution of the GMM

To show consistency for the GMM estimator, we will again rely on the regularity conditions from N&M (Appendix A.7). Suppose we have a sample of iid data $\boldsymbol{w}_i$, where $W_n \xrightarrow{p} W$, a positive definite matrix. Then let $\boldsymbol{\theta}_0$ be a $[K, 1]$ vector and be such that $q \geq K$ population moment conditions are satisfied at $\theta_0 : \mathbb{E}[g(\boldsymbol{w}_i, \boldsymbol{\theta}_0)] = 0$. Then, if the regularity conditions are met, the estimator is consistent, $\hat{\boldsymbol{\theta}} \xrightarrow{p} \boldsymbol{\theta}_0$.

Same as before, provided that the regularity conditions hold (Appendix A.8), if we let $\Omega \equiv \mathbb{E}[(g(\boldsymbol{w}_i, \boldsymbol{\theta}_0))(g(\boldsymbol{w}_i, \boldsymbol{\theta}_0))']$, and $G(\boldsymbol{\theta}) \equiv \mathbb{E}[\nabla_\theta g(\boldsymbol{w}_i, \boldsymbol{\theta}_0)]$ with $G = G(\boldsymbol{\theta}_0)$, then:

$$\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \xrightarrow{d} N(0, (G'WG)^{-1}G'W\Omega WG(G'WG)^{-1}) \tag{5.18}$$

If $W = \Omega^{-1}$, then it follows that:

$$\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \xrightarrow{d} N(0, (G'\Omega^{-1}G)^{-1}) \tag{5.19}$$

As with our previous estimators, it is worth spending some time deriving the asymptotic variance of the estimator. We begin with the definition of the moment conditions $\sqrt{n}g_n(\boldsymbol{\theta}) \xrightarrow{d} N(0, \Omega)$, and then follow the steps below:

Step 1: Take the FOCs of the GMM maximization problem:

$$\nabla_\theta Q_n(\hat{\boldsymbol{\theta}}) = -2G_n(\boldsymbol{w}_i, \hat{\boldsymbol{\theta}})'W_n g_n(\boldsymbol{w}_i, \hat{\boldsymbol{\theta}}) = 0 \tag{5.20}$$

with $G_n(\boldsymbol{w}_i, \hat{\boldsymbol{\theta}}) = \nabla_\theta g_n(\boldsymbol{w}_i, \hat{\boldsymbol{\theta}})$.

Step 2: Take a mean value expansion of the sample moment conditions around $\boldsymbol{\theta}_0$ and plug the FOC:

$$g_n(\boldsymbol{w}_i, \hat{\boldsymbol{\theta}}) = g_n(\boldsymbol{w}_i, \boldsymbol{\theta}_0) + G_n(\boldsymbol{w}_i, \bar{\boldsymbol{\theta}})(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \tag{5.21}$$

Substitute (5.21) into (5.20):

$$\begin{aligned}
\nabla_\theta Q_n(\hat{\boldsymbol{\theta}}) = -2G_n(\boldsymbol{w}_i, \hat{\boldsymbol{\theta}})'W_n \Big(g_n(\boldsymbol{w}_i, \boldsymbol{\theta}_0) + G_n(\boldsymbol{w}_i, \bar{\boldsymbol{\theta}})(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)\Big) = 0 \\
\iff G_n(\boldsymbol{w}_i, \hat{\boldsymbol{\theta}})'W_n G_n(\boldsymbol{w}_i, \bar{\boldsymbol{\theta}})(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) = G_n(\boldsymbol{w}_i, \hat{\boldsymbol{\theta}})'W_n g_n(\boldsymbol{w}_i, \boldsymbol{\theta}_0) \\
\implies (\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) = \Big(G_n(\boldsymbol{w}_i, \hat{\boldsymbol{\theta}})'W_n G_n(\boldsymbol{w}_i, \bar{\boldsymbol{\theta}})\Big)^{-1} G_n(\boldsymbol{w}_i, \hat{\boldsymbol{\theta}})'W_n g_n(\boldsymbol{w}_i, \boldsymbol{\theta}_0)
\end{aligned} \tag{5.22}$$

Using our moment condition, we also know that:

$$\sqrt{n}g_n(\boldsymbol{\theta}) \xrightarrow{d} N(0, \Omega) \tag{5.23}$$

The regularity conditions ensure that $W_n \xrightarrow{p} W$ and $G(\bullet) \to G$, so we can use Slutsky and the Central Moment theorem to get:

$$\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \xrightarrow{d} N(0, (G'WG)^{-1}G'W\Omega WG(G'WG)^{-1})$$

If $W = -\Omega$, or the efficient GMM estimator, then it is true that:

$$\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \xrightarrow{d} N(0, (G'\Omega^{-1}G)^{-1})$$

Nevertheless, recall that $\Omega = \mathbb{E}[(g(\boldsymbol{w}_i, \boldsymbol{\theta}_0))(g(\boldsymbol{w}_i, \boldsymbol{\theta}_0))']$, which is a function of that unknown parameter, so we need to estimate it. Suppose we estimate $\boldsymbol{\theta}_0$ using GMM with $W = I_q$. The resulting estimator of $\boldsymbol{\theta}_0$, $\tilde{\boldsymbol{\theta}}_0$,is consistent but inefficient. But, we can still use it as follows:

$$\hat{\Omega} = \frac{1}{n}\sum_{i=1}^{n}(g(\boldsymbol{w}_i, \tilde{\boldsymbol{\theta}}))(g(\boldsymbol{w}_i, \tilde{\boldsymbol{\theta}}))' \tag{5.24}$$

The GMM estimator using $\hat{\Omega}^{-1}$ will be consistent and efficient.

5.3 Extensions

Specification tests in over-identified models

GMM has the nice property of being able to deal with over identified in the most efficient way. The test statistic is based on the difference between the objective function and zero. The closer to 0, the more evidence towards the assumptions of the model; moreover, towards the validity of the assumed moment condition assumption $\mathbb{E}[g(\boldsymbol{w}_i, \boldsymbol{\theta}_0)] = 0$.

In particular, we have:

$$J = nQ_n(\boldsymbol{\theta}_0) = ng(\boldsymbol{w}_i, \boldsymbol{\theta}_0)'\Omega^{-1}g(\boldsymbol{w}_i, \boldsymbol{\theta}_0) \xrightarrow{d} \chi^2(q) \tag{5.25}$$

$$\hat{J} = nQ_n(\hat{\boldsymbol{\theta}}) = ng(\boldsymbol{w}_i, \hat{\boldsymbol{\theta}})'\Omega^{-1}g(\boldsymbol{w}_i, \hat{\boldsymbol{\theta}}) \xrightarrow{d} \chi^2(q-K) \tag{5.26}$$

Bear in mind that this property requires the estimates to be consistent.

GMM distance statistic

Consider the following null hypothesis: $H_0 : r(\boldsymbol{\theta}_0)$ with $H_1 : r(\boldsymbol{\theta}_0) \neq 0$. The efficient GMM estimator under the null (constrained) and alternative hypothesis (i.e., unconstrained) are:

$$\tilde{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}\in\Theta:r(\boldsymbol{\theta})=0}{\operatorname{argmax}} \; Q_n(\boldsymbol{\theta}) \tag{5.27}$$

$$\hat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}\in\Theta}{\operatorname{argmax}} \, Q_n(\boldsymbol{\theta}) \tag{5.28}$$

Where W_n converges in probability to Ω^{-1}. It follows that:

$$n\left(Q_n(\hat{\boldsymbol{\theta}}) - Q_n(\tilde{\boldsymbol{\theta}})\right) \xrightarrow{d} \chi^2_r \tag{5.29}$$

with r as the number of restrictions tested.

