Econometrics

These slides follow very closely
William H. Green, Econometric Analysis, 6th Edition
All graphics and formula are taken out of this book.

Contents

15.2 The Method of Moments I

Idea:

Sample statistics such as the mean or the variance can be treated as simple descriptive measures. However, in general, sample statistics each have a counterpart in the population, for example, the correspondence between the sample mean and the population. The natural next step in the analysis is to use this analogy to justify using the sample 'moments' as estimators of these population parameters. It remains to establish whether this approach is a good way to use the sample data to infer the characteristics of the population.

15.2 The Method of Moments II

The basis of the method of moments is as follows:

- In random sampling, under generally benign assumptions, a **sample statistic will converge** in probability **to some constant**. For example, with i.i.d sampling $\bar{m}_2' = \frac{1}{n}\sum_{i=1}^{n} y_i^2$ converges to $\sigma_y^2 - \mu_y^2$.
- This constant will, in turn, be a function of the unknown parameters of the distribution.
- In order to estimate K parameters we compute K such statistics whose probability limits are known functions of the parameters.

15.2.1 Random Sampling and Estimating the Parameters of Distributions I

Consider i.i.d. random sampling from a distribution $f(y|\theta_1, \theta_2, ..., \theta_K)$ with finite moments up to $E[y^{2K}]$. The random sample consist of n observations $y_1, ..., y_n$.

$$\bar{m}_k \equiv \frac{1}{n}\sum_{i=1}^{n} y_i^k \qquad k\text{th uncentered moment}$$

$$E[\bar{m}_k] = E[y_i^k]$$

$$Var[\bar{m}_k] = \frac{1}{n} Var[y_i^k]$$

$$\text{plim } \bar{m}_k = E[y_i^k]$$

$$(\bar{m}_k - E[\bar{m}_k]) \longrightarrow^d N(0, Var[\bar{m}_k])$$

15.2.1 Random Sampling and Estimating the Parameters of Distributions II

Example 1: Method of Moments Estimator for $N(\mu, \sigma^2)$

$$\left.\begin{array}{l} \bar{m}_1 = (1/n)\sum_i y_i \\ \bar{m}_2 = (1/n)\sum_i y_i^2 \end{array}\right\} \begin{array}{l} \text{plim } \bar{m}_1 = \mu \\ \text{plim } \bar{m}_2 = E[y_i^2] = \sigma^2 + \mu^2 \end{array}$$

$$\Longrightarrow \hat{\mu} = \bar{m}_1 \text{ and } \hat{\sigma}^2 = \bar{m}_2 - \bar{m}_1^2$$

Note that $\hat{\sigma}^2$ is biased, although both estimators are consistent.

15.2 The Method of Moments I

Although the moments based on powers of y provide a natural source of information about the parameters, other functions of the data may also be useful. Let $m_k(\cdot)$ be a continuous and differentiable function not involving the sample size n, and let

$$\bar{m}_k = \frac{1}{n}\sum_{i=1}^{n} m_k(y_i), \qquad k = 1, ..., K.$$

These are also 'moments' of the data.

$$\text{plim } \bar{m}_k = E[m_k(y_i)] = \mu_k(\theta_1, ..., \theta_K)$$

15.2 The Method of Moments II

We assume that $\mu_k(\cdot)$ involves some of or all the parameters of the distribution. With K parameters to be estimated, the K moment equations,

$$\begin{aligned}
\bar{m}_1 - \mu_1(\theta_1, ..., \theta_K) &= 0, \\
\bar{m}_2 - \mu_2(\theta_1, ..., \theta_K) &= 0, \\
&\vdots \\
\bar{m}_K - \mu_K(\theta_1, ..., \theta_K) &= 0,
\end{aligned}$$

provide K equations in K unknowns, $\theta_1, ..., \theta_K$. If the equations are continuous and functionally independent, then method of moments estimators can be obtained by solving the system of equations for $\hat{\theta}_k = \hat{\theta}_k(\bar{m}_1, ..., \bar{m}_K)$.

15.2 Example 15.4 Mixtures of Normal Distributions I

$$\begin{aligned} f(y) &= \lambda N\left(\mu_1, \sigma_1^2\right) + (1-\lambda) N\left(\mu_2, \sigma_2^2\right) \\ &= \frac{\lambda}{\sqrt{2\pi\sigma_1^2}} e^{-1/2[(y-\mu_1)/\sigma_1]^2} + \frac{1-\lambda}{\sqrt{2\pi\sigma_2^2}} e^{-1/2[(y-\mu_2)/\sigma_2]^2} \end{aligned}$$

The sample mean and second through fifth central moments,

$$\bar{m}_k = \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^k, \qquad k = 2, 3, 4, 5,$$

provide five equations in five unknowns that can be solved for consistent estimators of the five parameters.

15.2 Example 15.4 II

$$
\begin{aligned}
\bar{m}_k &= \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^k \qquad \text{and} \qquad \mu_k = E[(y_i - \mu)^k] \\
\mu &= E[y_i] = \lambda\mu_1 + (1-\lambda)\mu_2 \\
\sigma^2 &= Var[y_i] = \lambda\sigma_1^2 + (1-\lambda)\sigma_2^2 + \lambda(1-\lambda)(\mu_1 - \mu_2)^2 \\
&\vdots
\end{aligned}
$$

$$
\begin{aligned}
\bar{m}_1 - \mu_1 &= 0 \rightarrow \bar{m}_1 - [\lambda\mu_1 + (1-\lambda)\mu_2] = 0 \\
\bar{m}_2 - \mu_2 &= 0 \rightarrow \bar{m}_2 - [\lambda\sigma_1^2 + (1-\lambda)\sigma_2^2 + \\
&\qquad\qquad + \lambda(1-\lambda)(\mu_1 - \mu_2)^2] = 0 \\
\bar{m}_3 - \mu_3 &= 0 \\
&\vdots \\
&\Rightarrow \quad \hat{\lambda}, \hat{\mu}_1, \hat{\mu}_2, \hat{\sigma}_1^2, \hat{\sigma}_2^2
\end{aligned}
$$

15.2 Example 15.4 III

Computation of the variance:

$$
\begin{aligned}
\sigma^2 &= Var[y] = \int [y - E[y]]^2 f(y)dy \\
&= \int \left[y^2 - 2yE[y] + E[y]^2\right] f(y)dy = \int y^2 f(y)dy - E[y]^2 \\
&= \int y^2 f_1(y)dy + \int y^2 f_2(y)dy - E[y]^2 \\
&= \lambda\sigma_1^2 + \lambda\mu_1^2 + (1-\lambda)\sigma_2^2 + (1-\lambda)\mu_2^2 - [\lambda\mu_1 + (1-\lambda)\mu_2]^2 \\
&= \lambda\sigma_1^2 + \lambda\mu_1^2 + (1-\lambda)\sigma_2^2 + (1-\lambda)\mu_2^2 - \lambda^2\mu_1^2 - 2\lambda(1-\lambda)\mu_1\mu_2 \\
&\quad -(1-\lambda)^2\mu_2^2 \\
&= \lambda\sigma_1^2 + (1-\lambda)\sigma_2^2 + \lambda\mu_1^2(1-\lambda) + (1-\lambda)\mu_2^2(1-1+\lambda) \\
&\quad -2\lambda(1-\lambda)\mu_1\mu_2 \\
&= \lambda\sigma_1^2 + (1-\lambda)\sigma_2^2 + \lambda(1-\lambda)(\mu_1-\mu_2)^2
\end{aligned}
$$

Example 15.5 Gamma Distribution

$$f(y) = \frac{\lambda^P}{\Gamma(P)} e^{-\lambda y} y^{P-1}, \qquad y \geq 0, P > 0, \lambda > 0$$

$$\mu = E[y] = \frac{P}{\lambda} \quad \text{and} \quad \sigma^2 = Var[y] = \frac{P}{\lambda^2}$$

