

ECON 5350 Class Notes
Chapter 16-18. Alternative Estimation Frameworks

1 Introduction

We have already examined two different estimation frameworks – least squares and maximum likelihood. However, these are not the only types of estimations frameworks. With recent advancements in computing power, other types of estimation are becoming more common – such as Bayesian, generalized method of moments (GMM), simulation-based and kernel estimation. This section gives an introduction to these alternative types of estimation by placing them in one of three categories:

- Parametric Estimation. Makes the strongest assumptions about functional form and distribution of the errors. If the assumptions are correct, will generally be the most efficient and allows one to draw the sharpest conclusions. However, the assumptions may be incorrect.
- Semi-Parametric Estimation. Relaxes some assumptions but maintains others. Tradeoff between additional flexibility and reduced ability to draw sharp conclusions.
- Non-Parametric Estimation. Relaxes all parametric assumptions. Gives the ultimate flexibility in fitting the data and is robust to parametric assumptions, but provides limited ability to draw precise inferences.

2 Parametric Estimation

2.1 Maximum Likelihood

We have already discussed maximum likelihood estimation (and will further in the next two chapters) so I will keep this section brief.

2.1.1 Basic Framework

Begin by assuming the data-generating process (conditional pdf) is

$$y_i|x_i' \sim N[x_i'\beta, \sigma^2].$$

Maximum likelihood estimation (MLE) proceeds by writing down the joint pdf for a given sample of data $\{(y_1, x'_1), (y_2, x'_2), \dots, (y_n, x'_n)\}$

$$L(\theta) = f(y_1, \dots, y_n | x'_1, \dots, x'_n; \theta) = \prod_{i=1}^n f(y_i | x'_i; \theta) \quad (1)$$

and choosing $\theta = (\beta, \sigma^2)'$ to maximize (1). Typically, this problem is rewritten so as to maximize

$$\ln L(\theta) = \sum_{i=1}^n \ln f(y_i | x'_i; \theta)$$

and requires nonlinear optimization methods.

2.1.2 Properties of Maximum Likelihood Estimators

ML estimators are attractive because of their large-sample properties. Assuming certain regularity conditions (Greene, p. 474) hold, MLE has the following properties.

Properties

- The estimator is consistent: $\hat{\theta}_{ML} \xrightarrow{p} \theta$.
- The estimator is asymptotically normal: $\hat{\theta}_{ML} \overset{asy}{\sim} N[\theta, I(\theta)^{-1}]$.
- The estimator is asymptotically efficient: $asy.var.(\hat{\theta}_{ML})$ achieves the Cramer-Rao lower bound $I(\theta)^{-1} = -E[\partial^2 \ln L / (\partial \theta \partial \theta')]^{-1}$ for consistent estimators.
- The estimator is invariant: $g(\hat{\theta}_{ML})$ is the ML estimator of $g(\theta)$, provided g is continuous and differentiable.

2.1.3 A MLE Example

The Poisson distribution is

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

for $y_i = 0, 1, 2, \dots$. The log likelihood function for a sample of size n is

$$\ln L(\theta) = -n\theta + n \ln(\theta) \bar{y} - \sum_{i=1}^n \ln(y_i!).$$

The first-order condition for maximization (with respect to θ) is

$$\frac{\partial \ln L(\theta)}{\partial \theta} = -n + n \frac{\bar{y}}{\theta} = 0,$$

which implies that $\hat{\theta}_{ML} = \bar{y}$. The sample average is therefore a consistent estimator of θ . The second-order condition is

$$\frac{\partial^2 \ln L(\theta)}{\partial \theta^2} = -\frac{n\bar{y}}{\theta^2} \leq 0.$$

The Cramer-Rao lower bound is

$$E \left[\frac{n\bar{y}}{\theta^2} \right]^{-1} = \frac{\theta^2}{nE(\bar{y})}.$$

Using the moment-generating function, $m(t, \theta) = E(e^{ty})$, it is straightforward to show that the expected value of any y_i is

$$E(y_i) = \sum_{i=1}^{\infty} y_i \frac{e^{-\theta} \theta^{y_i}}{y_i!} = \theta.$$

Therefore the *asy.var.*($\hat{\theta}_{ML}$) = $\frac{1}{n}\theta$. Because, $\hat{\theta}_{ML} = \bar{y}$, this implies that $var(y_i) = E(y_i) = \theta$.