The distance statistic test can perform better than the Wald test for nonlinear hypotheses, although it is numerically equivalent to the Wald statistic for linear hypotheses. This result also holds for non-efficient GMM, as long as the same weighting matrix is used for the restricted and unrestricted estimation.

Two step estimation

Suppose we want to estimate the parameter $\boldsymbol{\theta}_0$ from the following moment condition:

$$\mathbb{E}[g(\boldsymbol{w}_i, \boldsymbol{\theta}_0, \boldsymbol{\gamma}_0)] = 0 \tag{5.30}$$

where $\boldsymbol{\gamma}_0$ is also an unknown parameter that is consistently estimated in a first step.

We are concerned with how the first stage estimation affects the consistency and asymptotic distribution of $\hat{\boldsymbol{\theta}}$. Some examples of this issue include the weighted least squares, IV in nonlinear models, Heckman sample selection, etc.

Suppose that we are able to consistently estimate γ_0 from the moment condition:

$$\mathbb{E}[m(\boldsymbol{w}_i, \boldsymbol{\gamma}_0)] = 0 \tag{5.31}$$

The estimation of $\boldsymbol{\theta}_0$ and $\boldsymbol{\gamma}_0$ can be done through GMM by stacking the vector of moment conditions:

$$\tilde{g}(\boldsymbol{w}_i, \boldsymbol{\theta}, \boldsymbol{\gamma}) = \begin{bmatrix} g(\boldsymbol{w}_i, \boldsymbol{\theta}, \boldsymbol{\gamma}) \\ m(\boldsymbol{w}_i, \boldsymbol{\gamma}) \end{bmatrix} \tag{5.32}$$

The GMM derived from this vector moment condition will be consistent, provided that the stacked vector satisfies the identification requirement from the regularity conditions (Appendix A.7).

Despite the convenience of implementing a two-step estimator as a stacked GMM estimator, it is common to estimate and analyze $\boldsymbol{\theta}_0$ by using the following 1-step sample moment condition to construct the GMM estimator:

$$\frac{1}{n}\sum_{i=1}^{n} g(\boldsymbol{w}_i, \boldsymbol{\theta}, \hat{\boldsymbol{\gamma}}) \tag{5.33}$$

That is, it is still common to estimate $\boldsymbol{\theta}_0$ while ignoring the effect of the sampling error that is included in $\hat{\boldsymbol{\gamma}}$ on the variance of $\hat{\boldsymbol{\theta}}$. The key question is whether the preliminary estimate of $\boldsymbol{\gamma}_0$ has an impact on the asymptotic variance of $\hat{\boldsymbol{\theta}}$. Can we derive $Avar(\hat{\boldsymbol{\theta}})$ as if we knew $\boldsymbol{\gamma}_0$?

If $\mathbb{E}[\nabla_\gamma g(\boldsymbol{w}_i, \boldsymbol{\theta}_0, \boldsymbol{\gamma}_0)] = 0$, then the first stage estimation error in $\hat{\boldsymbol{\gamma}}$ is not passed through $\hat{\boldsymbol{\theta}}$. In such case, we proceed with forming a GMM using $\mathbb{E}[g(\boldsymbol{w}_i, \boldsymbol{\theta}_0, \hat{\boldsymbol{\gamma}})] = 0$ and we treat $\hat{\boldsymbol{\gamma}}$ as non-stochastic. If the requirement is not satisfied then we have:

$$G_\theta = \mathbb{E}[\nabla_\theta g(\boldsymbol{w}_i, \boldsymbol{\theta}_0, \boldsymbol{\gamma}_0)] \tag{5.34}$$
$$G_\gamma = \mathbb{E}[\nabla_\theta g(\boldsymbol{w}_i, \boldsymbol{\theta}_0, \boldsymbol{\gamma}_0)] \tag{5.35}$$
$$M = \mathbb{E}[\nabla_\gamma m(\boldsymbol{w}_i, \boldsymbol{\gamma}_0)] \tag{5.36}$$
$$\Psi(\boldsymbol{w}_i) = -M^{-1} m(\boldsymbol{w}_i, \boldsymbol{\gamma}_0) \tag{5.37}$$

Then, provided that $\hat{\boldsymbol{\gamma}} \xrightarrow{p} \boldsymbol{\gamma}_0$ and $\mathbb{E}[g(\boldsymbol{w}_i, \boldsymbol{\theta}_0, \boldsymbol{\gamma}_0)] = 0$, in addition with all the regularity conditions for GMM estimators, we finally have:

$$\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \xrightarrow{d} N(0, V) \tag{5.38}$$
$$V = G_\theta^{-1}\mathbb{E}\left[(g(\boldsymbol{w}_i, \boldsymbol{\theta}_0, \boldsymbol{\gamma}_0) + G_\gamma \Psi(\boldsymbol{w}_i))(g(\boldsymbol{w}_i, \boldsymbol{\theta}_0, \boldsymbol{\gamma}_0) + G_\gamma \Psi(\boldsymbol{w}_i))'\right] G_\theta^{-1} \tag{5.39}$$

6 Binary Response Models

Binary response models deal with limited dependent variables such as binary responses, count/multinomial data, censored regression models and continuous duration analysis. We will cover each of those and see how they relate to the extremum estimators we have studied so far.

6.1 Binary Dependent Variables

Models with binary dependent variables are those such that $y_i \in \{0, 1\}$. These models include measures such application approval, currently employed, etc. Given this binary nature for the dependent variable, the common linear model $\mathbb{E}[y_i|\boldsymbol{x}_i]$ might be inappropriate. Accordingly, we have two ways of dealing with this issue, probit and logit.

In binary response models we care about estimating the probability that y_i takes the value 1, conditional on the covariates. This is often called **response probability** or **probability of success**, although the event might not necessarily be a success. Formally:

$$P(\boldsymbol{x}_i) = Pr(\boldsymbol{y}_i = 1|\boldsymbol{x}_i) \tag{6.1}$$

Because $y_i \in \{0, 1\}$, $y_i|\boldsymbol{x}_i$ is a Bernoulli random variable, $p(\boldsymbol{x}_i)$ fully characterizes the distribution and its properties:

$$\begin{aligned}
&Pr(y_i = 1|\boldsymbol{x}_i) = f(1|\boldsymbol{x}_i) = p(\boldsymbol{x}_i)\\
&Pr(y_i = 0|\boldsymbol{x}_i) = f(0|\boldsymbol{x}_i) = 1 - p(\boldsymbol{x}_i)\\
&\mathbb{E}[y_i|\boldsymbol{x}_i] = p(\boldsymbol{x}_i)\\
&Var(y_i|\boldsymbol{x}_i) = p(\boldsymbol{x}_i)(1 - p(\boldsymbol{x}_i))
\end{aligned}$$

The linear probability model (LPM) for a binary response model is then given by:

$$Pr(y_i = 1|\boldsymbol{x}_i) = \mathbb{E}[y|\boldsymbol{x}_i] = \boldsymbol{x}_i'\beta \tag{6.2}$$

So we could write the regression equation:

$$\begin{aligned}
y_i &= \mathbb{E}[y_i|\boldsymbol{x}_i] + \epsilon_i\\
&= \boldsymbol{x}_i'\beta + \epsilon_i
\end{aligned} \tag{6.3}$$

LPM is a linear regression when the dependent variable is $\{0, 1\}$. The marginal change in x_{ij} (when continuous), is given by:

$$\frac{\partial}{\partial x_{ij}} Pr(y_i = 1|\boldsymbol{x}_i) = \beta_j \tag{6.4}$$

with β_j measuring the increase in the response probability associated with a unit increase in x_{ij}, holding all other x's constant. Note that the marginal effect is constant, and therefore it does not change with the value of x_{ij}.That property can be hard to justify in many situations.

