Estimation Methods

Table of contents

1 Overview

In this handout we will look at several approaches to generate estimators:

- Least Squares
- Method of Moments
- Maximum Likelihood

We will discuss each approach in the context of the Classical Linear Regression Model discussed in Lecture 1. You may also wish to revise the notes on Linear Algebra.

Further reading can be found in:

- Section 5.6 of Cameron and Trivedi (2005)
- Section 6.1 of Verbeek (2017)

2 Review of CLRM

The classical linear regression model is states that the conditional expectation function $E[Y_i|X_i]$ is linear in parameters.

For the random sample $i = 1, ..., n$,

$$Y_i = X_i'\beta + u_i$$

where X_i is a random k-dimensional vector (k-vector) and β a non-random k-vector of population parameters. Both Y_i and u_i are random scalars.

As we saw, we can stack each observation into a column vector:

$$Y = \underbrace{\begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix}}_{n\times 1} = \underbrace{\begin{bmatrix} X_1'\beta \\ X_2'\beta \\ \vdots \\ X_n'\beta \end{bmatrix}}_{n\times 1} + \underbrace{\begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_n \end{bmatrix}}_{n\times 1} = \underbrace{\begin{bmatrix} X_{11} & X_{12} & \dots & X_{1k} \\ X_{21} & X_{22} & & \\ \vdots & & \ddots & \\ X_{n1} & & & X_{nk} \end{bmatrix}}_{n\times k} \underbrace{\begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{bmatrix}}_{k\times 1} + \begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_n \end{bmatrix} = X\beta + u$$

where X is now a $n \times k$ random matrix, but β remains a non-random k-vector of population parameters.

> **Data as a Matrix**
>
> If you have any experience working datasets in Stata/R, you will know that they tend to have a rectangular structure: each row *typically* represents an observation and each

column a variable. This is the structure depicted above in matrix notation: each row of the X matrix depicts an observation and each column a regressor. The dataset you are using contains is a matrix of observations for both the outcome variable and regressors: $[Y, X,]$. Of course, we do not observe the error term.

In this section, we will employ the CLRM assumptions as we discuss the three approaches to estimation.

3 (Ordinary) Least Squares

The Ordinary Least Squares (OLS) estimator is the 'work-horse' of applied economics research.[1] It is not the only Least Squares estimator, but as the simplest case is the most useful place to start. Other Least Squares estimators include Weighted Least Squares (WLS), (Feasible) Generalized Least Squares (GLS), Non-Linear Least Squares and Two-Stage Least Squares (2SLS).

OLS is "ordinary" in the sense that there is no additional features to the method. For example, WLS applies a unique weight to each observation while OLS weights each observation equally. While OLS is arguably 'vanilla' in this way, it is efficient (as we shall see).

In general, LS estimators minimize a measure of 'distance' between the observed outcomes and the fitted values of the model. The measure of distance is sum of squared deviations (squared Euclidean or ℓ_2 norm).[2]

3.1 Application to CLRM

In the case of OLS estimator to the CLRM, the goal is to find the b-vector that minimizes,

$$\sum_{i=1}^{n}(Y_i - \underbrace{\tilde{Y}_i}_{\tilde{X}_i'b})^2 = \sum_{i=1}^{n}\tilde{u}_i^2$$

This sum-of-squares can be written as the inner product of two-vectors:[3]

[1] Machine Learning techniques, such as neural networks, use non-linear operators such as the Softmax function and Rectified Linear Unit (ReLU). Who knows, in a few years, we may no longer think of OLS as the 'work-horse' of applied statistics and research.

[2] A measure of distance must be positive. You can use the (sum of) absolute-value deviations, but the least squares has nice properties including ease of differentiation and overall efficiency (of the estimator).

[3] The inner (or dot) product of two *equal-length* vectors, a and b, is defined as:

$$\langle a, b\rangle = a \cdot b = \sum_{i=1}^{k} a_i \times b_i = a'b$$

$$\sum_{i=1}^{n} \tilde{u}_i^2 = \tilde{u}'\tilde{u} = (Y - Xb)'(Y - Xb)$$

Applying the rules of matrix transposition, the inner product of these two matrices is given by,

$$(Y' - b'X')(Y - Xb) = Y'Y - b'X'Y - Y'Xb + b'X'Xb$$

Since all terms are scalars, $b'X'Y = Y'Xb$; which then gives us,

$$Y'Y - 2b'X'Y + b'X'Xb$$

Thus, the (ordinary) least-squares estimator the vector that solves this linear expression.