Furthermore,

$$\bar{m}_1 = \frac{1}{n}\sum_{i=1}^{n} y_i \quad \text{and} \quad \bar{m}_2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2$$

$$\left.\begin{array}{l} \bar{m}_1 - \mu_1 = 0 \rightarrow \bar{m}_1 - \frac{P}{\lambda} = 0 \\ \bar{m}_2 - \mu_2 = 0 \rightarrow \bar{m}_2 - \frac{P}{\lambda^2} = 0 \end{array}\right\} \begin{array}{l} \hat{\lambda} = \bar{m}_1/\bar{m}_2 \\ \hat{P} = \bar{m}_1^2/\bar{m}_2 \end{array}$$

15.2.2 Properties of the Method of Moments Estimator

- In sampling from the normal distribution exact distribution is available
- Method of Moments Estimator $\hat{\theta}$ is consistent
- $\hat{\theta} \sim^{a} N\left(\theta_0, Est.Asy.Var[\hat{\theta}]\right)$
- Generally it is not an efficient estimator (exception random sampling from exponential families of distributions)

Definition 15.1 Exponential Family

An exponential (parametric) family of distributions is one whose log-likelihood is of the form

$$\ln L(\theta|data) = a(data) + b(\theta) + \sum_{k=1}^{K} c_k(data)s_k(\theta)$$

where $a(\cdot), b(\cdot), c_k(\cdot)$, and $s_k(\cdot)$ are functions. The members of the 'family' are distinguished by the different parameter values.

If the log-likelihood function is of this form, then the functions $c_k(\cdot)$ are called sufficient statistics. Method of moments estimators can be functions of them. In this case the method of moments estimators will also be the maximum likelihood estimators, so, they will be efficient at least asymptotically.

The Gamma distribution is a member of the exponential family

$$
\begin{aligned}
f(y) &= \frac{\lambda^P}{\Gamma(P)} e^{-\lambda y} y^{P-1}, \qquad y \geq 0, P > 0, \lambda > 0 \\
\frac{1}{n} LL &= [P \ln \lambda - \ln \Gamma(P)] - \lambda \frac{1}{n} \sum_{i=1}^{n} y_i + (P-1) \frac{1}{n} \sum_{i=1}^{n} \ln y_i
\end{aligned}
$$

Two sufficient statistics: $\frac{1}{n}\sum_{i=1}^{n} y_i$ and $\frac{1}{n}\sum_{i=1}^{n} lny_i$. The method of moments estimators based on these sufficient statistics would be the maximum likelihood estimators and therefore efficient under normality.

More moment equations than parameters

For example for the Gamma distribution we have also

$$plim \frac{1}{n} \sum_{i=1}^{n} \begin{pmatrix} y_i \\ y_i^2 \\ \ln y_i \\ 1/y_i \end{pmatrix} = \begin{pmatrix} P/\lambda \\ P(P+1)/\lambda^2 \\ \Psi(P) - \ln(\lambda) \\ \lambda/(P-1) \end{pmatrix},$$

where $\Psi(P) = dln\Gamma(P)/dP$. Any two of these can be used to estimate λ and P.

15.3 Minimum Distance Estimation I

The preceding analysis has considered **exactly identified cases** (K parameters and K moments).

How should we proceed if we have more moments than we need?

A minimum distance estimator (MDE) is defined as follows:
Let $\bar{m}_{n,l}$ denote a sample statistic based on n observations such that plim $\bar{m}_{n,l} = g_l(\theta_0), l = 1, ..., L$, where θ_0 is a vector of $K \leq L$ parameters to be estimated. Arrange theses moments and functions in $L \times 1$ vectors $\bar{m}_n$ and $g(\theta_0)$ and further assume that the statistics are jointly asymptotically normally distributed with plim $\bar{m}_n = g(\theta_0)$ and $Asy.Var[\bar{m}_n] = (1/n)\Phi$. Consequently,

$$\hat{\theta}_{MDE} = \text{argmin}_\theta \, [q = [\bar{m}_n - g(\theta)]' W [\bar{m}_n - g(\theta)]]$$

for a positive definite weighting matrix, W.

15.3 Minimum Distance Estimation II

Different choices of W will produce different estimators, but the estimator has the following properties for any W:

Under the assumption that $\sqrt{n}[\bar{m}_n - g(\theta_0)] \longrightarrow^d N(0, \Phi)$, the **asymptotic properties of the minimum distance estimator** are as follows:

- plim $\hat{\theta}_{MDE} = \theta_0$
- $Asy.Var[\hat{\theta}_{MDE}] = \frac{1}{n}[\Gamma_0' W \Gamma_0]^{-1}[\Gamma_0' W \Phi W \Gamma_0][\Gamma_0' W \Gamma_0]^{-1}$, where $\Gamma_0 = \Gamma(\theta_0) = \text{plim } G(\hat{\theta}_{MDE}) = \text{plim } \frac{\partial g(\hat{\theta}_{MDE})}{\partial \hat{\theta}'_{MDE}}$
- Optimal weighting matrix $W^* = \Phi^{-1}$
- $\hat{\theta}_{MDE} \longrightarrow^d N(0, Asy.Var[\hat{\theta}_{MDE}])$

15.4.1 Estimation Based on Orthogonality Conditions

Consider the **least squares estimator** of the parameters in the classical regression model. An important assumption of the model is

$$E[x_i \varepsilon_i] = E[x_i(y_i - x_i'\beta)] = 0.$$

The sample analog is

$$\frac{1}{n}\sum_{i=1}^{n} x_i e_i = \frac{1}{n}\sum_{i=1}^{n} x_i(y_i - x_i'\hat{\beta}) = 0$$

The estimator of β is the one that satisfies these moment equations, which are just the normal equations for the least squares estimator. So, the OLS estimator is a method of moments estimator.

15.4.2 Generalizing the Method of Moments

Notation: $m_l(y_i, x_i, z_i, \theta) \equiv m_{il}(\theta), \bar{m}_{nl}(\theta_0) = \frac{1}{n}\sum_{i=1}^{n} m_{il}(\theta_0), \theta = (\theta_1, ..., \theta_K), \theta_0$

Population orthogonality conditions: $E[m_{il}(\theta_0)] = 0$
Corresponding sample moment equations: $E[\bar{m}_l(y, X, Z, \theta_0)] = 0$

Assume $L > K$, we need an objective function:

$$
\begin{aligned}
q &= \sum_{l=1}^{L} \bar{m}_l^2 = \bar{m}(\theta)'\bar{m}(\theta) \\
q &= \bar{m}(\theta)' W_n \bar{m}(\theta) \\
&\qquad W = (Asy.Var[\sqrt{n}\bar{m}])^{-1} = \Phi^{-1} \\
\hat{\theta}_{GMM} &= \text{argmin}_\theta \bar{m}(\theta)'\Phi^{-1}\bar{m}(\theta)
\end{aligned}
$$

15.4.3 Properties of the GMM Estimator

Theorem 15.2 Asymptotic Distribution of the GMM Estimator

Under the assumptions of convergence of the empirical moments, identification and asymptotic distribution of the empirical moments the GMM estimator has the following properties

- The GMM estimator of θ is consistent.
- Asymptotic covariance matrix $V_{GMM} = \frac{1}{n}(\Gamma'\Phi^{-1}\Gamma)^{-1}$, where Γ is the matrix of derivatives with jth row equal to $\Gamma^j = \text{plim } \frac{\partial \bar{m}_j(\theta)}{\partial \theta'}$.
- The estimator is asymptotically normally distributed $\hat{\theta}_{GMM} \sim^a N(\theta_0, V_{GMM})$.

Example

Assumption 1: Convergence of the Empirical Moments

The data generating process is assumed to meet the conditions for a law of large numbers to apply, so that we may assume that the empirical moments converge in probability to their expectation. What is required for this assumption is that

$$\bar{m}_n(\theta_o) = \frac{1}{n}\sum_{i=1}^{n} m_i(\theta_0) \longrightarrow^{p} 0.$$

The empirical moments are assumed to be continuous and continuously differentiable functions of the parameters.
Consequently, we will be able to assume that the derivatives of the moments

$$\bar{G}_n(\theta_0) = \frac{\partial \bar{m}_n(\theta_0)}{\partial \theta_0'} = \frac{1}{n}\sum_{i=1}^{n} \frac{\partial m_{i,n}(\theta_0)}{\partial \theta_0',}$$

converge to a probability limit, say, plim $\bar{G}_n(\theta_0) = \bar{G}(\theta_0)$.