The LPM has pros and cons, on one side it is the best linear predictor (BLP) of $y_i|\boldsymbol{x}_i$ in the MSE sense, which requires one weak functional assumptions. It imposes no distributional assumptions on $\epsilon|\boldsymbol{x}_i$ other than having mean 0. It is easy to interpret, and its limiting distribution is known and simple:

$$\sqrt{n}(\hat{\beta} - \beta) \xrightarrow{d} N(H^{-1}JH^{-1}) \tag{6.5}$$

$$H = \mathbb{E}[\boldsymbol{x}_i\boldsymbol{x}_i'] \tag{6.6}$$

$$J = \mathbb{E}[\epsilon^2\boldsymbol{x}_i\boldsymbol{x}_i'] \tag{6.7}$$

Nonetheless and given its structure, PLM can predict probabilities greater than 1 or smaller than 0 as a direct result of the linearity of the conditional expected function, which we saw rendered constant marginal effects. In other words, that a 1-unit increase in x_{ij} always changes $Pr(y_i = 1|\boldsymbol{x}_i)$ by the same amount regardless of the initial value of x_{ij}. We can fix this linearity issue by completely saturating the model, or having 1 dummy variable for every permutation vector $\boldsymbol{x}_i$, then the PLM will be completely general. In this case, the fitted probabilities are simply the average of y_i in each of the cells defined by vector $\boldsymbol{x}_i$, which are always between 0 and 1 by definition.

Another issue with LPM is that the model is heteroskedactic by construction since the variance of $y_i|\boldsymbol{x}_i$ is $p(\boldsymbol{x}_i|)(1-p(\boldsymbol{x}_i))$. This property affects inference, but not the consistency of the estimator. We could use heteroskedastic-robust inferences, or correct by using weighted least squares where the weight is given by:

$$\sqrt{p(\boldsymbol{x}_i)(1 - p(\boldsymbol{x}_i))} \tag{6.8}$$

6.2 Probit and Logit

To solve the shortcuts of a constant marginal effect on the covariates, we can use models that restrict the functional form of the response probability such that $0 < Pr(y_i = 1|\boldsymbol{x}_i) < 1$. Formally:

$$Pr(y_i = 1|\boldsymbol{x}_i) = F(\boldsymbol{x}_i'\beta), \quad 0 < F(z) < 1 \,\forall\, z \tag{6.9}$$

These are called index models because they efficiently restrict the way in which the response probability depends on $\boldsymbol{x}_i$. $Pr(y_i = 1|\boldsymbol{x}_i)$ depends on $\boldsymbol{x}_i$ only through the index $\boldsymbol{x}_i'\beta$. In most applications $F(\bullet)$ will be a cumulative distribution function. Now we can introduce two index models known as the probit and logit models:

- **Probit model**:

$$Pr(y_i = 1|\boldsymbol{x}_i) = \Phi(\boldsymbol{x}_i'\beta) \tag{6.10}$$

- **Logit model**:

$$Pr(y_i = 1|\boldsymbol{x}_i) = \Lambda(\boldsymbol{x}_i'\beta) = \frac{e^{\boldsymbol{x}_i'\beta}}{1 + e^{\boldsymbol{x}_i'\beta}} \tag{6.11}$$

Note that $\Phi(\bullet)$ is the cdf of a standard normal distribution $\Lambda(\bullet)$ is the cdf of the standard logistic distribution

<u>**Latent Variable**</u>

The specific functional form of $F(\bullet)$ can be derived from a latent variable model, where latent refers to partially observed. Consider the following model:

$$y_i^* = \boldsymbol{x}_i'\beta + \epsilon_i \tag{6.12}$$

$$y_i = \mathbb{I}(y_i^* > c) \tag{6.13}$$

with ϵ_i as a continuously iid randomly distributed variable around zero and independent of $\boldsymbol{x}_i$. The threshold c is a number, but we don't need to know this number, as it only changes the interpretation of the value and we normalize it to zero. To apply the latent variable idea, consider the applications below:

1. **Probit model**: Assume $\epsilon_i \sim N(0,1)$, note the normalization to $\sigma^2 = 1$:

$$\begin{aligned} y_i^* &= \boldsymbol{x}_i'\beta + \epsilon_i \\ y_i &= \mathbb{I}(y_i^* > 0) \end{aligned}$$

$$\begin{aligned} Pr(y_i = 1|\boldsymbol{x}_i) &= Pr(\epsilon_i > -\boldsymbol{x}_i'\beta|\boldsymbol{x}) \\ &= 1 - \Phi(-\boldsymbol{x}_i'\beta) \\ &= \Phi(\boldsymbol{x}_i'\beta) \end{aligned} \tag{6.14}$$

2. **Logit model**: Assume $\epsilon_i|\boldsymbol{x}_i \sim Log(0,1)$. Normalize $Var(\epsilon_i|\boldsymbol{x}_i) = \pi^2/3$. Proceeding:

$$\begin{aligned} Pr(y_i = 1|\boldsymbol{x}_i) &= Pr(\epsilon_i > -\boldsymbol{x}_i'\beta|\boldsymbol{x}_i) \\ &= 1 - \Lambda(-\boldsymbol{x}_i'\beta) \\ &= \Lambda(\boldsymbol{x}_i'\beta) \end{aligned} \tag{6.15}$$

Marginal Effects in Probit and Logit models

For continuous regressors, the marginal effects in these models are given by:

1. **Probit model**:

$$\frac{\partial Pr(y_i = 1|\boldsymbol{x}_i)}{\partial x_{ij}} = \frac{\partial \Phi(\boldsymbol{x}_i'\beta)}{\partial x_{ij}} = \beta_j \phi(\boldsymbol{x}_i'\beta) \tag{6.16}$$

2. **Logit model**:

$$\frac{\partial Pr(y_i = 1|\boldsymbol{x}_i)}{\partial x_{ij}} = \frac{\partial \Lambda(\boldsymbol{x}_i'\beta)}{\partial x_{ij}} = \beta_j \frac{e^{\boldsymbol{x}_i'\beta}}{1 + e^{\boldsymbol{x}_i'\beta}} \tag{6.17}$$

Note that both marginal effects are a function of the coefficients and the respective pdfs for each model. So, marginal effects change as x_{ij} changes since the pdf is not a linear function. Recall that by definition $f(\bullet) > 0$, so the sign of β_j will determine the sign of the marginal effect. It's important to note that the coefficients are informative about the sign of the marginal effect at the point of estimation. One should report the average marginal effect (value evaluated at some vector $\boldsymbol{x}$ such as $\bar{\boldsymbol{x}}$) instead of the ML estimates of β.