$$\hat{\beta}^{OLS} = \arg\min_{b} \quad Y'Y - 2b'X'Y + b'X'Xb$$

Using the rules of vector differentiation (see Linear Algrebra) we can find the first order conditions:

$$-2X'Y + 2X'X\hat{\beta}^{OLS} = 0$$

If the X matrix is full rank ($=k$), then $X'X$ is non-singular and its inverse exists. Recall, this was one of the CLRM assumptions. Then,

$$\hat{\beta}^{OLS} = (X'X)^{-1}X'Y$$

3.2 Bias

Is the OLS estimator unbiased? The answer will depend on the assumption of the model. Here, we have assumed that the model being estimated is a CLRM. This means that we have assumed conditional mean independence of the error term:

$$E[u|X] = 0$$

The OLS-estimator is given by,

$$\hat{\beta}^{OLS} = (X'X)^{-1}X'Y = (X'X)^{-1}X'(X\beta+u) = \underbrace{(X'X)^{-1}X'X}_{I_n}\beta+(X'X)^{-1}X'u = \beta+(X'X)^{-1}X'u$$

plugging in the definition of Y from the model.

Hence,

$$E[\hat{\beta}^{OLS}|X] = E[\beta + (X'X)^{-1}X'u|X] = \beta + E[(X'X)^{-1}X'u|X]$$

Here we apply the linearity of the expectation operator and the factor that β is a non-random vector. Next, we exploit the fact that conditional on X, any function of X is non-random and can come out of the expectation operator.

$$E[\hat{\beta}^{OLS}|X] = \beta + (X'X)^{-1}X' \underbrace{E[u|X]}_{=0} = \beta$$

Notice, we require the stronger assumption of conditional mean independence, $E[u|X] = 0$. Uncorrelateness, $E[X'u] = 0$, is insufficient for unbiasedness.

> **!** Important
>
> Notice, unbiasedness depends on the assumptions of the model and not any properties of the estimator. The estimator is simply a calculation using observed data. The properties and interpretation of this computation depend on the *assumptions* we make regarding the underlying model.

It is also worth noting that the unbiasedness of the OLS estimator does NOT depend on any assumptions regarding the variance or distribution of the error term.

3.3 Efficiency

The OLS estimator is a k-dimensional random vector. The variance of this vector is a $k \times k$ variance-covariance matrix.

$$Var(\hat{\beta}) = E[\underbrace{(\hat{\beta} - E[\hat{\beta}])}_{k\times 1}\underbrace{(\hat{\beta} - E[\hat{\beta}])'}_{1\times k}]$$

The off-diagonals of the matrix are the covariances: $Cov(\hat{\beta}_j, \hat{\beta}_k)$ for $j \neq k$.

We have just shown that $E[\hat{\beta}] = \beta$ and

$$\hat{\beta}^{OLS} - \beta = (X'X)^{-1}X'u$$

Thus, the (conditional) variance of this estimator is then given by,

$$\begin{aligned} Var(\hat{\beta}^{OLS}|X) =& E\big[(X'X)^{-1}X'uu'X(X'X)^{-1}|X\big] \\ =& (X'X)^{-1}X'E[uu'|X]X(X'X)^{-1} \\ =& (X'X)^{-1}X'Var(u|X)X(X'X)^{-1} \end{aligned}$$

The variance of the estimator depends on the variance of the error term, the unexplained part of the model. In order to any further expressions for this variance calculation, we need to go back to the model. What assumptions did we make concerning the variance in the CLRM?

Under the assumption CLRM 3 of homoskedasticity,

$$Var(u|X) = \sigma^2 I_n = \begin{bmatrix} \sigma^2 & 0 & ... & 0 \\ 0 & \sigma^2 & & \\ \vdots & & \ddots & \\ 0 & & & \sigma^2 \end{bmatrix}$$

the above expression simplies to

$$\begin{aligned} Var(\hat{\beta}^{OLS}|X) =& (X'X)^{-1}X'\sigma^2 I_n X(X'X)^{-1} \\ =& \sigma^2(X'X)^{-1}X'X(X'X)^{-1} \\ =& \sigma^2(X'X)^{-1} \end{aligned}$$

If we made a different assumption of heteroskedasticty (CLRM 3), then

$$Var(u|X) = \begin{bmatrix} \sigma_1^2 & 0 & ... & 0 \\ 0 & \sigma_2^2 & & \\ \vdots & & \ddots & \\ 0 & & & \sigma_n^2 \end{bmatrix} = \Omega$$

the variance matrix does not reduce to a scalar multiplied by the identity matrix. And,

$$Var(\hat{\beta}^{OLS}|X) = (X'X)^{-1}X'\Omega X(X'X)^{-1}$$

This is commonly referred to as a 'sandwich' formula, given the way $Var(u|X) = \Omega$ is sandwiched between two linear transformations. The Eicker-Huber-White estimator for heteroskedastic standard errors of $\hat{\beta}^{OLS}$ replaces $Var(u|X) = E[uu'|X]$ with $\hat{u}\hat{u}'$, the OLS residuals.