2.2 Bayesian Estimation

Bayesian estimation is fundamentally different than standard Classical estimation (e.g., maximum likelihood, least squares), in which we attempt to provide estimates of fixed population parameters. Rather, under a Bayesian philosophy, we are continually updating our beliefs about the distribution of the parameters.

Bayesian estimation relies on Bayes' theorem

$$P(A|B)P(B) = P(B|A)P(A)$$

or written more familiarly as

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

where A and B are events. Treating A as the parameters (θ) and B as the data (Y), we have

$$P(\theta|Y) = \frac{P(Y|\theta)P(\theta)}{P(Y)} \propto P(Y|\theta)P(\theta), \quad (2)$$

where \propto indicates "proportional to". The left-hand side of (2) is called the posterior density, the first term on the right-hand side is the likelihood function and the last right-hand side term is the prior density. The

Bayesian estimator is then the mean of the posterior distribution.

2.2.1 An Example

The likelihood function for the classical regression model ($Y = X\beta + \epsilon$) with normally distributed errors is

$$\ln L(y, X|\beta, \sigma^2) = [2\pi\sigma^2]^{-0.5n} \exp[-0.5\sigma^{-2}(y - X\beta)'(y - X\beta)]. \quad (3)$$

Under Classical sampling theory, we interpreted this as the joint probability of the data conditional on the fixed (albeit unknown) values for the parameters. Under Bayesian estimation theory, we interpret (3) as the joint probability of the parameters given a new sample of information on y and X . The posterior density is then proportional to

$$L(y, X|\theta) \times g(\theta)$$

where $\theta = (\beta, \sigma^2)'$ and g is the prior distribution for θ . Setting $g(\theta)$ equal to the uniform density results in a noninformative or "flat" prior. An informative prior would have a more interesting shape. For more details regarding Bayesian estimation, see Mittelhammer, Judge and Miller's (2000) text *Econometric Foundations*.

3 Semi-Parametric Estimation

3.1 Generalized Method of Moments (GMM)

I begin by outlining the classical method of moments technique (Fisher, 1925) and then proceed to generalized method of moments (Hansen, 1982).

3.1.1 Traditional Method of Moments

The idea is to match the population moments of a distribution to the sample moments, using as many moments as necessary to estimate the unknown parameters. Let $\{X_1, X_2, \dots, X_n\}$ be a random sample from the pdf $f(x; \theta_1, \dots, \theta_r)$. Also, let

$$m'_k = \frac{1}{n} \sum_{i=1}^n x_i^k$$

be the k^{th} sample moment and $\mu'_k = E(X^k)$ the k^{th} population moment. The method of moments estimator for $\theta = (\theta_1, \dots, \theta_r)'$ is therefore the solution to the equations

$$m'_i = \mu'_i(\theta)$$

for $i = 1, \dots, r$. Method of moments can be modified to use centered, as opposed to raw, moments. While consistent, method of moments estimators are not generally efficient.

Example Suppose $\{X_1, X_2, \dots, X_n\}$ is a random sample from a $gamma(\alpha, \beta)$ distribution. The likelihood function

$$L(\theta) = (\Gamma(\alpha)\beta^\alpha)^{-n} (x_1 x_2 \cdots x_n)^{\alpha-1} \exp\left(-\sum_{i=1}^n x_i/\beta\right)$$

is difficult to evaluate without using numerical methods. A method of moments estimator jointly solves

$$\begin{aligned}\bar{X} &= \frac{1}{n} \sum_i X_i = E(X) = \alpha\beta \\ \hat{\sigma}^2 &= \frac{1}{n} \sum_i (X_i - \bar{X})^2 = E[(X - \mu_1)^2] = \alpha\beta^2\end{aligned}$$

for $\hat{\alpha}$ and $\hat{\beta}$. This gives

$$\begin{aligned}\hat{\alpha} &= \bar{X}^2/\hat{\sigma}^2 \\ \hat{\beta} &= \hat{\sigma}^2/\bar{X}.\end{aligned}$$