6.3 Estimation

Binary response models are usually estimated using ML techniques, given that we have already specified a distribution in the latent variable representation. Suppose we have $\boldsymbol{w}_i$, where each observation on $y_i|\boldsymbol{x}_i$ is drawn from the Bernoulli distribution $f(0|\boldsymbol{x}_i) = 1 - p(\boldsymbol{x}_i)$ and $f(1|\boldsymbol{x}_i) = p(\boldsymbol{x}_i)$. In this case, the pdf of $y_i|\boldsymbol{x}_i$ is given by:

$$\begin{aligned} f(y_i|\boldsymbol{x}_i;\theta_0) &= Pr(y_i = 1|\boldsymbol{x}_i)^{y_i} Pr(y_i = 0|\boldsymbol{x}_i)^{1-y_i} \\ &= F(\boldsymbol{x}_i'\theta_0)^{y_i}(1 - F(\boldsymbol{x}_i'\theta_0))^{1-y_i} \end{aligned} \tag{6.18}$$

The joint pdf of the sample is given by the product of the marginal effects:

$$\begin{aligned} f(\boldsymbol{w}_1, ..., \boldsymbol{w}_n; \theta_0) &= \prod_{i=1}^{n} F(\boldsymbol{x}_i'\theta_0)^{y_i}(1 - F(\boldsymbol{x}_i'\theta_0))^{1-y_i} \\ \ln\left[L(\theta)\right] &= \sum_{i=1}^{n} y_i \ln[F(\boldsymbol{x}_i'\theta_0)] + (1 - y_i)\ln[(1 - F(\boldsymbol{x}_i'\theta_0))] \qquad \text{(Taking logs)} \end{aligned} \tag{6.19}$$

6.4 Large Sample Properties

We define the binary response extremum estimator as:

$$\hat{\theta} = \operatorname*{argmax}_{\theta\in\Theta} \frac{1}{n}\sum_{i=1}^{n} y_i F(\boldsymbol{x}_i'\theta_0) + (1 - y_i)(1 - F(\boldsymbol{x}_i'\theta_0)) \tag{6.20}$$

The probit and logit models estimators of θ are obtained by plugging their respective form of $F(\boldsymbol{x}_i'\theta)$. Provided that we satisfy the usual regularity conditions (i.e., continuity, uniform convergence, etc. Appendix A.5), the estimator is consistent.

On the other hand, assuming that we specified the correct distribution, and that we satisfy the required regularity conditions (Appendix A.6), the MLE with binary dependent variables is asymptotically normal, with limiting distribution:

$$\sqrt{n}(\hat{\theta} - \theta_0) \xrightarrow{d} N(0, J^{-1}) \tag{6.21}$$

Because we proceeded with MLE, the information matrix equality holds, and the asymptotic variance is given by:

$$J = \mathbb{E}\left[\frac{f(\boldsymbol{x}_i'\theta_0)^2\boldsymbol{x}_i\boldsymbol{x}_i'}{F(\boldsymbol{x}_i'\theta_0)(1 - F(\boldsymbol{x}_i'\theta_0))}\right] \tag{6.22}$$

7 Multinomial Response Models

Multinomial responses deal with cases in which $y_i = \{0, 1, ..., J\}$, and the choices or alternatives are mutually exclusive. For these type of models, we have two cases: (1) ordered multinomial responses, where the values attached to outcomes matter, and (2) unordered multinomial responses where the values attached to the outcomes is arbitrary, possessing no cardinal significance.

Just like binary response models, the goal is to model the response probabilities as functions of covariates. This time, however, we do this for each alternative j:

$$Pr(y_i = j|\boldsymbol{x}_i) \equiv p_j(\boldsymbol{x}_i|\theta) \tag{7.1}$$

It is important to note that different models will lead to different parametric forms for $p_j(\boldsymbol{x}_i, \theta)$. The object of interest is to estimate the unknown parameter θ by MLE, and its marginal effects:

$$\frac{\partial}{\partial x_{ij}} p_j(\boldsymbol{x}_i, \theta) \tag{7.2}$$

7.1 Estimation

For observation i, the contribution to the log-likelihood is given by:

$$\ln[I_i(\theta)] = \sum_{j=0}^{J} \mathbb{I}(y_i = j) \ln[p_j(\boldsymbol{x}_i, \theta)] \tag{7.3}$$

Note that for each i, only one of the indicator functions $\mathbb{I}(y_i = j)$, $j = \{0, 1, ..., J\}$ equals 1, so we've taken care of any double counting. The log-likelihood function then takes the form:

$$\ln[L(\theta)] = \sum_{i=1}^{n} \ln[I_i(\theta)] = \sum_{i=1}^{n} \sum_{j=1}^{J} \mathbb{I}(y_i = j) \ln[p_j(\boldsymbol{x}_i, \theta)] \tag{7.4}$$

In some models, the vector of regressors can be alternative-specific (x_{ij}), and the vector of parameters can be alternative specific as well (θ_j).

7.2 Large Sample Properties

Multinomial models are consistent and asymptotically normal under the usual regularity conditions we established for MLE (Appendix A.5-A.6). In particular:

$$\sqrt{n}(\hat{\theta} - \theta_0) \xrightarrow{d} N(0, HJ^{-1}H) \tag{7.5}$$

with:

$$J^{-1} = \mathbb{E}\left[(\nabla_\theta \ln[I_i(\theta_0)])(\nabla_\theta \ln[I_i(\theta_0)])'\right] \tag{7.6}$$

Nevertheless and due to misspecification, it is recommended to use the more robust formula $Avar(\hat{\theta}) = H^{-1}JH^{-1}$, where $H = \mathbb{E}[\nabla_{\theta\theta} \ln[I_i(\theta_0)]]$.

7.3 Ordered Multinomial Response Models

Suppose y corresponds to a an order response, taking values $j = \{0, 1, ..., J\}$. The ordered probit can be derived using the standard latent variable model:

$$y_i^* = \boldsymbol{x}_i'\beta + \epsilon_i, \quad \epsilon_i|\boldsymbol{x}_i \sim N(0,1)$$

where $\boldsymbol{x}_i$ is a $[K \times 1]$ vector that excludes a constant term. The ordered logit is derived similarly, replacing the normal distribution by the logistic distribution.

Now, define $\alpha_1 < \alpha_2 < ... < \alpha_J$ be unknown thresholds or cut points such that:

$$\begin{aligned} y_i &= 0; \quad \text{if } y_i^* \leq \alpha_1 \\ y_i &= j; \quad \text{if } \alpha_j < y_i^* \leq \alpha_{j+1},\ \forall\, j = 1, ..., J-1 \\ y_i &= J; \quad \text{if } y_i^* \geq \alpha_J \end{aligned} \tag{7.7}$$

For example, if $y = \{0, 1, 2\}$, there are two thresholds α_1 and α_2. One could also interpret having one $J+1$ cut points, with $\alpha_0 = -\infty$ and $\alpha_{J+1} = +\infty$. Under the normality assumption, we can derive the $J+1$ response probabilities:

$$\begin{aligned} Pr(y_i = 0|\boldsymbol{x}_i) &\equiv p_0(\boldsymbol{x}_i, \theta) = \Phi(\alpha_1 - \boldsymbol{x}_i'\beta) \\ Pr(y_i = j|\boldsymbol{x}_i) &\equiv p_j(\boldsymbol{x}_i, \theta) = \Phi(\alpha_{j+1} - \boldsymbol{x}_i'\beta) - \Phi(\alpha_j - \boldsymbol{x}_i'\beta),\ \forall\, j = 1, 2, ..., J-1 \\ Pr(y_i = J|\boldsymbol{x}_i) &\equiv p_J(\boldsymbol{x}_i, \theta) = 1 - \Phi(\alpha_J - \boldsymbol{x}_i'\beta) \end{aligned} \tag{7.8}$$

Note that if $J = 1$, we return to the binary probit model:

$$Pr(y_i = 1|\boldsymbol{x}_i) = 1 - Pr(y_i = 0|\boldsymbol{x}_i) = 1 - \Phi(\alpha_1 - \boldsymbol{x}_i'\beta) = \Phi(\boldsymbol{x}_i'\beta - \alpha_1) \tag{7.9}$$

and so α_1 is the intercept.