3.4 Finite-sample distribution

The finite sample distribution of the OLS estimator depends on the assumptions of the model. Under CLRM 5,

$$u|X \sim N(0, \sigma^2 I_n)$$

And we have already shown that the OLS estimator is simply a linear transformation of the error term,

$$\hat{\beta}^{OLS} = \beta + (X'X)^{-1}X'u$$

Then, using the properties of the Normal distribution[4]

$$\hat{\beta}^{OLS}|X = N(\beta, \sigma^2(X'X)^{-1})$$

assuming homoskedasticity. With heteroskedastic variance, you simply change the variance, as the assumption has no implications for biasedness.

3.5 Consistency

Recall from Lecture 2 that an estimator is consistent if it converges in probability to the parameter. In this case, we want to show that

$$\hat{\beta}^{OLS} \to_p \beta \qquad \text{as} \qquad n \to \infty$$

Using the derivation $\hat{\beta}^{OLS} - \beta = (X'X)^{-1}X'u$, we need to show that $(X'X)^{-1}X'u \to_p 0$. To emphasize the fact that $\hat{\beta}$ is a function of the sample size, I am going to switch to the notation $\hat{\beta}_n$ for this section.

For the consistency of the OLS estimator we require a few assumptions,

- CLRM 1: linear in parameters
- CLRM 2^b: uncorrelatedness, $E[X_i u_i] = 0$
- CLRM 4: $rank(X) = k$
- CLRM 6: data is iid
- (NEW) CLRM 7: $E[X_i X_i']$ is a finite, positive-definite matrix.

[4] If $Y \sim N(\mu, \Sigma)$, $Y \in \mathbf{R}^k$, then $AY + b \sim N(A\mu + b, A\Sigma A')$ for any non-random $m \times k$ A-matrix and $m \times 1$ b-vector.

We begin by re-writing the expression, $\hat{\beta}^{OLS} = \beta + (X'X)^{-1}X'u$ in summation notation and then scaling by n,

$$\hat{\beta}_n = \beta + \left(\sum_{i=1}^{n} X_i X_i'\right)^{-1} \sum_{i=1}^{n} X_i u_i = \beta + \left(n^{-1}\sum_{i=1}^{n} X_i X_i'\right)^{-1} n^{-1}\sum_{i=1}^{n} X_i u_i$$

By the WLLN,[5]

$$n^{-1}\sum_{i=1}^{n} X_i u_i \to_p E[X_1 u_1] = 0$$

Similarly, by WLLN, underpinned by finiteness of $E[X_1 X_1']$ (CLRM 7),

$$n^{-1}\sum_{i=1}^{n} X_i X_i' \to_p E[X_1 X_1']$$

Since $E[X_1 X_1']$ is also positive definite (CLRM 7), then by Slutzky's Theorem

$$\left(n^{-1}\sum_{i=1}^{n} X_i X_i'\right)^{-1} \to_p \left(E[X_1 X_1']\right)^{-1}$$

Hence, by Slutzky's Theorem, which says that the product of two consistent estimators converges in probability to the product of their targets,

$$(X'X)^{-1}X'u \to_p \left(E[X_1 X_1']\right)^{-1} E[X_1 u_1] = 0$$

Thus,

$$p\lim(\hat{\beta}_n) = \beta$$

! Important

Scaling each term by n is very important, as without it, both terms do not have a finite mean. Consider, under iid,

$$E\left[\sum_{i=1}^{n} X_i X_i'\right] = nE[X_1 X_1']$$

[5]The assumptions are fulfilled by CLRM 2[b] and CLRM 6.

while,

$$E\left[n^{-1}\sum_{i=1}^{n} X_i X_i'\right] = E[X_1 X_1']$$

3.6 Asymptotic Distribution

To derive the asymptotic distribution of the OLS estimator we will need to apply the Central Limit Theorem. We will need to scale by $\sqrt{n}$, to derive the distribution of,

$$\sqrt{n}(\hat{\beta}_n - \beta)$$

Recall from Lecture 2 that by Cramer's Convergence Theorem, $Y_n X_n \to_d cX$ where $X_n \to_d$ and $Y_n \to_p c$. This result holds for the case where Y_n is a random matrix.

$$\sqrt{n}(\hat{\beta}_n - \beta) = \left(n^{-1}X'X\right)^{-1} n^{-1/2} Xu$$

We have already established that,

$$\left(n^{-1}X'X\right)^{-1} \to_p \left(E[X_1 X_1']\right)^{-1}$$

under assumptions CLRM 1, 2$^{\text{b}}$, 6, and 7.