◂ Return Button

Assumption 2: Identification

For any $n \geq K$, if θ_1 and θ_2 are two different parameter vectors, then there exist data sets such that $\bar{m}_n(\theta_1) \neq \bar{m}_n(\theta_2)$. Formally, identification is defined to imply that the probability limit of the GMM criterion function is uniquely minimized at the true parameters, θ_0.

The identification condition has three important implications:

1. Order condition: $L \geq K$.
2. Rank condition. The $L \times K$ matrix of derivatives, $\bar{G}_n(\theta_0)$, will have row rank equal to K.
3. Uniqueness. With the continuity assumption, the identification assumption implies that the parameter vector that satisfies the population moment condition is unique.

◂ Return Button

Assumption 3: Asymptotic Distribution of Empirical Moments

We assume that the empirical moments obey a central limit theorem. This assumes that the moments have a finite asymptotic covariance matrix, $(1/n)\Phi$, so that

$$\sqrt{n}\bar{m}_n(\theta_0) \longrightarrow^d N(0, \Phi).$$

◂ Return Button

Example 15.7 GMM Estimation of the Parameters of a Gamma Distribution I

For the gamma distribution
$f(y) = \frac{\lambda^P}{\Gamma(P)} e^{-\lambda y} y^{P-1}, Y \geq 0, P > 0, \lambda > 0$ we have for example the following four moment equations ($\Psi(P) = d \ln \Gamma(P)/dP$)

$$E \begin{bmatrix} y_i - P/\lambda \\ y_i^2 - P(P+1)/\lambda^2 \\ \ln y_i - \Psi(P) + \ln \lambda \\ 1/y_i - \lambda/(P-1) \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}$$

The sample means of these will provide the moment equations for estimation. Let $y_1 = y, y_2 = y^2, y_3 = \ln y$, and $y_4 = 1/y$. Then

$$\begin{aligned} \bar{m}_1(P,\lambda) &= \frac{1}{n}\sum_{i=1}^{n}(y_{i1} - P/\lambda) = \frac{1}{n}\sum_{i=1}^{n}(y_{i1} - \mu_1(P/\lambda)) \\ &= \bar{y}_1 - \mu_1(P,\lambda) \\ \bar{m}_2(P,\lambda) &= ..., \bar{m}_3(P,\lambda) = ..., \bar{m}_4(P,\lambda) = ..., \end{aligned}$$

Example 15.7 II

For our initial set of estimators, we will use OLS. The optimization problem is

$$\text{Minimize}_{P,\lambda} \sum_{i=1}^{4} \bar{m}_l(P,\lambda)^2 = \sum_{i=1}^{4} [(\bar{y}_l - \mu_l(P,\lambda)]^2 = \bar{m}(P,\lambda)'\bar{m}(P,\lambda).$$

This estimator will be the minimum distance estimator with $W = I$. This is a nonlinear optimization problem (initial values are the ML estimates $\hat{P}_{ML} = 2.4106, \hat{\lambda}_{ML} = 0.0771$), for which we got $\hat{P} = 2.0583$ and $\hat{\lambda} = 0.0658$. Using these we get our estimate for W:

$$\hat{\Phi} = \frac{1}{20} \sum_{i=1}^{20} \begin{pmatrix} y_{i1} - \hat{P}/\hat{\lambda} \\ y_{i2} - \hat{P}(\hat{P}+1)/\hat{\lambda}^2 \\ y_{i3} - \Psi(\hat{P}) + \ln(\hat{\lambda}) \\ y_{i4} - \hat{\lambda}/(\hat{P}-1) \end{pmatrix} \begin{pmatrix} y_{i1} - \hat{P}/\hat{\lambda} \\ y_{i2} - \hat{P}(\hat{P}+1)/\hat{\lambda}^2 \\ y_{i3} - \Psi(\hat{P}) + \ln(\hat{\lambda}) \\ y_{i4} - \hat{\lambda}/(\hat{P}-1) \end{pmatrix}'$$

Example 15.7 III

The GMM estimator is now obtained by minimizing

$$q = \bar{m}(P, \lambda)' \hat{\Phi}^{-1} \bar{m}(P, \lambda).$$

The two estimates are $\hat{P}_{GMM} = 3.3589$, and $\hat{\lambda}_{GMM} = 0.1245$.
To obtain an asymptotic covariance matrix for the two estimates, we first recompute $\hat{\Phi}$. Then we require the derivatives matrix,

$$\begin{aligned}
\Gamma'(\theta) &= \begin{pmatrix} \partial\bar{m}_1/\partial P & \partial\bar{m}_2/\partial P & \partial\bar{m}_3/\partial P & \partial\bar{m}_4/\partial P \\ \partial\bar{m}_1/\partial\lambda & \partial\bar{m}_2/\partial\lambda & \partial\bar{m}_3/\partial\lambda & \partial\bar{m}_4/\partial\lambda \end{pmatrix} \\
&= \begin{pmatrix} -1/\lambda & -(2P+1)/\lambda^2 & -\Psi'(P) & \lambda/(P-1)^2 \\ P/\lambda^2 & 2P(P+1)/\lambda^3 & 1/\lambda & -1/(P-1) \end{pmatrix}
\end{aligned}$$

Finally, we get the estimate $\bar{G}(\hat{P}, \hat{\lambda})$ of $\Gamma(\theta)$ and

$$V_{GMM} = \frac{1}{20}[\bar{G}(\hat{P}, \hat{\lambda})' \hat{\Phi}^{-1} \bar{G}(\hat{P}, \hat{\lambda})]^{-1}.$$

Example 15.5 Testing Hypothesis in the GMM Framework
Testing the Validity of the Moment Restrictions

Criterion for GMM estimation $q = \bar{m}(\theta)' W \bar{m}(\theta)$. If $L = K$ $\bar{m}(\theta) = 0$ W irrelevant to the solution.
If $L > K$ substantive restrictions, Wald statistic with H_0 : validity of the model:

$$
\begin{aligned}
nq &= \left(\sqrt{n}\bar{m}(\hat{\theta})\right)' \left(Est.Asy.Var[\sqrt{n}\bar{m}(\hat{\theta})]\right)^{-1} \left(\sqrt{n}\bar{m}(\hat{\theta})\right) \\
nq &\longrightarrow^{d} \chi^2(L-K)
\end{aligned}
$$

This is a specification test, not a test of parametric restrictions. However, there is a symmetry between the moment restrictions and restrictions on the parameters: $(nq_R - nq) \rightarrow^{d} \chi^2(J)$.

Example 15.5.2 GMM Counterparts to the Wald, LM, and LR Tests

$$H_0 : r(\theta) = 0$$

Define c_1 the GMM estimates of θ without the restrictions, and c_0 the restricted GMM estimates; J possibly nonlinear restrictions on K parameters.

$$\begin{aligned}
LR_{GMM} &= nq_0 - nq_1 \longrightarrow^d \chi^2(J) \\
Wald &= r(c_1)'(Est.Asy.Var[r(c_1)])^{-1} r(c_1) \\
&= r(c_1)' \left(R_1 \frac{1}{n} \bar{G}(c_1)' \Phi^{-1} \bar{G}(c_1) R_1 \right) r(c_1), \\
&\qquad \text{where } R_1 = \partial r(c_1)/\partial c_1' \\
LM_{GMM} &= g_1(c_0)' \left(Est.Asy.Var[g_1(c_0)]\right)^{-1} g_1(c_0), \\
&\qquad \text{where } g_1(c_0) = \partial q/\partial c_0 \\
&= n \left[\bar{m}(c_0)' \hat{\Phi}_1^{-1} \bar{G}_1(c_0)\right] \left[\bar{G}_1(c_0)' \hat{\Phi}_1^{-1} \bar{G}_1(c_0)\right]^{-1} \cdot \\
&\qquad \left[\bar{G}_1(c_0)' \hat{\Phi}_1^{-1} \bar{m}(c_0)\right]
\end{aligned}$$

Example 15.6.1 Single Equation Linear Models I

$$\begin{aligned} y_i &= x_i'\beta + \varepsilon_i \\ E[z_i\varepsilon_i] &= 0, \end{aligned}$$

for K variables in x_i and for some set of L instrumental variables, z_i, where $L \geq K$.
In the generalized regression model $z_i = x_i$, and in the classical regression framework $\Omega = I$.