The marginal effects change in response probability for small changes in x_{ik}, and are similar to the ones derived in the binary model:

$$\begin{aligned} \frac{\partial p_0(\boldsymbol{x}_i, \theta)}{\partial x_{ik}} &= -\beta_k \phi(\alpha_1 - \boldsymbol{x}_i'\beta) \\ \frac{\partial p_j(\boldsymbol{x}_i, \theta)}{\partial x_{ik}} &= \beta_k[\phi(\alpha_j - \boldsymbol{x}_i'\beta) - \phi(\alpha_{j+1} - \boldsymbol{x}_i'\beta)] \qquad j = 1, 2, ..., J-1 \\ \frac{\partial p_J(\boldsymbol{x}_i, \theta)}{\partial x_{ik}} &= \beta_k \phi(\alpha_J - \boldsymbol{x}_i'\beta) \end{aligned} \tag{7.10}$$

Similar assumptions and derivations hold for the logit model, except that it is based on the logistic distribution:

$$\begin{aligned}
&Pr(y_i = 0|\boldsymbol{x}_i) \equiv p_0(\boldsymbol{x}_i, \theta) = \Lambda(\alpha_1 - \boldsymbol{x}_i'\beta) \\
&Pr(y_i = j|\boldsymbol{x}_i) \equiv p_j(\boldsymbol{x}_i, \theta) = \Lambda(\alpha_{j+1} - \boldsymbol{x}_i'\beta) - \Lambda(\alpha_j - \boldsymbol{x}_i'\beta),\ \forall\, j = 1, 2, ..., J-1 \\
&Pr(y_i = J|\boldsymbol{x}_i) \equiv p_J(\boldsymbol{x}_i, \theta) = 1 - \Lambda(\alpha_J - \boldsymbol{x}_i'\beta)
\end{aligned} \tag{7.11}$$

with $\Lambda(z) = e^z/(1+e^z)$.

7.4 Unordered Multinomial Response Models

In this case, the choice variable y takes non-negative integer values with more than 2 outcomes $y_i = \{0, 1, ..., J\}$, and the order is irrelevant. The goal is to model response probabilities as a function of the covariates.

There are several ways to model this problem, a key issue is to correctly identify the nature of the regressors. More specifically, if the regressors are (1) constant across alternatives (i.e. age, education), (2) varying across alternatives, same for individuals (i.e. cost bus ticket in a transportation model), or (3) varying across alternatives and individuals (i.e. commuting time in a transportation choice model).

1. **Multinomial Logit**: When the choice depends on the characteristics of individual i, but not on attributes of the alternatives, it is typical to use a multinomial logit model (MNL):

$$Pr(y_i = j|\boldsymbol{x}) \equiv p_j(\boldsymbol{x}_i, \beta) = \frac{e^{\boldsymbol{x}_i'\beta_j}}{1 + \sum\limits_{m=1}^{J} e^{\boldsymbol{x}_i'\beta_m}} \qquad j = 1, 2, ..., J \tag{7.12}$$

Because the response probabilities sum to 1, we impose the natural restriction:

$$Pr(y_i = 0|\boldsymbol{x}) \equiv p_0(\boldsymbol{x}_i, \beta) = \frac{1}{1 + \sum\limits_{m=1}^{J} e^{\boldsymbol{x}_i'\beta_m}} \tag{7.13}$$

with marginal effects:

$$\frac{\partial p_j(\boldsymbol{x}_i, \beta)}{\partial x_{ik}} = p_j(\boldsymbol{x}_i, \beta) \left(\frac{\beta_{jk} - \sum\limits_{m=1}^{n} \beta_{mk} e^{\boldsymbol{x}_i, \beta}}{1 + \sum\limits_{m=1}^{n} e^{\boldsymbol{x}_i, \beta}} \right) \tag{7.14}$$

In this case, the sign of β_{jk} is uninformative about directional effects, unlike the binary and ordered models. An easier interpretation can be obtained by using the log of the odds-ratio. The odds-ratio between the base “0” alternative, and the jth alternative is given by:

$$\frac{p_j(\boldsymbol{x}_i, \beta)}{p_0(\boldsymbol{x}_i, \beta)} = e^{\boldsymbol{x}_i, \beta_j} \qquad j = 1, 2, ..., J \tag{7.15}$$

Thus, the change in the odds-ratio for a small change in x_{ik} is given by:

$$\frac{\partial}{\partial x_{ik}} \frac{p_j(\boldsymbol{x}_i, \beta)}{p_0(\boldsymbol{x}_i, \beta)} = \beta_{ik} e^{\boldsymbol{x}_i, \beta_j} \qquad j = 1, 2, ..., J \tag{7.16}$$

A positive β_{ik} means that an increase in x_{ik} also increases the probability of choosing option j relative to the base option 0. The log odds-ratio is then given by:

$$\ln\left[\frac{p_j(\boldsymbol{x}_i,\beta)}{p_0(\boldsymbol{x}_i,\beta)}\right] = \boldsymbol{x}_i, \beta_j \qquad\qquad j = 1, 2, ..., J \tag{7.17}$$

with β_{jk} as the marginal effect of x_{ik} on the log-odds of choosing alternative j relative to the base alternative "0".

2. **Conditional Logit**: The conditional logit model (CL) is appropriate when choices depend on the characteristics of each alternative, or possibly when they depend on individuals across alternatives. The CL is similar to MNL, in fact MNL can be derived from the CL model under restrictions about the types of covariates and the alternative vector; likewise, CL can be derived from a utility maximization/latent variable framework, like the binary logit model.

   Consider the probabilistic choice model:

$$y^*_{ij} = \boldsymbol{x}'_{ij}\beta + \epsilon_{ij} \qquad\qquad j = 0, 1, 2, ..., J \tag{7.18}$$

   where ϵ_{ij} are unobservable factors affecting tastes for each alternative j (independent across j), and $\boldsymbol{x}_{ij}$ is a $[K \times 1]$ vector that varies across alternatives and possibly across individuals across alternatives. For instance, suppose j indexes alternative modes of transportation, and $\boldsymbol{x}_{ij}$ the time associated with transportation alternative j for individual i and the price of bus tickets.

   We restrict $\boldsymbol{x}_{ij}$ to include only elements that vary across j. For example, $\boldsymbol{x}_{ij}$ excludes a constant term. Individuals choose the alternative j that maximizes their utility:

$$y_i = \text{argmax}(y^*_{i0}, y^*_{i1}, ..., y^*_{iJ}) \tag{7.19}$$

   Depending on the distribution of ϵ_{ij}, evaluating the probability that y_i takes values between 0 and J can be challenging, and in general requires the computation of a $J-1$ dimensional integral.

   In 1974, McFadden showed that if errors ϵ_{ij} follow an iid type I extreme value distribution, are independent across alternatives, and independent of $\boldsymbol{x}_{ij}$, then:

$$Pr(y_i = j|\boldsymbol{x}_i) \equiv p_j(\boldsymbol{x}_i,\beta) = \frac{e^{\boldsymbol{x}'_{ij}}}{\sum\limits_{m=0}^{J} e^{\boldsymbol{x}'_{ij}\beta}} \qquad\qquad j = 0, 1, 2, ..., J \tag{7.20}$$

   The extreme value type I distribution has a unique mode at 0, and a variance of 1.65. The cdf is given by: $F(z) = e^{-e^{z}}$.

   The marginal effects in the conditional logit model have the usual interpretation. That is, the sign on β_k is informative about the directional effect of a change in the regressor x_{ijk}:

$$\frac{\partial p_j(\boldsymbol{x}_i,\beta)}{\partial x_{ijk}} = \beta_k p_j(\boldsymbol{x}_i,\beta)(1 - p_j(\boldsymbol{x}_i,\beta)) \qquad\qquad j = 1, 2, ..., J \tag{7.21}$$

   The main issue with the MNL and the CL models is the property of Independence of Irrelevant Alternatives (IIA). Note that the odd-ratios between 2 alternatives (say h and m) only depend on the characteristics of the h & m alternatives:

$$\frac{p_h(\boldsymbol{x}_i,\beta)}{p_m(\boldsymbol{x}_i,\beta)} = \frac{e^{\boldsymbol{x}'_{ih}\beta}}{e^{\boldsymbol{x}'_{im}\beta}} \tag{7.22}$$

This issue implies that adding another alternative, no matter how close a substitute it is for h or m, it will not change odds-ratio between h & m. This problems arises directly as a consequence of the errors assumed to be uncorrelated across alternatives.