We therefore need to consider the asymptotic distribution of $n^{-1/2}Xu$. By CLRM 2$^{\text{b}}$ $E[X_1 u_1] = 0$, fulfilling one of the CLT conditions. We then need the second moment to be finite: $Var(X_1 u_1) = E[u_1^2 X_1 X_1']$. This is a $k \times k$ matrix.

We will need to make some additional assumptions:

- (NEW) CLRM 8: $E[u_1^2 X_1 X_1']$ is a finite positive-definite matrix.[6]

Under assumptions CLRM 1, 2, 6, and 8, by CLT,

$$n^{-1/2}Xu \to N(0, E[u_1^2 X_1 X_1'])$$

There,

$$\sqrt{n}(\hat{\beta}_n - \beta) \to_d N\left(0, \left(E[X_1 X_1']\right)^{-1} E[u_1^2 X_1 X_1'] \left(E[X_1 X_1']\right)^{-1}\right)$$

[6]The finiteness of this matrix requires two assumptions:

- $E[X_{i,j}^4] < \infty$ for all $j = 1, \ldots k$ (i.e. each regressor has finite fourth moment)
- $E[U_j^4] < \infty$

These assumptions are sufficient for all elements of matrix $E[u_1^2 X_1 X_1']$ to be finite. The proof is an application of the Cauchy-Schwartz Inequality, which we haven't covered.

Under the homoskedasticity, $E[u_1^2 X_1 X_1'] = \sigma^2 E[X_1 X_1']$, giving us,

$$\sqrt{n}(\hat{\beta}_n - \beta) \to_d N\left(0, \sigma^2 (E[X_1 X_1'])^{-1}\right)$$

We can approximate the asymptotic distribution $\hat{\beta}_n$ by multiplying by $\sqrt{n}$ and replacing $(E[X_1 X_1'])^{-1}$ with the approximation $(n^{-1} X'X)^{-1}$.

$$\hat{\beta}_n \overset{a}{\sim} N(\beta, \sigma^2 (X'X)^{-1})$$

The n in the variance formula is cancelled out by the pre-multiply of $\sqrt{n}$.

3.7 Other properties

- Among the class of unbiased linear estimators of the CLRM, the OLS is the Best Linear Unbiased Estimator (BLUE). "Best" here means lowest variance. Can you show this?

4 Method of Moments

The method of moments (MM) approach is to match assumed 'moments' given by the model with their sample analogue. This is a very general approach and is used extensively in applied macroeconomics, where a structural model gives rise to moments between economic variables that can be matched in the data.

A general principle of MM is that the number of moments, m, must be $\geq k$, the number of parameters being estimated. This akin to saying, the number of equations must be greater or equal to the number of variables being solved for. If the number of moments exceeds the number of parameters, we say that the model is *overidentified*. In the case of instrumental variables, overidentification allows you two test certain model assumptions.

In term 2, you will study instrumental variables which adopts a GMM approach to estimation. MM approaches are also used extensively in time series. For now, we will apply the MM approach to the CLRM.

4.1 General setup

The observed data is given by, $W_1, ..., W_n$, read W_i is a p-dimension random vector. Let $g(W_i, \theta)$ be a l-dimension function (i.e. $\in \mathbf{R}^l$) and $\theta \in \mathbf{R}^k$:

$$g(W_i, \theta) = \begin{bmatrix} g_1(W_i, \theta) \\ \vdots \\ g_l(W_i, \theta) \end{bmatrix}$$

We assume that the true value of the parameter $\theta_0 \in \Theta \subset \mathbf{R}^k$ satisifies the condition,

$$E[g(W_i, \theta_0)] = 0$$

We say that the model is *identified* if there is a unique solution to the above equations. That is, $E[g(W_i, \theta)] = E[g(W_i, \tilde{\theta})] = 0 \;\Rightarrow\; \theta = \tilde{\theta}$. A necessary condition for identifiction is $l \geq k$; i.e. the number of equations is at least as large as the number of unknown parameters. A model can be underidentified, which typically means that there is not a unique solution for some of the parameters.

4.2 Estimator

The basic principle of MM estimation is to replacing the expectation operator with the average function and solve for $\hat{\theta}$.

$$n^{-1} \sum_{i=1}^{n} g(W_i, \hat{\theta}^{MM}) = 0$$

However, this only works when $l = k$ (exactly identified cases). For $l > k$ (overidentified cases) there is no unique vector that solves all l equations.

The *Generalized* Method of Moments (GMM) approach applies a set of weights to the minimization problem. Let A_n be a $l \times l$ weight matrix, such that $A_n \to_p A$. Then,

$$\hat{\theta}^{GMM} = \arg\min_{\theta \in \Theta} \left\| A_n n^{-1} \sum_{i=1}^{n} g(W_i, \hat{\theta}^{MM}) \right\|^2$$

Here, $||v||$ denotes the Euclidean norm of vector v: $||v|| = \sqrt{v'v}$.