Variance of Method of Moments Estimator Let the sample moments be

$$\bar{g}_k = \frac{1}{n} \sum_i g_k(X_i)$$

for $k = 1, \dots, K$ and $\bar{g} = (\bar{g}_1, \dots, \bar{g}_K)$ have asymptotic variance-covariance matrix V , with elements

$$V_{jk} = \frac{1}{n} \left\{ \frac{1}{n} \sum_i (g_j(X_i) - \bar{g}_j)(g_k(X_i) - \bar{g}_k) \right\}$$

where $j, k = 1, \dots, K$. Now let G be the matrix

$$G = \begin{bmatrix} \frac{\partial \bar{g}_1}{\partial \theta_1} & \frac{\partial \bar{g}_1}{\partial \theta_2} & \dots & \frac{\partial \bar{g}_1}{\partial \theta_K} \\ \frac{\partial \bar{g}_2}{\partial \theta_1} & \frac{\partial \bar{g}_2}{\partial \theta_2} & & \frac{\partial \bar{g}_2}{\partial \theta_K} \\ \vdots & & \ddots & \vdots \\ \frac{\partial \bar{g}_K}{\partial \theta_1} & \frac{\partial \bar{g}_K}{\partial \theta_2} & \dots & \frac{\partial \bar{g}_K}{\partial \theta_K} \end{bmatrix}_{K \times K}.$$

Since the population moments $\mu(\theta)$ are typically a nonlinear function in θ , we will linearize using a first-order Taylor approximation to $\bar{g}_k = \mu_k(\theta)$ around the true value θ

$$\begin{aligned} \bar{g} &\cong \mu(\theta) + G(\theta)(\hat{\theta} - \theta) \Rightarrow \\ (\hat{\theta} - \theta) &= G^{-1}(\theta)(\bar{g} - \mu(\theta)). \end{aligned}$$

Therefore, our estimate of the asymptotic variance is

$$est.asy.var.(\hat{\theta}) = \hat{G}^{-1}V(\hat{G}^{-1})'.$$

Gamma Example Continued In the gamma distribution example above, where $g_1 = X_i$ and $g_2 = (X_i - \bar{X})^2$, we have

$$\hat{G} = \begin{bmatrix} \hat{\beta} & \hat{\alpha} \\ \hat{\beta}^2 & 2\hat{\alpha}\hat{\beta} \end{bmatrix}$$

and

$$V = \frac{1}{n} \begin{bmatrix} \widehat{var}(g_1) & \widehat{cov}(g_1, g_2) \\ \widehat{cov}(g_2, g_1) & \widehat{var}(g_2) \end{bmatrix}.$$

3.1.2 Generalized Method of Moments

GMM extends the classical method of moments estimator to handle cases where there are more moment conditions than parameters to estimate (i.e., the model is overidentified).

Basic Framework Suppose there are K parameters to estimate $\theta = (\theta_1, \dots, \theta_K)'$ and $L \geq K$ moment conditions

$$E[m_l(y_i, X_i, Z_i; \theta)] = 0 \tag{4}$$

for $l = 1, \dots, L$. The sample analog of (4) is

$$\bar{m}_l(y_i, X_i, Z_i; \theta) = \frac{1}{n} \sum_{i=1}^n m_l(y_i, X_i, Z_i; \theta) = 0$$

which will generally have a unique solution if $L = K$ and multiple solutions if $L > K$. To reconcile the multiple solutions, consider minimization of

$$q = \bar{m}(\theta)' W_n \bar{m}(\theta)$$

where $\bar{m}(\theta) = (\bar{m}_1, \dots, \bar{m}_L)'$ and W_n is a positive definite weighting matrix. If $W_n = I_n$, then minimization of q is simply a least squares criterion. If $W_n \neq I_n$, then minimization of q is similar in spirit to GLS, which re-weights the observations according to the variance-covariance matrix of the errors. Again, in the spirit of GLS, Hansen (1982) shows that the optimal criterion (weighting matrix) is to minimize

$$q = \bar{m}(\theta)' \Phi^{-1} \bar{m}(\theta)$$

where

$$\Phi = \text{Asy.Var.}(\sqrt{n}\bar{m}).$$

The resulting estimator, $\hat{\theta}_{GMM}$, will have an asymptotic variance-covariance matrix equal to

$$\text{Est.Asy.Var.}(\hat{\theta}_{GMM}) = \frac{1}{n} [\Gamma' \hat{\Phi}^{-1} \Gamma]^{-1}$$

where Γ is a matrix of partial derivatives similar in spirit to G above.