To solve this problem we need to allow for correlation across alternatives, where there are several approaches to do so:

(a) Nested Logit Model: Nest similar alternatives in subsets, where every alternative must be assigned to only one subset.

(b) Random Coefficient Logit Model: It allows for person-specific marginal utilities in a latent variable model:

$$y^*_{ij} = \boldsymbol{x}'_{ij}\beta_i + \epsilon_{ij} \qquad j = 0, 1, 2, ..., J \tag{7.23}$$

which we can rewrite as:

$$y^*_{ij} = \boldsymbol{x}'_{ij}\beta_i + (\epsilon_{ij} + \boldsymbol{x}'_{ij}(\beta_i - \bar{\beta})) \qquad j = 0, 1, 2, ..., J \tag{7.24}$$

and then require instruments to proceed.

(c) Multinomial Probit: We can allow for correlated errors in a multivariate probit

$$y^*_{ij} = \boldsymbol{x}'_{ij}\beta_i + \boldsymbol{\epsilon}_{ij} \qquad j = 0, 1, 2, ..., J \tag{7.25}$$

with:

$$\boldsymbol{\epsilon} = vec(\epsilon_{ij}) \tag{7.26}$$

$$\boldsymbol{\epsilon} \sim (\mathbf{0}, \boldsymbol{\Sigma}) \tag{7.27}$$

where $\boldsymbol{\Sigma}$ includes non zero entries on the off-diagonal to allow for correlations across alternatives. This method is also computationally challenging, and it's not as commonly used.

8 Censored Regression Models

These models deal with data in which the dependent variable is censored, but the independent variable is perfectly observed. Consider the latent variable model:

$$y^*_i = \boldsymbol{x}'_{ij}\beta_i + \epsilon_i \tag{8.1}$$

$$\epsilon_i|\boldsymbol{x}_i \sim N(0, \sigma^2) \tag{8.2}$$

Accordingly, we can have data that is censored from above $y_i = \min\{y^*_i, U\}$, censored form below $y_i = \max\{y^*_i, L\}$, or both.

8.1 Tobit Models

The tobit model is obtained by setting the threshold L to 0 in the censored regression model: $y_i = \max\{y^*_i, 0\}$. Otherwise, the issue is that without normalization, the intercept of the index $\boldsymbol{x}'\beta$

is meaningless. Note that changing from min to max (accommodating the censoring from above or below) does not change the magnitude of the estimator for β, but it does change the sign:

$$\begin{aligned} & y_i = \min(y_i^*, U) \\ \iff & y_i - U = \min(y_i^* - U, 0) \\ \iff & -(y_i - U) = \max(U - y_i^*, 0) \end{aligned}$$

If we use a linear regression of y_i on $\boldsymbol{x}_i$, it yields an inconsistent estimator for β, since $\mathbb{E}[y_i|\boldsymbol{x}_i]$ is not linear in this model:

$$\mathbb{E}[y_i|\boldsymbol{x}_i] = Pr(y_i = 0|\boldsymbol{x}_i) \times 0 + Pr(y_i > 0|\boldsymbol{x}_i) \times \mathbb{E}[y_i|y_i > 0, \boldsymbol{x}_i] \tag{8.3}$$

Now consider the following expression:

$$\begin{aligned} Pr(y_i > 0|\boldsymbol{x}_i) &= Pr(y_i^* > 0|\boldsymbol{x}_i) \\ &= Pr(\frac{\epsilon_i}{\sigma} > -\frac{\boldsymbol{x}_i'\beta}{\sigma}|\boldsymbol{x}_i) \\ &= 1 - \Phi\left(-\frac{\boldsymbol{x}_i'\beta}{\sigma}\right) = \Phi\left(\frac{\boldsymbol{x}_i'\beta}{\sigma}\right) \end{aligned} \tag{8.4}$$

Invoking the Mills Ration[2] on theconditional mean:

$$\begin{aligned} \mathbb{E}[y_i|y_i > 0, \boldsymbol{x}_i] &= \mathbb{E}[\boldsymbol{x}_i'\beta + \epsilon_i|\epsilon_i > -\boldsymbol{x}_i'\beta, \boldsymbol{x}_i] \\ &= \boldsymbol{x}_i'\beta + \mathbb{E}[\epsilon_i|\epsilon_i > -\boldsymbol{x}_i'\beta, \boldsymbol{x}_i] \\ &= \boldsymbol{x}_i'\beta + \sigma\mathbb{E}\left[\frac{\epsilon_i}{\sigma}|\frac{\epsilon_i}{\sigma} > -\frac{\boldsymbol{x}_i'\beta}{\sigma}, \boldsymbol{x}_i\right] \\ &= \boldsymbol{x}_i'\beta + \sigma\frac{\phi\left(\frac{-\boldsymbol{x}_i'\beta}{\sigma}\right)}{1 - \Phi\left(\frac{-\boldsymbol{x}_i'\beta}{\sigma}\right)} \\ &= \boldsymbol{x}_i'\beta + \sigma\frac{\phi\left(\frac{\boldsymbol{x}_i'\beta}{\sigma}\right)}{\Phi\left(\frac{\boldsymbol{x}_i'\beta}{\sigma}\right)} \end{aligned} \tag{8.7}$$

which implies:

$$\begin{aligned} \mathbb{E}[y_i|\boldsymbol{x}_i] &= Pr(y_i > 0|\boldsymbol{x}_i) \times \mathbb{E}[y_i|y_i > 0, \boldsymbol{x}_i] \\ &= \Phi\left(\frac{\boldsymbol{x}_i'\beta}{\sigma}\right)\left(\boldsymbol{x}_i'\beta + \sigma\frac{\phi\left(\frac{\boldsymbol{x}_i'\beta}{\sigma}\right)}{\Phi\left(\frac{\boldsymbol{x}_i'\beta}{\sigma}\right)}\right) \end{aligned} \tag{8.8}$$

$$\tag{8.9}$$

[2] Let $Z \sim N(\mu, \sigma^2)$ and d be a constant, then:

$$\mathbb{E}[Z|Z > d] = \mu + \sigma\frac{\phi\left(\frac{d-\mu}{\sigma}\right)}{1 - \Phi\left(\frac{d-\mu}{\sigma}\right)} \tag{8.5}$$

if $d = 0$:

$$\mathbb{E}[Z|Z > 0] = \mu + \sigma\frac{\phi\left(\frac{\mu}{\sigma}\right)}{\Phi\left(\frac{\mu}{\sigma}\right)} \tag{8.6}$$

In the statistics literature, equation (8.5) is known as the Hazard Function, while equation (8.6) is known as the Inverse Mills Ratio.