4.3 Application to CLRM

Assumption CLRM 2$^{\text{b}}$ tells us that the regressors are uncorrelated with the error term.

$$E[X_i u_i] = 0$$

This is the moment that gives rise to identification in the CLRM. Given CLRM 1, we can replace the error term in the above moment with $Y_i - X_i'\beta$.

$$E[X_i(Y_i - X_i'\beta)] = 0$$

Thus, the $g(W_i, \beta) = X_i(Y_i - X_i'\beta)$ for $W_i = [Y_i, X_i']'$.

How many equations are there? Recall, X_i is a k-dimension random vector. So,

$$E[X_i(Y_i - X_i'\beta)] = \begin{bmatrix} E[X_{i1}(Y_i - X_i'\beta)] \\ E[X_{i2}(Y_i - X_i'\beta)] \\ \vdots \\ E[X_{ik}(Y_i - X_i'\beta)] \end{bmatrix} = 0$$

The are k-moments (or equations), meaning that we can estimate *up to* k parameters. In this instance, we have a failure of identification if $rank(E[X_iX_i']) < k$; which is required for the invertibility of $E[X_iX_i']$. This condition is met by assumption CLRM 4; ensuring exact identification.

The MM estimator for the CLRM is then given by the solution to,

$$n^{-1}\sum_{i=1}^{n} X_i(Y_i - X_i'\widehat{\beta}^{MM}) = 0$$

The solution is equivalent to the OLS estimator:

$$\widehat{\beta}^{MM} = \left(n^{-1}\sum_{i=1}^{n} X_iX_i'\right)^{-1} n^{-1}\sum_{i=1}^{n} X_iY_i = (X'X)^{-1}X'Y = \widehat{\beta}^{OLS}$$

4.4 Consistency

As we have already established the consistency of the OLS estimator, we will briefly review the case of the GMM estimator here. A more detailed discussion will be provided in term 2.

Recall, the assumption $A_n \to_p A$. Then,

$$\left\| A_n n^{-1}\sum_{i=1}^{n} g(W_i,\theta) \right\|^2 \to_p \left|\left| AE[g(W_i,\theta)] \right|\right|^2$$

In instance where identification is exact/unique, $E[g(W_i,\theta)] = 0 \iff \theta = \theta_0$. Which is to say that the true value of θ is the unique minimizer.

Then,

$$\begin{aligned} \hat{\theta}^{GMM} =& \arg\min_{\theta\in\Theta} \left\| A_n n^{-1}\sum_{i=1}^{n} g(W_i,\theta) \right\|^2 \\ \to_p & \arg\min_{\theta\in\Theta} \left|\left| AE[g(W_i,\theta)] \right|\right|^2 \\ =& \theta_0 \end{aligned}$$

The formal proof requires a number of additional regularity assumptions; including, the compactness of Θ.

4.5 Asymptotic Normality

The GMM estimator is asymptotically normal.

$$\sqrt{n}(\hat{\theta}^{GMM} - \theta_0) \to_d N(0, V)$$

where,

$$\begin{aligned} V =& (Q'A'AQ)^{-1}QA'A\Omega A'AQ(Q'A'AQ)^{-1} \\ Q =& E\left[\frac{\partial g(W_i, \theta_0)}{\partial \theta'}\right] \\ \Omega =& E[g(W_i, \theta)g(W_i, \theta)'] \end{aligned}$$

Where does this result come from? We won't go through the proof in details. However, it starts from the FOCs. The GMM estimator solves,

$$\Bigg[\underbrace{n^{-1}\sum_{i=1}^{n}\frac{\partial g(W_i, \hat{\theta}^{GMM})}{\partial \theta'}}_{Q_n(\hat{\theta}^{GMM})}\Bigg]' A_n' A_n n^{-1}\sum_{i=1}^{n} g(W_i, \hat{\theta}^{GMM}) = 0$$

In the above expression, the matrix of derivatives will converge (under some regularity conditions) in probability: $Q_n(\hat{\theta}^{GMM}) \to_p Q$. Second, since $E[g(W_i, \theta)] = 0$, by CLT we know,

$$n^{-1/2}\sum_{i=1}^{n} g(W_i, \theta_0) \to_d N(0, \underbrace{E[g(W_i, \theta)g(W_i, \theta)']}_{\Omega})$$

We can therefore see where the components of the variance come from. The proof requires a bit more work. First, we need the distribution of $\sqrt{n}(\hat{\theta}^{GMM} - \theta_0)$, not $n^{-1/2}\sum_{i=1}^{n} g(W_i, \theta_0)$. Second, the FOCs contain $n^{-1}\sum_{i=1}^{n} g(W_i, \hat{\theta}^{GMM})$ and not $n^{-1}\sum_{i=1}^{n} g(W_i, \theta_0)$.