Assumption about the DGP:

- Classical regression: $Var[\varepsilon_i|X, Z] = \sigma^2$.
- Heteroscedasticity: $Var[\varepsilon_i|X, Z] = \sigma_i^2$.
- Generalized model: $Cov[\varepsilon_t, \varepsilon_s|X, Z] = \sigma^2\omega_{ts}$,

where Z and X are the $n \times L$ and $n \times K$ observed data matrices.
No specific distribution is assumed for the disturbances.

Example 15.6.1 Single Equation Linear Models II

The assumption $E[z_i\varepsilon_i] = 0$ implies the following orthogonality condition $E[z_i(y_i - x_i'\beta)] = 0$. This further implies the

Population moment equation

$$E\left[\frac{1}{n}\sum_{i=1}^{n} z_i(y_i - x_i'\beta)\right] \quad = \quad E[\bar{m}(\beta)] = 0$$

Sample counterpart

$$\frac{1}{n}\sum_{i=1}^{n} z_i(y_i - x_i'\hat{\beta}) \quad = \quad \frac{1}{n}\sum_{i=1}^{n} m_i(\hat{\beta}) = \bar{m}(\hat{\beta}) = 0$$

1. **Underidentified** $L < K$: Impossible to find a solution.

Example 15.6.1 Single Equation Linear Models III

2. **Exactly identified** $L = K$:

$$\begin{aligned} \bar{m}(\hat{\beta}) &= \frac{1}{n} Z'y - \frac{1}{n} Z'X\hat{\beta} \\ \hat{\beta} &= (Z'X)^{-1} Z'y \end{aligned}$$

3. **Overidentified** $L > K$: There is no unique solution to the equations system $\bar{m}(\hat{\theta}) = 0$.

$$\begin{aligned} \text{Min}_{\beta} q &= \bar{m}(\beta)' \bar{m}(\beta) \\ \frac{\partial q}{\partial \beta} &= 2 \left(\frac{\partial \bar{m}(\hat{\beta})'}{\partial \beta} \right) \bar{m}(\hat{\beta}) = 2 \bar{G}(\hat{\beta})' \bar{m}(\hat{\beta}) \\ &= 2 \left(\frac{1}{n} X'Z \right) \left(\frac{1}{n} Z'y - \frac{1}{n} Z'X\hat{\beta} \right) = 0 \\ \hat{\beta} &= [(X'Z)(Z'X)]^{-1} (X'Z)(Z'y) \end{aligned}$$

Example 15.6.1 Single Equation Linear Models IV

It remains to establish consistency and to obtain the asymptotic distribution for the estimator, therefore we need to check/formulate the assumptions 15.1 till 15.3.

1. **Convergence of moments**: plim $\bar{m}(\beta) = 0$.
2. **Identification**: The order condition ($L \geq K$) is already assumed. We must state the rank condition
   The $L \times K$ matrix
   $\Gamma(\beta) = E[\bar{G}(\beta)] = \text{plim } \bar{G}(\beta) = \text{plim } \frac{\partial \bar{m}}{\partial \beta'} = \text{plim } \frac{1}{n} \sum_{i=1}^{n} \frac{\partial m_i}{\partial \beta'}$
   must have row rank equal to K.
3. **Limiting Normal Distribution for the Sample Moments**. The population moment obeys a central limit theorem or some similar variant.

Under these assumptions we obtain the desired properties of our GMM estimator $\hat{\beta}$.

Example 15.6.1 Single Equation Linear Models V

$$
\begin{aligned}
Asy.Var[\hat{\beta}] &= \frac{1}{n}[\Gamma'\Gamma]^{-1}\Gamma'\left(Asy.Var[\sqrt{n}\bar{m}(\beta)]\right)\Gamma[\Gamma'\Gamma]^{-1} \\
\bar{m}(\beta) &= (1/n)(Z'y - Z'X\beta) \\
\bar{G}(\beta) &= (1/n)Z'X \\
\Gamma(\beta) &= Q_{ZX} \\
Asy.Var[\sqrt{n}\bar{m}(\beta)] &= V = \frac{1}{n}Var[\sum_{i=1}^{n} z_i\varepsilon_i] = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n}\sigma^2\omega_{ij}z_i z_j' \\
&= \sigma^2\frac{Z'\Omega Z}{n} \\
Est.Asy.Var[\hat{\beta}] &= \frac{1}{n}[\bar{G}(\hat{\beta})'\bar{G}(\hat{\beta})]^{-1}\bar{G}(\hat{\beta})'\hat{V}\bar{G}(\hat{\beta})[\bar{G}(\hat{\beta})'\bar{G}(\hat{\beta})]^{-1} \\
&= n[(X'Z)(Z'X)]^{-1}(X'Z)\hat{V}(Z'X)[(X'Z)(Z'X)]^{-1}.
\end{aligned}
$$

Example 15.6.1 Single Equation Linear Models VI

- Classical regression:

$$\hat{V} = \frac{(e'e/n)}{n} \sum_{i=1}^{n} z_i z_i' = \frac{e'e/n}{n} Z'Z.$$

- Heteroscedastic regression: $\hat{V} = \frac{1}{n} \sum_{i=1}^{n} e_i^2 z_i z_i'$.
- Generalized regression:

$$\begin{aligned} \hat{V} &= \frac{1}{n} \left[\sum_{i=1}^{n} e_i^2 z_i z_i' + \right. \\ &+ \left. \sum_{l=1}^{p} \left(1 - \frac{l}{p+1} \right) \sum_{t=l+1}^{n} e_t e_{t-l} (z_z z_{t-l}' + z_{t-l} z_t') \right]. \end{aligned}$$

- If it is known only that the terms in $\bar{m}(\beta)$ are uncorrelated, then $\hat{V} = \frac{1}{n} \sum_{i=1}^{n} m_i(\hat{\beta}) m_i(\hat{\beta})'$.

Example 15.6.1 Single Equation Linear Models VII

Let's go a step back:

$\text{Min}_\beta\ q = \bar{m}(\beta)' W \bar{m}(\beta)$, where W is any positive matrix.

- **Exactly identified case** $L = K$: W is irrelevant to the solution.
- **Overidentified case**: In this case the 'optimal' weighting matrix, that is, the W that produces the most efficient estimator, is $W = V^{-1}$. Consequently,

$$\begin{aligned} \hat{\beta}_{GMM} &= [(X'Z)\hat{V}^{-1}(Z'X)]^{-1}(X'Z)\hat{V}^{-1}(Z'y) \\ Est.Asy.Var[\hat{\beta}_{GMM}] &= \frac{1}{n}[\bar{G}'\hat{V}^{-1}\bar{G}]^{-1} = [(X'Z)\hat{V}^{-1}(Z'X)]^{-1} \end{aligned}$$

Example 15.6.1 Single Equation Linear Models VIII

Practical Implementation:

$$\text{Min}_\beta \; q = \bar{m}(\beta)' V^{-1} \bar{m}(\beta)$$

The process of GMM estimation will have to proceed in two steps:

Step 1: Use $W = I$ to obtain a consistent estimator of β. Then estimate V, e.g. in the heteroscedastic case by $\hat{V} = \frac{1}{n} \sum_{i=1}^{n} e_i^2 z_i z_i'$.