Therefore, a linear regression of y_i on $\boldsymbol{x}_i$ yields an inconsistent estimator for β, but the non-linear regression based on the expression for $\mathbb{E}[y_i|\boldsymbol{x}_i]$ yields a consistent estimator. Because of this property, it makes sense to use a maximum likelihood estimator, note that $y_i|\boldsymbol{x}_i \sim N(\boldsymbol{x}_i'\beta, \sigma^2)$ for $y_i > 0$, so:

$$\begin{aligned} f(y_i|y_i > 0, \boldsymbol{x}_i; \beta, \sigma) &= \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\left(\frac{y_i - \boldsymbol{x}_i'\beta}{\sigma}\right)^2} \\ &= \frac{1}{\sigma}\phi\left(\frac{y_i - \boldsymbol{x}_i'\beta}{\sigma}\right) \end{aligned} \tag{8.10}$$

Let $y_{0i} = \mathbb{I}(y_i = 0)$. The density of $y_i|\boldsymbol{x}_i$ is then given by:

$$f(y_i|y_i > 0, \boldsymbol{x}_i; \beta, \sigma) = \left(1 - \Phi\left(\frac{\boldsymbol{x}_i\beta}{\sigma}\right)\right)^{y_{i0}} \left(\frac{1}{\sigma}\phi\left(\frac{y_i - \boldsymbol{x}_i\beta}{\sigma}\right)\right)^{1-y_{0i}} \tag{8.11}$$

with the following log-likelihood function:

$$\ln[L(\beta, \sigma)] = \sum_{i=1}^{n} \left[\underbrace{y_{0i} \ln\left(1 - \Phi\left(\frac{\boldsymbol{x}_i\beta}{\sigma}\right)\right)}_{\text{Censored data}} + \underbrace{(1 - y_{0i}) \ln\left(\frac{1}{\sigma}\phi\left(\frac{y_i - \boldsymbol{x}_i\beta}{\sigma}\right)\right)}_{\text{Not censored data}} \right] \tag{8.12}$$

It is important to note that the first section of the tobit model is from a probit model, and the second component is from a regression model with normal errors. The tobit model allows for the estimation of β and σ while the intercept remains meaningless. The MLE is consistent and asymptotically normal under the usual assumptions. Nonetheless, the tobit MLE is generally inconsistent under mis-specification, or in the presence of heteroskedasticity.

8.2 Semi-parametric Estimators

Semi-parametric estimators rely less on distributional assumption and are now the standard approach for dealing with censored data. The main observation is that for a convex function $g(X)$, $\mathbb{E}[g(X)] \geq g(\mathbb{E}[X])$ but $med[g(X)] = g(med[X])$. The Censored Least Absolute Deviations (CLAD)estimator uses this property by:

$$y_i^* = \boldsymbol{x}_i'\beta + \epsilon_i \tag{8.13}$$

$$med(\epsilon_i|\boldsymbol{x}_i) = 0 \tag{8.14}$$

$$y_i = \max(y_i^*, 0) \tag{8.15}$$

Therefore, we replace the normality assumption with the $med(\epsilon_i|\boldsymbol{x}_i) = 0$. It follows that:

$$\begin{aligned} med(y_i|\boldsymbol{x}_i) &= med(\max(y_i^*, 0)|\boldsymbol{x}_i) \\ &= \max(0, med(y_i^*, 0)|\boldsymbol{x}_i) \\ &= \max(0, \boldsymbol{x}_i\beta) \end{aligned} \tag{8.16}$$

To proceed with the estimation of β, we now use the fact that the mean absolute prediction error $\mathbb{E}[|y_i - g(\boldsymbol{x}_i)|]$ is minimized at $g(\boldsymbol{x}_i) = med(y_i|\boldsymbol{x}_i)$. Therefore, the CLAD estimator is defined as:

$$\hat{\beta} = \operatorname*{argmin}_{b \in B} \frac{1}{n} \sum_{i=1}^{n} |y_i - \max(0, \boldsymbol{x}_i\beta)| \tag{8.17}$$

This estimator is consistent and asymptotically normal, but only if additional conditions are met because the function is not twice continuously differentiable.

9 Appendix

A Main Theorems and Lemmas from Newey and Mc Fadden (1994)

This appendix summarizes the main theorems and lemmas that give the regularity conditions for extremum estimators. Let $\{\boldsymbol{w}_i = (y_i, \boldsymbol{x}_i) : \ i = 1, .., n\}$ be a sample of iid data. Let $\boldsymbol{\theta}_0$ be the true value parameter of interest, and $\hat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}\in\Theta}{\operatorname{argmax}} = \frac{1}{n}\sum\limits_{i=1}^{n} a(\boldsymbol{w}_i, \boldsymbol{\theta})$ be an estimator of $\boldsymbol{\theta_0}$.

A.1 Uniform Law of Large Numbers

Let $a(\boldsymbol{w}_i, \boldsymbol{\theta})$ be a function from $\Theta \to \mathbb{R}$. Lemma (2.4) says that if:

1. $\boldsymbol{w}_i$ is iid.
2. Θ is compact.
3. $a(\boldsymbol{w}_i, \boldsymbol{\theta})$ is continuous $\forall\ \boldsymbol{\theta} \in \Theta$.
4. $\exists d(\boldsymbol{w}_i) : ||a(\boldsymbol{w}_i, \boldsymbol{\theta})|| \leq d(\boldsymbol{w}_i), ;\ \forall\ \boldsymbol{\theta} \in \Theta$ and $\mathbb{E}[d(\boldsymbol{w}_i)] < \infty$.

Then $\mathbb{E}[a(\boldsymbol{w}_i, \boldsymbol{\theta})]$ is continuous in $\boldsymbol{\theta}$, and:

$$\frac{1}{n}\sum_{i=1}^{n} a(\boldsymbol{w}_i, \boldsymbol{\theta}) \xrightarrow{u.p} \mathbb{E}[a(\boldsymbol{w}_i, \boldsymbol{\theta})] \text{ over } \Theta \tag{A-1}$$

A.2 Basic Consistency Theorem

Theorem (2.1) says that if:

1. **Identification**: $Q_0(\boldsymbol{\theta}_0) \equiv \mathbb{E}[a(\boldsymbol{w}_i, \boldsymbol{\theta})]$ is uniquely maximized at $\boldsymbol{\theta}_0$.
2. **Compactness**: Θ is compact.
3. **Continuity**: $Q_n(\boldsymbol{\theta})$ is continuous in $\boldsymbol{\theta}$ over Θ.
4. **Uniform convergence**: $Q_n(\boldsymbol{\theta}) \xrightarrow{u.p.} Q_0(\boldsymbol{\theta})\ \forall\ \boldsymbol{\theta}$ over Θ.

Then:

$$\hat{\boldsymbol{\theta}} \xrightarrow{p} \boldsymbol{\theta}_0 \tag{A-2}$$

A.3 Consistency with Concave Objective Functions

Theorem (2.7) says that if:

1. **Identification**: $Q_0(\boldsymbol{\theta}_0)$ uniquely maximized at $\boldsymbol{\theta}_0$.
2. **Convex parameter space**: $\boldsymbol{\theta}_0$ is an element of the interior of a convex set $\Theta \subset \mathbb{R}^K$.
3. **Concave objective function**: $Q_n(\boldsymbol{\theta})$ is concave in $\boldsymbol{\theta}$.
4. **Pointwise convergence in probability**: $Q_n(\boldsymbol{\theta}) \xrightarrow{p} Q_0(\boldsymbol{\theta})\ \forall\ \boldsymbol{\theta} \in \Theta$.

Then:

$$\hat{\boldsymbol{\theta}} \xrightarrow{p} \boldsymbol{\theta}_0 \tag{A-3}$$

A.4 Asymptotic Normality

Suppose that $\hat{\boldsymbol{\theta}} \xrightarrow{p} \boldsymbol{\theta}_0$. Theorem (3.1) says that if:

1. **Parameter space**: $\boldsymbol{\theta}_0 \in \text{int } \Theta$.
2. **Differentiability**: $Q_n(\boldsymbol{\theta})$ is twice differentiable (in a neighborhood on $\boldsymbol{\theta}_0$).
3. **Central limit theorem**: $\sqrt{n}\nabla_\theta Q_n(\boldsymbol{\theta}_0) \xrightarrow{d} N(0, J)$
4. **Continuity and uniform convergence of expected Hessian**: $H(\boldsymbol{\theta}) \equiv \mathbb{E}[\nabla_{\theta\theta} a(\boldsymbol{w}_i, \boldsymbol{\theta})]$ is continuous in $\boldsymbol{\theta}$ and $\nabla_{\theta\theta} Q_n(\boldsymbol{\theta}) \xrightarrow{u.p.} H(\boldsymbol{\theta})$ (in a neighborhood on $\boldsymbol{\theta}_0$).
5. **Limit Hessian invertible**: $H \equiv H(\boldsymbol{\theta}_0)$ is non-singular.