Properties of the GMM Estimator Assuming that the

1. parameters are identifiable,
2. empirical moments converge in probability to their population counterparts (i.e., $\bar{m}(\theta) \xrightarrow{p} 0$), and
3. the empirical moments obey the central limit theorem (i.e., $\sqrt{n}\bar{m}(\theta) \xrightarrow{d} N[0, \Phi]$),

then

$$\hat{\theta}_{GMM} \stackrel{asy}{\sim} N\left[\theta, \frac{1}{n} (\Gamma' \Phi^{-1} \Gamma)^{-1}\right].$$

Example #1. Ordinary Least Squares – Exactly Identified Case Nearly all estimators we have covered can be posed as method of moment estimators. Consider GMM estimation of the bivariate linear regression model

$$y_i = \alpha + \beta x_i + \epsilon_i.$$

Two moment conditions arising from the Classical assumptions are

$$\begin{aligned} E[m_1(y_i, x_i; \alpha, \beta)] &= E(\epsilon_i) = 0 \\ E[m_2(y_i, x_i; \alpha, \beta)] &= E(\epsilon_i x_i) = 0. \end{aligned}$$

The sample analog of these population moment conditions are

$$\begin{aligned} \frac{1}{n} \sum_i e_i &= \frac{1}{n} \sum_i (y_i - \hat{\alpha} - \hat{\beta} x_i) = 0 \\ \frac{1}{n} \sum_i e_i x_i &= \frac{1}{n} \sum_i (y_i - \hat{\alpha} - \hat{\beta} x_i) x_i = 0, \end{aligned}$$

which are, of course, the normal equations for OLS estimation of the classical linear regression model. In this instance, the weighting matrix W is irrelevant because both moment conditions can be satisfied exactly. Therefore, we have

$$\begin{aligned} \hat{\alpha}_{GMM} &= \hat{\alpha}_{OLS} = \bar{y} - b\bar{x} \\ \hat{\beta}_{GMM} &= \hat{\beta}_{OLS} = \frac{\sum_i (y_i - \bar{y})(x_i - \bar{x})}{\sum_i (x_i - \bar{x})^2}. \end{aligned}$$

Example #2. Hall’s Random-Walk Consumption Hypothesis In a famous 1978 article in the *Journal of Political Economy*, Robert Hall showed that under certain conditions, consumption should be expected to follow a random walk. Consider an agent that chooses consumption c_t to maximize discounted, expected lifetime utility

$$E_0 \sum_{t=0}^T (1 + \rho)^{-t} 0.5 \phi [\bar{c} - c_t]^2,$$

where ρ is the subjective discount rate, ϕ is a constant, and \bar{c} is the bliss level of consumption, subject to

$$A_0 = \sum_{t=0}^T (1 + r)^{-t} (c_t - w_t)$$

where A_0 is initial assets, r is the interest rate and w_t is the wage rate. Hall shows that if $\rho = r$, then consumption follows a random walk

$$c_t = c_{t-1} + \epsilon_t$$

where $E_{t-1}[\epsilon_t] = 0$. Campbell and Mankiw (1989) test Hall's hypothesis by posing a specific alternative – agents simply consume a given fraction λ of their current income (i.e., $c_t = \lambda w_t$). The two hypotheses can be nested according to

$$\begin{aligned} c_t - c_{t-1} &= \lambda(w_t - w_{t-1}) + (1 - \lambda)\epsilon_t \\ \Delta c_t &= \lambda\Delta w_t + \nu_t. \end{aligned}$$

In principle, one could just run a regression of the change in consumption on the change in income and test whether the coefficient λ is different than zero. The problem is that Δw_t and ν_t are likely to be correlated so that instrumental variables need to be found. Consider using the first four lagged changes in consumption: $\Delta c_{t-1}, \dots, \Delta c_{t-4}$. The moment conditions are therefore

$$\begin{aligned} E[m_1(\Delta c_t, \Delta w_t, \Delta c_{t-1}; \lambda)] &= E[\nu_t \Delta c_{t-1}] = 0 \\ E[m_2(\Delta c_t, \Delta w_t, \Delta c_{t-2}; \lambda)] &= E[\nu_t \Delta c_{t-2}] = 0 \\ E[m_3(\Delta c_t, \Delta w_t, \Delta c_{t-3}; \lambda)] &= E[\nu_t \Delta c_{t-3}] = 0 \\ E[m_4(\Delta c_t, \Delta w_t, \Delta c_{t-4}; \lambda)] &= E[\nu_t \Delta c_{t-4}] = 0. \end{aligned}$$