Step 2: Use $W = V^{-1}$ to compute the GMM estimator.

Note: Continuously updated GMM estimator.

Contents

16.2 The Likelihood Function and Identification of the Parameters I

The probability function, or pdf, for a random variable, y, conditioned on a set of parameters, θ, is denoted $f(y|\theta)$.This function identifies the data-generating process.

The joint density of n independent and identically distributed observations from this process is the product of the individual densities:

$$f(y_1, ..., y_n|\theta) = \prod_{i=1}^{n} f(y_i|\theta) = L(\theta|y).$$

This joint density is the likelihood function, defined as a function of the unknown parameter vector, θ, where y is used to indicate the collection of sample data.

It is usually simpler to work with the log of the likelihood function:

$$\ln L(\theta|y) = \sum_{i=1}^{n} \ln f(y_i|\theta).$$

16.2 The Likelihood Function and Identification of the Parameters II

It will usually be necessary to generalize the concept of the likelihood function to allow the density to depend on other conditioning variables.

Consider as example the classical linear regression model: Suppose the disturbances are normally distributed. Then, conditioned on its specific x_i, y_i is normally distributed with mean $\mu_i = x_i'\beta$ and variance σ^2. That means the observed random variables are not i.i.d., they have different means. Nonetheless the observations are independent, and consequently

$$\begin{aligned} \ln L(\theta|y, X) &= \sum_{i=1}^{n} \ln f(y_i|x_i, \theta) \\ &= -\frac{1}{2}\sum_{i=1}^{n} \left(\ln \sigma^2 + \ln(2\pi) + (y_i - x_i'\beta)^2/\sigma^2\right), \end{aligned}$$

where X is the $n \times K$ matrix of data with ith row equal to x_i'.

16.2 Identification

The parameter vector θ is identified (estimable) if for any other parameter vector, $\theta^* \neq \theta$, for some data y, $L(\theta^*|y) \neq L(\theta|y)$.

Example 16.1 **Multicollinearity in the classical regression model**
Suppose that there is a nonzero vector a such that $x_i'a = 0$ for every x_i (perfect multicollinearity). Then there is another parameter vector $\gamma = \beta + a$ such that $x_i'\beta = x_i'\gamma$ for every x_i. Consider

$$\ln L(\theta|y, X) = -\frac{1}{2}\sum_{i=1}^{n}\left(\ln\sigma^2 + \ln(2\pi) + (y_i - x_i'\beta)^2/\sigma^2\right).$$

If this perfect multicollinearity is the case, then the log-likelihood function is the same whether it is evaluated at β or at γ. As such it is not possible to consider estimation of β in this model because β cannot be distinguished from γ.

16.3 Efficient Estimation: The Principle of Maximum Likelihood I

The logic of the technique is easily demonstrated in the setting of a discrete distribution. Consider a random sample of the following 10 observations from a Poisson distribution: $5, 0, 1, 1, 0, 3, 2, 3, 4,$ and 1. The density of each observation is

$$f(y_i|\theta) = \frac{e^{-\theta}\theta^{y_i}}{y_i!}$$

$$f(y_1, y_2, ..., y_{10}|\theta) = \prod_{i=1}^{10} f(y_i|\theta) = \frac{e^{-10\theta}\theta^{\sum_{i=1}^{10} y_i}}{\prod_{i=1}^{10} y_i!} = \frac{e^{-10\theta}\theta^{20}}{207,360}$$

The last result gives the probability of observing this particular sample.

What value of θ would make this sample most probable?

16.3 Efficient Estimation: The Principle of Maximum Likelihood II

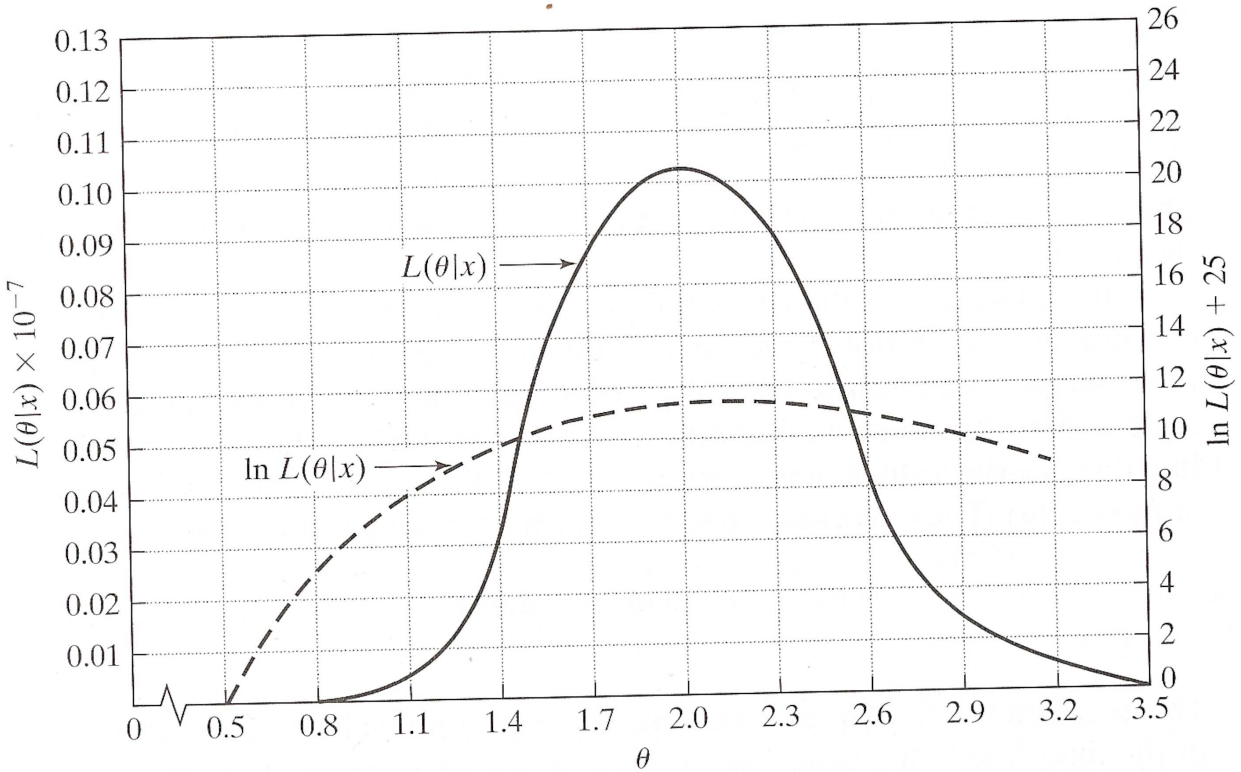

FIGURE 16.1 Likelihood and Log-Likelihood Functions for a Poisson Distribution.

16.3 Efficient Estimation: The Principle of Maximum Likelihood III

$$
\begin{aligned}
f(y_1, y_2, ..., y_n|\theta) &= \frac{e^{-n\theta}\theta^{\sum_{i=1}^{n} y_i}}{\prod_{i=1}^{n} y_i!} \\
\ln L(\theta|y) &= -n\theta + \ln\theta \sum_{i=1}^{n} y_i - \sum_{i=1}^{n} \ln(y_i!) \\
\frac{\partial \ln L(\theta|y)}{\partial\theta} &= -n + \frac{1}{\theta}\sum_{i=1}^{n} y_i = 0 \Rightarrow \hat{\theta}_{MLE} = \bar{y}_n \\
\frac{\partial^2 \ln L(\theta|y)}{\partial\theta^2} &= -\frac{1}{\theta^2}\sum_{i=1}^{n} y_i
\end{aligned}
$$

For the given sample we get $\hat{\theta}_{MLE} = 2$ and $\frac{\partial^2 \ln L(\theta|y)}{\partial\theta^2} = -\frac{20}{\theta^2} < 0$.

Note: Reference to the probability of observing the given sample in a continuous distribution.