Then,

$$\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \xrightarrow{d} N(0, H^{-1}JH^{-1}) \tag{A-4}$$

A.5 Consistency of Maximum Likelihood Estimators

Suppose that $(y_i, \boldsymbol{x}_i)$ are continuous random variables with conditional pdf $f(y_i|\boldsymbol{x}_i; \boldsymbol{\theta}_0)$ where $f(y_i|\boldsymbol{x}_i; \boldsymbol{\theta}) > 0$. Let $\hat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}\in\Theta}{\text{argmax}}\, Q_n(\boldsymbol{\theta}) = \underset{\boldsymbol{\theta}\in\Theta}{\text{argmax}}\, \frac{1}{n}\sum_{i=1}^{n} f(y_i|\boldsymbol{x}_i; \boldsymbol{\theta})$ be an estimator of $\boldsymbol{\theta}_0$. Theorem (2.5) says that if:

1. **Identification**: $Q_0(\boldsymbol{\theta}) = \mathbb{E}[\ln[f(y_i, \boldsymbol{x}_i; \boldsymbol{\theta})]]$ is uniquely maximized at $\boldsymbol{\theta}_0 \in \Theta$.
2. **Compactness**: $\boldsymbol{\theta}_0 \in \Theta$, a compact set.
3. **Continuity**: $\ln[f(y_i, \boldsymbol{x}_i; \boldsymbol{\theta})]$ is continuous at each $\boldsymbol{\theta}_0 \in \Theta$.
4. **Uniform convergence**: $\mathbb{E}[\sup_{\boldsymbol{\theta}_0\in\Theta} |\ln[f(y_i, \boldsymbol{x}_i; \boldsymbol{\theta})]|] < \infty$. This condition impies uniform convergence, i.e., $Q_n(\boldsymbol{\theta}) \xrightarrow{u.p.} Q_0(\boldsymbol{\theta})$ over Θ and allows to interchange the differentiation and integration operators.

Then,

$$\hat{\boldsymbol{\theta}} \xrightarrow{p} \boldsymbol{\theta}_0 \tag{A-5}$$

A.6 Asymptotic Normality of MLE

Suppose that $\hat{\boldsymbol{\theta}} \xrightarrow{p} \boldsymbol{\theta}_0$. Theorem (3.3) says that if:

1. **Parameter space**: $\boldsymbol{\theta}_0 \in \text{int } \Theta$.
2. **Differentiability**: $\ln[f(y_i, \boldsymbol{x}_i; \boldsymbol{\theta})]$ has continuous 1st and 2nd derivatives and $f(y_i, \boldsymbol{x}_i; \boldsymbol{\theta}) > 0$.
3. **Central limit theorem**: $\sqrt{n}\nabla_\theta Q_n(\boldsymbol{\theta}_0) = \sqrt{n}\frac{1}{n}\sum_i \nabla_\theta \ln[f(y_i, \boldsymbol{x}_i; \boldsymbol{\theta}_0)] \xrightarrow{d} N(0, J)$, and $J = \mathbb{E}[(\nabla_\theta \ln[f(y_i, \boldsymbol{x}_i; \boldsymbol{\theta}_0)])(\nabla_\theta \ln[f(y_i, \boldsymbol{x}_i; \boldsymbol{\theta}_0)])'] = -H = -\mathbb{E}[\nabla_{\theta\theta} \ln[f(y_i, \boldsymbol{x}_i; \boldsymbol{\theta}_0)]]$
4. **Continuity and uniform convergence of expected Hessian**: $H(\boldsymbol{\theta}) \equiv \mathbb{E}[\nabla_{\theta\theta} \ln[f(y_i, \boldsymbol{x}_i; \boldsymbol{\theta}_0)]]$ is continuous in $\boldsymbol{\theta}$ and $\nabla_{\theta\theta} Q_n(\boldsymbol{\theta}) \xrightarrow{u.p.} H(\boldsymbol{\theta})$.
5. **Limit Hessian invertible**: $H \equiv H(\boldsymbol{\theta}_0)$ is non-singular.

Then,

$$\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \xrightarrow{d} N(0, H^{-1}JH^{-1}) = N(0, J^{-1}) \tag{A-6}$$

A.7 Consistency of GMM Estimators

Suppose that $W_n \xrightarrow{p} W$, a positive semi-definite matrix. Let $\boldsymbol{\theta}_0$ be the true value of the parameter of interests, and $\hat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}\in\Theta}{\operatorname{argmax}} Q_n(\boldsymbol{\theta}) = \underset{\boldsymbol{\theta}\in\Theta}{\operatorname{argmax}} -g(\boldsymbol{w}_i, \boldsymbol{\theta})' W_n g(\boldsymbol{w}_i, \boldsymbol{\theta})$ be an estimator for $\boldsymbol{\theta}_0$. Theorem (2.6) says that if:

1. **Identification**: $\mathbb{E}[g(\boldsymbol{w}_i, \boldsymbol{\theta})] = 0$ only if $\boldsymbol{\theta} = \boldsymbol{\theta}_0$
2. **Compactness** $\boldsymbol{\theta}_0$ in $\boldsymbol{\theta}_0$, a compact set.
3. **Continuity** $g(\boldsymbol{w}_i, \boldsymbol{\theta})$ is continuous at each $\boldsymbol{\theta}$ in Θ
4. **Uniform convergence** $\mathbb{E}[\underset{\boldsymbol{\theta}\in\Theta}{\sup}]||g(\boldsymbol{w}_i, \boldsymbol{\theta})|| < \infty$, so that $Q_n(\boldsymbol{\theta}) \xrightarrow{Q}_0 (\boldsymbol{\theta})$ over Θ.

Then,

$$\hat{\boldsymbol{\theta}} \xrightarrow{p} \boldsymbol{\theta}_0 \tag{A-7}$$

A.8 Asymptotic Normality of GMM

Suppose that the assumptions for consistency hold, so $\hat{\boldsymbol{\theta}} \xrightarrow{p} \boldsymbol{\theta}_0$. Theorem (3.2) says that if:

1. **Parameter Space** $\theta_0 \in$ in θ
2. **Differentiability** $g(\boldsymbol{w}_i, \boldsymbol{\theta})$ has continuous derivatives.
3. **CLT** $\sqrt{n} g_n(\boldsymbol{\theta}) \xrightarrow{d} N(0, \Omega)$, where $\Omega = \mathbb{E}[(g(\boldsymbol{w}_i, \boldsymbol{\theta}))(g(\boldsymbol{w}_i, \boldsymbol{\theta}))]$
4. **Uniform convergence** $G(\boldsymbol{\theta}) \equiv \mathbb{E}[\nabla_\theta g(\boldsymbol{w}_i, \boldsymbol{\theta})]$ is continuous in $\boldsymbol{\theta}$ and $\nabla_\theta g_n(\boldsymbol{\theta}) \xrightarrow{u.p} G(\boldsymbol{\theta})$
5. **Full rank** For $G \equiv G(\boldsymbol{\theta}_0)$, $G'WG$ is non-singular.

Then,

$$\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \xrightarrow{d} N(0, (G'WG)^{-1} G'W\Omega WG(G'WG)^{-1}) \tag{A-8}$$

Note that this result suggests that the optimal weighting matrix is Ω^{-1}, as it simplifies the asymptotic variance to $(G'\Omega^{-1}G)^{-1}$.