This is resolve using a mean value expansion:

$$g(W_i, , \hat{\theta}^{GMM}) = g(W_i, \theta_0) + \frac{\partial g(W_i, \hat{\theta}^*)}{\partial \theta'}(\hat{\theta}^{GMM} - \theta_0)$$

Plugging this expansion into the FOCs, you can rearrange to solve,

$$\sqrt{n}(\hat{\theta}^{GMM} - \theta_0) = -[Q_n(\hat{\theta}^{GMM})' A_n' A_n Q_n(\hat{\theta}^*)]^{-1} Q_n(\hat{\theta}^{GMM})' A_n' A_n n^{-1/2}\sum_{i=1}^{n} g(W_i, \theta_0)$$

Since $\hat{\theta}^*$ is a mean value, it is a also consistent and $Q_n(\hat{\theta}^*) \to_p Q$.

4.6 Additional comments

- The targetted moments may be highly non-linear. For example, the Lucas Model pins down the rate of return on a risky asset $R_{j,t}$ using the relative utility of consumption today and tomorrow. The equilibrium condition for assets $j = 1, ..., m$ is given by,

$$E\Big[\underbrace{\delta\Big(\frac{C_{t+1}}{C_t}\Big)^{-\alpha}(1+R_{j,t})-1}_{g(W_{j,t},\theta)}\Big]=0$$

  where $W_{j,t} = [C_t, C_{t+1}, R_{j,t}]'$ and $\theta = [\alpha, \delta]'$. As the moment must hold for each asset, θ is identified so long as $m \geq 2$.

  Given the non-linearity of the g-function, there is no closed for solution. Instead, the GMM estimator must be solved for using numerical optimization.

- Some (macroeconomic) models are not identified (or underidentified). For example, a simple RBC model (with a random government component) yields the following moment condition from Euler equation,[7]

$$E\Big[\underbrace{\beta\frac{C_{t+1}}{C_t}(f_K+(1-\delta))-1}_{g(W_i,\theta)}\Big]=0$$

  where $W_{j,t} = [C_t, C_{t+1}, f_K]$ and $\theta = [\beta, \delta]$. In this application, there is only a single moment but two unknown parameters. For this reason, you will need to find an additional instrument that introduces an additional moment to identify both parameters.

5 Maximum Likelihood

Maximum Likelihood (ML or MLE) are a general class of estimators that exploit a knowledge of the underlying distribution of unobservables in the model. As the name suggests, the goal will be to maximize the likelihood (i.e. probability) of observing the a given sample of data, given the assumed distribution of the data, governed by a fixed set of parameters.

[7] This example is taken from Canova (2007, ch. 5, p. 167).

5.1 General setup

Consider an iid random sample of data: $W_1, ..., W_n$. We will assume that the data is drawn from a **known** distribution, $f(w_i; \theta)$, where $\theta \in \Theta \subset \mathbf{R}^k$ is an unknown vector of population parameters.[8]

> **Notation**
>
> The notation used to describe ML estimation varies quite a bit across texts. One key difference appears to be how to denote a parameterized distribution. The density function, $f(w_i; \theta)$, is the density at w_i (the realized value for observation i), where the distribution is *parameterized by* θ. Some texts use the conditional notation, $f(w_i|\theta)$, as the distribution depends on θ. However, probabilistic conditions tend to be based on random variables and not non-random parameters. I found this StackExchange discussion on the topic quite interesting. Needless to say, there is much disagreement and notation appears to differ across Mathematics and Statistics, and among the Statisticians, between frequentists and Bayesians. Even the Wikipedia page on MLE uses a combination of the two notations. I will use ';'; which also happens to be the notation used by Wooldridge (2010).

As the sample is iid, the joint density (or pdf) of the realized observations is given by the product of marginals,

$$f(w; \theta) = \prod_{i=1}^{n} f(w_i; \theta)$$

This is referred to as the likelihood function.

Suppose $W_i = [Y_i, X_i']'$, a vector contain a single outcome variable and a set of covariates. We can then define the joint conditional likelihood as,

$$\begin{aligned} \ell_i(\theta) &= f(Y_i|X_i; \theta) \\ L_n(\theta) &= \prod_{i=1}^{n} f(Y_i|X_i; \theta) \end{aligned}$$

Here my notation differs from Wooldridge (2010), who uses $\ell_i(\theta)$ to denote the conditional log-likelihood for observation i (see Wooldridge 2010, 471). Take note of the fact that the likelihood function is a random function of θ, since it depends on the random variables $W_i = [Y_i, X_i']'$.[9]

[8] Θ is a parameter space and is *typically* assumed to be compact: a closed and bounded subset of Euclidean space.