MLEs for the normal distribution

In sampling from a normal distribution with mean μ and variance σ^2, the log-likelihood function is given as

$$\ln L(\mu, \sigma^2) = -\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln(\sigma^2) - \frac{1}{2}\sum_{i=1}^{n}\frac{(y_i - \mu)^2}{\sigma^2}$$

Compute the MLEs, $\hat{\mu}_{MLE}$ and $\hat{\sigma}^2$.

16.4 Properties of Maximum Likelihood Estimators

Notation: $\hat{\theta}$... maximum likelihood estimator; θ_0... true value of the parameter vector; θ... another possible value of the parameter vector, not the MLE and not necessarily the true values. Expectation based on the true values of the parameters is denoted as $E_0[\cdot]$.

Under regularity, the MLE has the following asymptotic properties:

1. Consistency: plim $\hat{\theta} = \theta_0$.
2. Asymptotic normality: $\hat{\theta} \sim^a N\left(\theta_0, I(\theta_0)^{-1}\right)$, where $I(\theta_0) = -E_0\left[\frac{\partial^2 \ln L}{\partial \theta_0 \partial \theta_0'}\right]$.
3. Asymptotic efficiency: $\hat{\theta}$ is asymptotically efficient and achieves the **Cramer-Rao lower bound** for consistent estimators.
4. Invariance: The maximum likelihood estimator of $\gamma_0 = c(\theta_0)$ is $c(\hat{\theta})$ if $c(\theta_0)$ is a continuous and continuously differentiable function.

16.4.1 Regularity Conditions

We assume that $(y_1, ..., y_n)$ is a random sample from the population with density function $f(y_i|\theta_0)$ and that the following regularity conditions hold.

Definition of Regularity conditions

1. R1. The first three derivatives of $\ln f(y_i|\theta)$ with respect to θ are continuous and finite for almost all y_i and for all θ. This condition ensures the existence of a certain Taylor series approximation and the finite variance of the derivatives of $\ln L$.
2. R2. The conditions necessary to obtain the expectations of the first and second derivatives of $\ln f(y_i|\theta)$ are met.
3. R3. For all values of $\theta, |\partial^3 \ln f(y_i|\theta)/\partial\theta_j\partial\theta_k\partial\theta_l|$ is less than a function that has a finite expectation. This condition will allow us to truncate the Taylor series.

16.4.2 Properties of Regular Densities I

Densities that are regular by the above Definition have three properties that are used in establishing the properties of maximum likelihood estimators.

Theorem: Moments of the Derivatives of the Log-Likelihood

1. D1. $\ln f(y_i|\theta)$, $g_i = \partial \ln f(y_i|\theta)/\partial\theta$, and $H_i = \partial^2 \ln f(y_i|\theta)/\partial\theta\partial\theta'$, $i = 1, ..., n$ are all random samples of random variables. This statement follows from our assumption of random sampling. The notation $g_i(\theta_0)$ and $H_i(\theta_0)$ indicates the derivative evaluated at θ_0.
2. D2. $E_0[g_i(\theta_0)] = 0$.
3. D3. $Var[g_i(\theta_0)] = -E[H_i(\theta_0)]$.

16.4.2 Properties of Regular Densities II

$$\int_{A(\theta_0)}^{B(\theta_0)} f(y_i|\theta_0)dy_i = 1$$

$$\frac{\partial \int_{A(\theta_0)}^{B(\theta_0)} f(y_i|\theta_0)dy_i}{\partial \theta_0} = \int_{A(\theta_0)}^{B(\theta_0)} \frac{\partial f(y_i|\theta_0)}{\partial \theta_0} dy_i + f(B(\theta_0)|\theta_0)\frac{\partial B(\theta_0)}{\partial \theta_0}$$
$$-f(A(\theta_0)|\theta_0)\frac{\partial A(\theta_0)}{\partial \theta_0} = 0$$

Necessary conditions: $lim_{y_i \to A(\theta_0)} f(y_i|\theta_0) = lim_{y_i \to B(\theta_0)} f(y_i|\theta_0) = 0$
Sufficient conditions: $\frac{\partial A(\theta_0)}{\partial \theta_0} = \frac{\partial B(\theta_0)}{\partial \theta_0} = 0$ or the density is zero at the terminal points. (R2) Hence,

$$\frac{\partial \int f(y_i|\theta_0)dy_i}{\partial \theta_0} = \int \frac{\partial f(y_i|\theta_0)}{\partial \theta_0} dy_i = \int \frac{\partial \ln f(y_i|\theta_0)}{\partial \theta_0} f(y_i|\theta_0)dy_i$$
$$= E_0\left[\frac{\partial \ln f(y_i|\theta_0)}{\partial \theta_0}\right] = 0. \qquad \text{This proves D2.}$$

16.4.2 Properties of Regular Densities III

$$\int \left[\frac{\partial^2 \ln f(y_i|\theta_0)}{\partial \theta_0 \partial \theta_0'} f(y_i|\theta_0) + \frac{\partial \ln f(y_i|\theta_0)}{\partial \theta_0} \frac{\partial f(y_i|\theta_0)}{\partial \theta_0'} \right] dy_i = 0$$

$$\frac{\partial f(y_i|\theta_0)}{\partial \theta_0'} = f(y_i|\theta_0) \frac{\partial \ln f(y_i|\theta_0)}{\partial \theta_0'}$$

$$-\int \left[\frac{\partial^2 \ln f(y_i|\theta_0)}{\partial \theta_0 \partial \theta_0'} \right] f(y_i|\theta_0) dy_i =$$

$$\int \left[\frac{\partial \ln f(y_i|\theta_0)}{\partial \theta_0} \frac{\partial \ln f(y_i|\theta_0)}{\partial \theta_0'} \right] f(y_i|\theta_0) dy_i$$

$$-E\left[\frac{\partial^2 \ln f(y_i|\theta_0)}{\partial \theta_0 \partial \theta_0'} \right] = E\left[\left(\frac{\partial \ln f(y_i|\theta_0)}{\partial \theta_0} \right) \left(\frac{\partial \ln f(y_i|\theta_0)}{\partial \theta_0'} \right) \right]$$

$$= Var\left[\frac{\partial \ln f(y_i|\theta_0)}{\partial \theta_0} \right]$$

Proves D3.

16.4.3 The Likelihood Equation

The log-likelihood function is

$$\ln L(\theta|y) = \sum_{i=1}^{n} \ln f(y_i|\theta)$$

The first derivative vector, or score vector, is

$$g = \frac{\partial \ln L(\theta|y)}{\partial \theta} = \sum_{i=1}^{n} \frac{\partial \ln f(y_i|\theta)}{\partial \theta} = \sum_{i=1}^{n} g_i.$$

Because we are just adding terms, it follows from D1 and D2 that at θ_0,

$$E_0 \left[\frac{\partial \ln L(\theta_0|y)}{\partial \theta_0} \right] = E[g_0] = 0.$$

which is the likelihood equation mentioned earlier.

16.4.4 The Information Matrix Equality I

The Hessian of the log-likelihood is

$$H = \frac{\partial^2 \ln L(\theta|y)}{\partial\theta_0\partial\theta_0'} = \sum_{i=1}^{n} \frac{\partial^2 \ln f(y_i|\theta)}{\partial\theta_0\partial\theta_0'} = \sum_{i=1}^{n} H_i.$$

Evaluating once again at θ_0, by taking

$$E_0[g_0 g_0'] = E_0\left[\sum_{i=1}^{n}\sum_{j=1}^{n} g_{0i} g_{0j}'\right],$$

and, because of D1, dropping terms with unequal subscripts we obtain

$$E_0[g_0 g_0'] \quad = \quad E_0\left[\sum_{i=1}^{n} g_{0i} g_{0i}'\right] = E_0\left[\sum_{i=1}^{n}(-H_{0i})\right] = -E_0[H_0],$$

16.4.4 The Information Matrix Equality II

$$
\begin{aligned}
Var_0\left[\frac{\partial \ln L(\theta_0|y)}{\partial \theta_0}\right] &= E_0\left[\left(\frac{\partial \ln L(\theta_0|y)}{\partial \theta_0}\right)\left(\frac{\partial \ln L(\theta_0|y)}{\partial \theta_0'}\right)\right] \\
&= -E_0\left[\frac{\partial^2 \ln L(\theta_0|y)}{\partial \theta_0 \partial \theta_0'}\right]
\end{aligned}
$$

The information matrix equality.