The sample analogs are

$$\begin{aligned} \frac{1}{T} \sum_{t=0}^T m_{1t}(\hat{\lambda}) &= \frac{1}{T} \sum_{t=0}^T (\Delta c_t - \hat{\lambda} \Delta w_t) \Delta c_{t-1} = 0 \\ \frac{1}{T} \sum_{t=0}^T m_{2t}(\hat{\lambda}) &= \frac{1}{T} \sum_{t=0}^T (\Delta c_t - \hat{\lambda} \Delta w_t) \Delta c_{t-2} = 0 \\ \frac{1}{T} \sum_{t=0}^T m_{3t}(\hat{\lambda}) &= \frac{1}{T} \sum_{t=0}^T (\Delta c_t - \hat{\lambda} \Delta w_t) \Delta c_{t-3} = 0 \\ \frac{1}{T} \sum_{t=0}^T m_{4t}(\hat{\lambda}) &= \frac{1}{T} \sum_{t=0}^T (\Delta c_t - \hat{\lambda} \Delta w_t) \Delta c_{t-4} = 0. \end{aligned}$$

The GMM estimate $\hat{\lambda}_{GMM}$ minimizes

$$q = \bar{m}(\lambda)' W_T \bar{m}(\lambda)$$

where $W_T^{-1} = \Phi$ is the asymptotic variance of $\sqrt{n}\bar{m}(\lambda)$. See [Gauss example 16-18.1](#) for OLS, 2SLS and GMM estimates of λ .

Testing the Validity of the Overidentification Restrictions In an exactly identified system, $q = 0$. In an overidentified system, the moment restrictions implied by theory will not all be satisfied exactly in the data. Therefore, $q > 0$. This observation forms the basis for a test of overidentifying restrictions. If q is substantially greater than zero, then this suggests that at least one of the overidentifying restrictions is likely to be false. Similar to the Wald test introduced in earlier chapters, we have

$$nq = [\sqrt{n}\bar{m}(\hat{\theta})]' \hat{\Phi}^{-1} [\sqrt{n}\bar{m}(\hat{\theta})] \overset{asy}{\rightsquigarrow} \chi^2[L - K].$$

3.2 Simulation-Based Estimation

This section outlines a relatively new econometric method for estimating parameters when the criterion function does not have a closed-form solution. This often occurs when latent variables are involved – perhaps due to situations such as measurement error, censoring, or qualitative choices. In these instances, it is often necessary to examine expectations of random variables that will involve integral expressions that cannot be written in a closed form. Simulation-based methods, relying on advances in computing power, are one possible solution to this problem.

3.2.1 An Example. Random Parameters

Consider the fixed-effects panel-data model

$$y_{it} = \alpha_i + x'_{it}\beta_i + \epsilon_{it}$$

where $\beta_i = \beta + z'_i\gamma + \nu_i$ is a random variable since the individual heterogeneity term ν_i is stochastic with density

$$\nu_i \sim g(0, \Sigma).$$

The unconditional density for y_i is

$$f(y_i) = \int_{\nu_i} f(y_i | x_i, \beta + z'_i\gamma + \nu_i) g(\nu_i | \Sigma) d\nu_i.$$

The log-likelihood function is therefore

$$\ln L(\beta, \gamma, \Sigma) = \sum_{i=1}^n \ln \left\{ \int_{\nu_i} f(y_i | x_i, \beta + z_i' \gamma + \nu_i) g(\nu_i | \Sigma) d\nu_i \right\} \quad (5)$$

and is maximized by choosing β , γ and Σ . Since (5) involves an integral and cannot be written in a closed-form solution, it may be difficult to maximize.

One solution is to generate artificial data on ν_i by taking R random draws from $g(\nu_i | \Sigma)$. The simulated log-likelihood function is then

$$\ln L(\beta, \gamma) = \sum_{i=1}^n \ln \left\{ \frac{1}{R} \sum_{r=1}^R f(y_i | x_i, \beta + z_i' \gamma + \nu_{ir}) \right\}, \quad (6)$$

which is maximized by choosing β and γ . In many instances, it may be easier to maximize (6) than (5), especially when multiple integrals are involved. Details of simulation-based estimation, including simulated method of moments, can be found in Gourieroux and Monfort's (1996) book *Simulation-Based Econometric Methods*.

4 Non-Parametric Estimation

Will be covered later, time permitting.