[9] You may also see the equivalent notation $L(W_i; \theta) \equiv L_n(\theta)$. The subscript-$n$ implies that the function depends on the sample.

5.2 Estimator

The goal of ML estimation is to solve the value of $\hat{\theta}$ that maximizes the likelihood of observing the data.

$$\hat{\theta}^{ML} = \arg\max_{\theta} L_n(\theta)$$

In practice, we apply a monotonic transformation to the likelihood function. By taking the log of the likelihood, the product of marginal distributions becomes a sum. As the transformation is monotonic, the solution to the above problem is equivalent to the solution to,

$$\hat{\theta}^{ML} = \arg\max_{\theta} n^{-1} \log L_n(\theta) = \arg\max_{\theta} n^{-1} \sum_{i=1}^{n} \log \ell_i(\theta)$$

In addition, the division by n makes this problem the sample analogue of,

$$\max_{\theta \in \Theta} E[\log \ell_i(\theta)]$$

It turns out that the true value of the parameter, θ_0, is the solution to the above problem [see Wooldridge (2010), pp. 473].[10] We will prove this for the unconditional case when we discuss consistency of ML.

Assuming a continuous, concave density function, we can solve for $\hat{\theta}^{ML}$ using first-order conditions.

$$\frac{1}{n} \frac{\partial \log L_n(\hat{\theta})}{\partial \theta} = \frac{1}{n} \sum_{i=1}^{n} \frac{\partial \log \ell_i(\hat{\theta})}{\partial \theta} = n^{-1} S(\hat{\theta}) = 0$$

The *vector* of partial derivatives is referred to as the score function: $S(\theta)$. When evaluated at the ML estimator, the score function is 0. This is a k-dimensional vector in which row is the partial derivative with respect to θ_k.

5.3 Application to CLRM

Under CLRM 5, $U|X \sim N(0, \sigma^2 I_n)$. Together with CLRM 1 and 6, we know the conditional distribution of Y.

$$Y_i|X_i \sim_{iid} N(X_i'\beta, \sigma^2)$$

[10] This is actually a non-trivial issue and beyond the scope of this module. As noted by Wooldridge (2010), we can arrive at the ML estimator by picking the value of θ to maximize the joint likelihood. However, this approach assumes that the true value of $\theta \in \Theta$, θ_0, maximizes the joint likelihood. This is not immediately evident. Once established, we have a more robust basis of the ML estimator.

where $X_i'\beta$ is the conditional mean of Y_i. Therefore, the conditional likelihood of the data is given by,

$$L_n(\beta, \sigma^2) = \prod_{i=1}^{n} \left[\frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2}\left(\frac{Y_i - X_i'\beta}{\sigma} \right)^2 \right) \right]$$

:::{.callout-important} We are working with the *conditional* likelihood. To define the likelihood of observing the entire sample, $W_1, ..., W_n$, we would also need to consider the distribution of X_i. We would then define $f(W_i; \theta) = f(Y_i|X_i; \theta) \cdot f(X_i)$, where $f(X_i)$ may be parameterized by its own set of parameters. θ is the set of parameters that parameterize the conditional distribution of $Y|X$. ::: Taking the log transformation and divide by n, we get,

$$n^{-1} \log L_n(\beta, \sigma^2) = -\frac{1}{2}\log(\sigma^2) - \frac{1}{2}\log(2\pi) - \frac{1}{2n\sigma^2}\sum_{i=1}^{n}(Y_i - X_i'\beta)^2$$

It should be immediately clear that maximizing this expression will be equivalent to minimizing the sum of squared errors.

Consider the FOC's set to 0 at the optimal point. First, w.r.t. β,

$$\begin{aligned}
\frac{\partial n^{-1} \log L_n(\hat{\beta}, \hat{\sigma}^2)}{\partial \beta} =& -\frac{1}{n\sigma^2}\sum_{i=1}^{n} X_i(Y_i - X_i'\hat{\beta}) = 0 \\
\Rightarrow \hat{\beta}^{ML} =& \left(\sum_{i=1}^{n} X_i X_i' \right) \sum_{i=1}^{n} X_i Y_i \\
=& (X'X)^{-1} X'Y
\end{aligned}$$

In the case of the CLRM, $\hat{\beta}^{ML} = \hat{\beta}^{MM} = \hat{\beta}^{OLS}$.

Second, w.r.t. σ^2,

$$\begin{aligned}
\frac{\partial n^{-1} \log L_n(\hat{\beta}, \hat{\sigma}^2)}{\partial \sigma^2} =& -\frac{1}{2\hat{\sigma}^2} + \frac{1}{n2\hat{\sigma}^4}\sum_{i=1}^{n}(Y_i - X_i'\hat{\beta})^2 = 0 \\
\Rightarrow \hat{\sigma}^2_{ML} =& n^{-1}\sum_{i=1}^{n}(Y_i - X_i'\hat{\beta})^2
\end{aligned}$$

This estimator for the variance is consistent, but biased for small samples. This is because it scales by n and not $n - k$, a distinction that is ignorable as $n \to \infty$. For this reason, when conducting inference you should use the asymptotic distribution of the ML estimator.