16.4.6 Estimating the Asymptotic Variance of the Maximum Likelihood Estimator

$$
\begin{aligned}
[I(\theta_0)]^{-1} &= \left(-E_0\left[\frac{\partial^2 \ln L(\theta_0)}{\partial\theta_0\partial\theta_0'}\right]\right)^{-1} \\
\left[\hat{I}(\hat{\theta})\right]^{-1} &= \left(-\frac{\partial^2 \ln L(\hat{\theta})}{\partial\theta\partial\theta'}\right) \\
\left[\hat{\hat{I}}(\hat{\theta})\right]^{-1} &= \left(\sum_{i=1}^{n}\frac{\partial \ln L(\hat{\theta}|y_i)}{\partial\theta}\frac{\partial \ln L(\hat{\theta}|y_i)}{\partial\theta'}\right)^{-1}
\end{aligned}
$$

Example 16.4 Variance Estimators for MLE I

Find the maximum likelihood estimate of β for the following density

$$f(y_i, x_i, \beta) = \frac{1}{\beta + x_i} e^{-y_i/(\beta + x_i)}.$$

$$\begin{aligned}
\ln L(\beta) &= -\sum_{i=1}^{n} \ln(\beta + x_i) - \sum_{i=1}^{n} \frac{y_i}{\beta + x_i} \\
\frac{\partial \ln L(\beta)}{\partial \beta} &= -\sum_{i=1}^{n} \frac{1}{(\beta + x_i)} + \sum_{i=1}^{n} \frac{y_i}{(\beta + x_i)^2} = 0 \\
\frac{\partial^2 \ln L(\beta)}{\partial \beta^2} &= \sum_{i=1}^{n} \frac{1}{(\beta + x_i)^2} - 2\sum_{i=1}^{n} \frac{y_i}{(\beta + x_i)^3}
\end{aligned}$$

Variance Estimators for MLE II

$$
\begin{aligned}
I(\hat{\theta})^{-1} &= \left(-E_0 \left[\frac{\partial^2 \ln L(\theta_0)}{\partial \theta_0 \partial \theta_0'} \right] \right)^{-1}_{\hat{\beta}} \\
&= \left(\sum_{i=1}^{n} \frac{1}{(\hat{\beta} + x_i)^2} - 2 \sum_{i=1}^{n} \frac{E_0[y_i]_{\hat{\beta}}}{(\hat{\beta} + x_i)^3} \right)^{-1} \\
&= \left(\sum_{i=1}^{n} \frac{1}{(\hat{\beta} + x_i)^2} - 2 \sum_{i=1}^{n} \frac{1}{(\hat{\beta} + x_i)^2} \right)^{-1} = 44.255 \\
\hat{I}(\hat{\theta})^{-1} &= \sum_{i=1}^{n} \frac{1}{(\hat{\beta} + x_i)^2} - 2 \sum_{i=1}^{n} \frac{y_i}{(\hat{\beta} + x_i)^3} = 46.163 \\
\hat{\hat{I}}(\hat{\theta})^{-1} &= \frac{1}{\sum_{i=1}^{n} [-1/(\hat{\beta} + x_i) + y_i/(\hat{\beta} + x_i)^2]^2} = 100.512
\end{aligned}
$$

For the sample data in Example C.1 in Greene, 6th Edition, the above estimates are obtained.

16.5 Likelihood and Econometric Models I

Till now the analysis was done in terms of the density of an observed random variable and a vector of parameters, $f(y_i|\alpha)$. However, econometric models will involve exogenous or predetermined variables, x_i, so the results must be extended.

By partitioning the joint density of y_i and x_i into the product of the conditional and the marginal, the log-likelihood function may be written:

$$\ln L(\alpha|data) = \sum_{i=1}^{n} \ln f(y_i, x_i|\alpha) = \sum_{i=1}^{n} \ln f(y_i|x_i, \alpha) + \sum_{i=1}^{n} \ln g(x_i|\alpha)$$

$$\ln L(\theta, \delta|data) = \sum_{i=1}^{n} \ln f(y_i, x_i|\alpha) = \sum_{i=1}^{n} \ln f(y_i|x_i, \theta) + \sum_{i=1}^{n} \ln g(x_i|\delta)$$

16.5 Likelihood and Econometric Models II

Asymptotic results for the MLE must now account for the presence of x_i in the functions and derivatives of $\ln f(y_i|x_i, \theta)$. We will proceed under the assumption of well-behaved data so that sample averages such as

$$(1/n) \ln L(\theta|y, X) = (1/n) \sum_{i=1}^{n} \ln f(y_i|x_i, \theta)$$

and its gradient with respect to θ will converge in probability to their population expectations.

16.6 Hypothesis and Specification Test and Fit Measures I

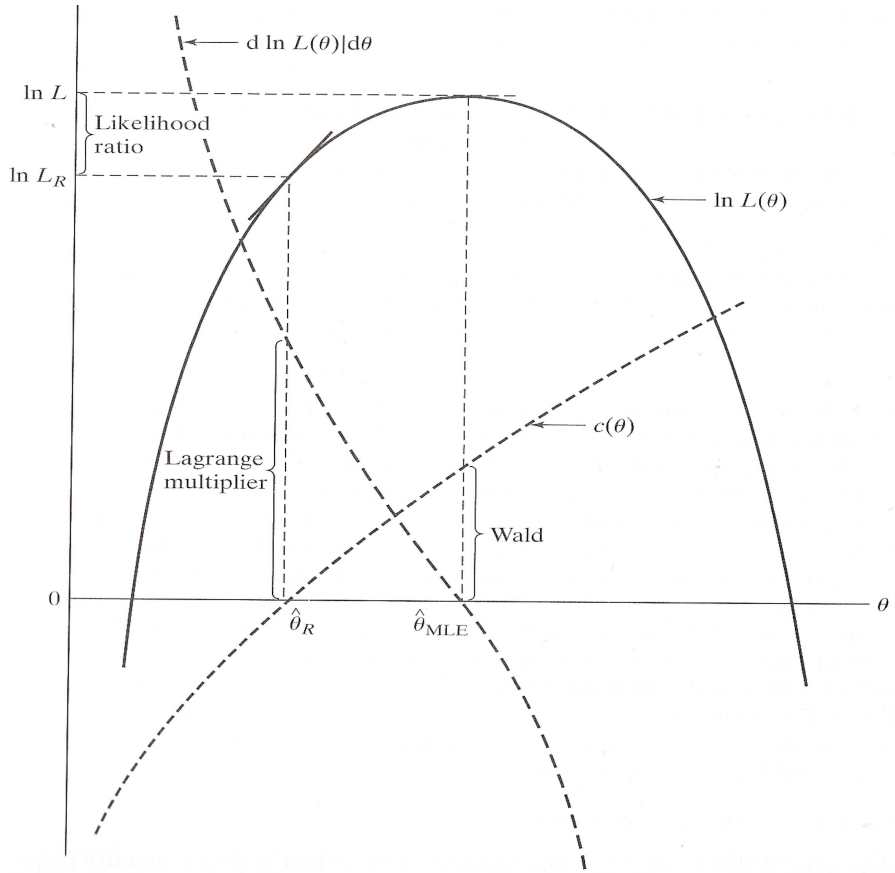

FIGURE 16.2 Three Bases for Hypothesis Tests.

16.6 Hypothesis and Specification Test and Fit Measures II

- Likelihood ratio test. If the restriction $c(\theta) = 0$ is valid, then imposing it should not lead to a large reduction in the log-likelihood function. Therefore, we base the test on the difference, $\ln L_U - \ln L_R$, where L_U is the value of the likelihood function at the unconstrained value of θ and L_R is the value of the likelihood function at the restricted estimate.

- Wald test. If the restriction is valid, then $c(\hat{\theta}_{MLE})$ should be close to zero because the MLE is consistent. Therefore, the test is based on $c(\hat{\theta}_{MLE})$. We reject the hypothesis if this value is significantly different from zero.

16.6 Hypothesis and Specification Test and Fit Measures III

- Lagrange multiplier test. If the restriction is valid, then the restricted estimator should be near the point that maximizes the log-function. Therefore, the slope of the log-likelihood function should be near zero at the restricted estimator. The test based on the slope of the log-likelihood at the point where the function is maximized subject to the restriction.

These three tests are asymptotically equivalent under the null hypothesis, but they can behave rather differently in a small sample.

16.6 Hypothesis and Specification Test and Fit Measures IV

- Likelihood ratio test. $LR = -2 \ln \frac{\hat{L}_R}{\hat{L}_U} \sim^a \chi^2(df)$.
- Wald test.
  $W = [c(\hat{\theta}) - q]'(Asy.Var.[c(\hat{\theta}) - q])^{-1}[c(\hat{\theta}) - q] \sim^a \chi^2(df)$.
  The $Est.Asy.Var.[c(\hat{\theta}) - q] = \hat{C}\left(Est.Asy.Var[\hat{\theta}]\right)\hat{C}'$, where
  $\hat{C} = \left(\frac{\partial c(\hat{\theta})}{\partial \hat{\theta}'}\right)$.
- Lagrange multiplier test.
  $LM = \left(\frac{\partial \ln L(\hat{\theta}_R)}{\partial \theta_R}\right)' [I(\hat{\theta}_R)]^{-1} \left(\frac{\partial \ln L(\hat{\theta}_R)}{\partial \theta_R}\right) \sim^a \chi^2(df)$.

Where df equals to the number of restrictions imposed.

16.6.5 Comparing Models and Computing Model Fit

For nonnested models, the computation is a comparison of one model to another based on an estimation criterion to discern which is to be preferred.

1. Information criteria. Two common measures that are based on the same logic as the adjusted R^2 for the linear model are
   Akaike information criterion (AIC) $= -2 \ln L + 2K$,
   Bayes (Schwarz) information criterion (BIC)
   $= -2 \ln L + K \ln n$.
2. Vuong statistic. $V = \frac{\sqrt{n}\bar{m}}{s_m}$, where $m_i = \ln L_{i,1} - \ln L_{i,2}$. Under the hypothesis that the models are equivalent $V \to N(0,1)$, large positive (negative) values indicate that model 1 (2) is 'better'.
3. Correlation between prediction and actual value. $Corr(y, \hat{y})$.

16.7 Two-Step Maximum Likelihood Estimation I

The literature contains a large and increasing number of models in which **one model is embedded in another**, which produces what are broadly known as 'two-step' estimation problems.

$$y_2 = f(x_2, \theta_2, E[y_1|x_1, \theta_1])$$

There are two ways to proceed:

1. Full information maximum likelihood (FIML). Forming the joint distribution $f(y_{i1}, y_{i2}|x_{i1}, x_{i2}, \theta_1, \theta_2)$ of the two random variables and then maximizing the full log-likelihood function, $\ln L = \sum_{i=1}^{n} f(y_{i1}, y_{i2}|x_{i1}, x_{i2}, \theta_1, \theta_2)$.
2. Limited information maximum likelihood (LIML). Estimating the parameters of model 1, and then maximizing a conditional log-likelihood function using the estimates from step 1: $\ln \hat{L} = \sum_{i=1}^{n} f(y_{i2}|x_{i2}, \theta_2, (x_{i1,\theta_1})$.

16.7 Two-Step Maximum Likelihood Estimation II

Asymptotic Distribution of the Two-Step MLE

If the standard regularity conditions are met for both log-likelihood functions, then the second-step maximum likelihood estimator θ_2 is consistent and asymptotically normally distributed with asymptotic covariance matrix

$$V_2^* = \frac{1}{n}\left[V_2 + V_2[CV_1C' - RV_1C' - CV_1R']V_2\right],$$

where

$$V_1 = Asy.Var[\sqrt{n}(\hat{\theta}_1 - \theta_1)] \text{ based on } \ln L_1,$$

$$V_2 = Asy.Var[\sqrt{n}(\hat{\theta}_2 - \theta_2)] \text{ based on } \ln L_2|\theta_1,$$

$$C = E\left[\frac{1}{n}\left(\frac{\partial \ln L_2}{\partial \theta_2}\right)\left(\frac{\partial \ln L_2}{\partial \theta_1'}\right)\right],\ R = E\left[\frac{1}{n}\left(\frac{\partial \ln L_2}{\partial \theta_2}\right)\left(\frac{\partial \ln L_1}{\partial \theta_1'}\right)\right].$$

16.7 Two-Step Maximum Likelihood Estimation III

The correction of the asymptotic covariance matrix at the second step requires some additional computation. Matrices V_1 and V_2 are estimated by the respective uncorrected covariance matrices. Typically, the outer product estimators,

$$\hat{V}_1 = \left[\frac{1}{n}\sum_{i=1}^{n}\left(\frac{\partial \ln f_{i1}}{\partial \hat{\theta}_1}\right)\left(\frac{\partial \ln f_{i1}}{\partial \hat{\theta}_1'}\right)\right]^{-1},$$

$$\hat{V}_2 = \left[\frac{1}{n}\sum_{i=1}^{n}\left(\frac{\partial \ln f_{i2}}{\partial \hat{\theta}_2}\right)\left(\frac{\partial \ln f_{i2}}{\partial \hat{\theta}_2'}\right)\right]^{-1}$$

are used.

16.7 Two-Step Maximum Likelihood Estimation III

The matrices R and C are obtained by summing the individual observations on the cross products of the derivatives. These are estimated with

$$\hat{C} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{\partial \ln f_{i2}}{\partial \hat{\theta}_2}\right)\left(\frac{\partial \ln f_{i2}}{\partial \hat{\theta}_1'}\right),$$

$$\hat{R} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{\partial \ln f_{i2}}{\partial \hat{\theta}_2}\right)\left(\frac{\partial \ln f_{i1}}{\partial \hat{\theta}_1'}\right).$$

16.8.1 Maximum Likelihood and GMM Estimation I

In order to obtain the maximum likelihood estimator,

$$\text{Maximize}_\beta (1/n) \ln L(\beta|y, X) = (1/n) \sum_{i=1}^{n} \ln f(y_i|x_i, \beta)$$

We **maximize** the **log-likelihood function** by equating its **derivatives** to **zero**, so the MLE is obtained by solving the set of empirical moment equations

$$\frac{1}{n} \sum_{i=1}^{n} \frac{\partial \ln f(y_i|x_i, \beta)}{\partial \beta} = \frac{1}{n} \sum_{i=1}^{n} d_i(\beta) = \bar{d}(\beta) = 0$$

The population counterpart to the sample moment equation is

$$E\left[\frac{1}{n} \frac{\partial \ln L}{\partial \beta}\right] = E\left[\frac{1}{n} \sum_{i=1}^{n} d_i(\beta)\right] = E[\bar{d}(\beta)] = 0$$

16.8.1 Maximum Likelihood and GMM Estimation II

Using what we know about GMM estimators, then $\hat{\beta}_{ML}$ is consistent and asymptotically normally distributed with asymptotic covariance matrix equal to

$$V = [G(\beta)'G(\beta)]^{-1}G(\beta)'\{Var[\bar{d}(\beta)]\}G(\beta)[G(\beta)'G(\beta)]^{-1},$$

where $G(\beta) = \text{plim } \partial\bar{d}(\beta)/\partial\beta'$.

In the MLE case $G(\beta) = (1/n)E[H(\beta)] = \bar{H}(\beta)$, consequently $V = (-E[H(\beta)])^{-1}$.

Hence, we have developed an efficient, generalized method of moments estimator that has the same asymptotic properties as the MLE under the assumption of normality.