5.4 Consistency

The ML estimator is consistent. This can be shown in a couple of steps. To simplify notation we will examine the proof for the unconditional likelihood, but the same will hold for the conditional. The proof will require Jensen's inequality:

Theorem 5.1. *For $h(\cdot)$ concave, then $E[h(X)] \leq h(E[X])$.*

Proof. By the WLLN, for ALL values of $\theta \in \Theta$,

$$\begin{aligned} n^{-1}\sum_{i=1}^{n} \log f(W_i;\theta) \rightarrow_p & E\big[\log f(W_i;\theta)\big] \\ &= \int (\log f(w;\theta)) f(w;\theta_0) dw \end{aligned}$$

Note, an important distinction in the last line: the expectation is based on the density function parameterized by the true value, θ_0. This is because the data is generated by the true density.

We have convergence for ALL values of θ, but now need to establish convergence to the θ_0. Consider the difference,

$$\begin{aligned} E\big[\log f(W_i;\theta)\big] - E\big[\log f(W_i;\theta_0)\big] =& E\left[\log \frac{f(W_i;\theta)}{f(W_i;\theta_0)}\right] \\ \leq& \log\left[\frac{f(W_i;\theta)}{f(W_i;\theta_0)}\right] \qquad \text{by Jensen's} \\ =& \log \int \left(\frac{f(w;\theta)}{f(w,\theta_0)}\right) f(w;\theta_0) dw \\ =& \log \int f(w;\theta) dw \\ =& \log 1 \\ =& 0 \end{aligned}$$

The inequality can be made strict if we assume that $Pr(f(W_i;\theta_0) \neq f(W_i;\theta)) > 0 \ \forall \theta \neq \theta_0$. This ensures that θ_0 is a *unique* solution. Since the difference is ≤ 0, it follows that,

$$\theta_0 = \arg\max_{\theta \in \Theta} E[\log f(W_i;\theta)]$$

Which implies,

$$\begin{aligned} \widehat{\theta}_n^{ML} &= \arg\max_{\theta \in \Theta} n^{-1} \log L_n(\theta) \\ &\rightarrow_p \arg\max_{\theta \in \Theta} E\big[\log f(W_i,\theta)\big] \\ &= \theta_0 \end{aligned}$$

□

5.5 Asymptotic Normality

The ML estimator is asymtotically normal. We will not prove this result, but rather focus on the form of the asymptotic variance and its estimator. The proof uses the Mean Value Theorem and CLT.

$$\sqrt{n}(\hat{\theta}_n^{ML} - \theta_0) \to_d N(0, V)$$

where $V = [J(\theta_0)]^{-1}$. $J(\theta)$ is referred to as the *information matrix*, given by the expectation of the (Hessian) matrix of second-order derivatives:

$$J(\theta) = -E\left[\frac{\partial^2}{\partial\theta\partial\theta'}\log f(W_i, \theta_0)\right]$$

$[nJ(\theta_0)]^{-1}$ is used to approximate the variance, but since J is not observed, it must be estimated. This is done by replacing the expectation in the information matrix with sample average:

$$\hat{V}_H = \left[\frac{1}{n}\sum_{i=1}^{n}\frac{\partial^2}{\partial\theta\partial\theta'}\log f(W_i, \hat{\theta})\right]^{-1}$$

5.6 Additional comments

- ML estimators are invariant. If $\hat{\theta}$ is the ML-estimator for θ, then $\ln(\hat{\theta})$ is the ML-estimator for $\ln(\theta)$.
- In general, there are no closed form solutions for ML estimators; the CLRM being one exception. For this reason, ML estimation requires numerical optimization.
- The ML estimator is efficient. That is, its variance is at least as small as any other consistent (and asymptotically normal) estimator.
- ML estimators require as to know the true PDF, up to its parameters. For example, probit (logit) models assumes that the error term is normally (logistically) distributed.
- In some cases, the estimator may be consistent even if the PDF is misspecified. As is the case for the OLS estimator of the linear model. These estimators are referred to as a quasi-ML estimators.

References

Cameron, A Colin, and Pravin K Trivedi. 2005. *Microeconometrics: Methods and Applications*. Cambridge university press.

Verbeek, Marno. 2017. *A Guide to Modern Econometrics*. John Wiley & Sons.

Wooldridge, Jeffrey M. 2010. *Econometric Analysis of Cross Section and Panel Data*. MIT press.