Advanced Econometrics II
TA Session Problems No. 4

Agnieszka Borowska

27.01.2015

Note: this is only a draft of the problems discussed on Tuesday and might contain some typos or more or less imprecise statements. If you find some, please let me know.

1. ML Basic Concepts
2. Asymptotic Efficiency of the ML Estimator

ML Basic Concepts

Recall

$$
\begin{aligned}
f(y,\theta) &= \prod_{t=1}^{n} f_t(y_t,\theta), && \text{(joint PDF)}\\
\ell(y,\theta) &\equiv \log f(y,\theta) = \sum_{t=1}^{n} \underbrace{\ell_t(y_t,\theta)}_{\text{contribution}}, && \text{(loglikelihood)}\\
g(y,\theta) &= \big(g_i(y,\theta)\big)_{i=1,\dots,k}, && \text{(gradient/score vector)}\\
g_i(y,\theta) &\equiv \frac{\partial \ell(y,\theta)}{\partial\theta_i} = \sum_{t=1}^{n}\frac{\partial \ell_t(y_t,\theta)}{\partial\theta_i}, && \text{(typical element of } g(y,\theta))\\
\mathbf{H}(\theta) &= \Big(\frac{\partial^2\ell(\theta)}{\partial\theta_i\partial\theta_j}\Big)_{i,j=1,\dots,k}, && \text{(Hessian matrix)}\\
G(y,\theta) &= \big(G_{ti}(y^t,\theta)\big)_{i=1,\dots,k,\, t=1,\dots,n}, && \text{(matrix of contributions to the gradient)}\\
G_{ti}(y^t,\theta) &\equiv \frac{\partial \ell(y^t,\theta)}{\partial\theta_i} = \sum_{t=1}^{n}\frac{\partial \ell_t(y_t,\theta)}{\partial\theta_i}, && \text{(typical element of } G(y,\theta))\\
g_i(y,\theta) &= \sum_{t=1}^{n} G_{ti}(y^t,\theta), && \text{(10.27)}\\
\mathbf{I}(\theta) &\equiv \sum_{t=1}^{n} \underbrace{\mathbf{I}_t(\theta)}_{\text{contribution}} && \text{(information matrix)}\\
\mathbf{I}_t(\theta) &= \Big(\mathbb{E}_\theta\big(G_{ti}(y^t,\theta)G_{ti}(y^t,\theta)\big)\Big)_{i,j=1,\dots,k}, && \text{(covariance matrix of } G_t(y^t,\theta))\\
& && (G_t(y^t,\theta) - t^{th} \text{ row of } G(y,\theta))\\
\mathbf{I}_t(\theta) &\equiv \mathbb{E}_\theta\big(g(y,\theta)g^T(y,\theta)\big) && \text{(covariance matrix of the score vector)} \qquad (\star)\\
\mathcal{I}(\theta) &\equiv \operatorname{plim}_{n\to\infty}\frac{1}{n}\mathbf{I}(\theta), && \text{(asymptotic information matrix)}\\
\mathcal{H}(\theta) &\equiv \operatorname{plim}_{n\to\infty}\frac{1}{n}\mathbf{H}(\theta), && \text{(asymptotic Hessian matrix)}\\
\mathcal{I}(\theta) &= -\mathcal{H}(\theta), && \text{(information matrix equality)}
\end{aligned}
$$

Below, we will prove $(\star)$.

Covariance matrix of the gradient vector

DM, 10.5

Prove that the definition

$$\mathbf{I}(\theta) \equiv \sum_{t=1}^{n} \mathbf{I}_t(\theta) = \sum_{t=1}^{n} \mathbb{E}_\theta(G_t^T(y,\theta)G_t(y,\theta)) \tag{10.31}$$

of the information matrix is equivalent to the definition

$$\mathbf{I}(\theta) = \mathbb{E}_\theta(g(y,\theta)g^T(y,\theta)).$$

Hint: Use the result

$$\begin{aligned}\mathbb{E}_\theta(G_{ti}^T(y^t,\theta)G_{sj}(y^s,\theta)) &= \mathbb{E}_\theta\left(\mathbb{E}_\theta(G_{ti}(y^t,\theta)G_{sj}(y^s,\theta)|y^t)\right)\\ &= \mathbb{E}_\theta\left(\mathbb{E}_\theta(G_{ti}(y^t,\theta)G_{sj}(y^s,\theta)|y^t)\right) = 0.\end{aligned} \tag{10.30}$$

The task is to show that the following relation holds

$$\mathbb{E}_\theta(g(y,\theta)g^T(y,\theta)) = \sum_{t=1}^{n} \mathbb{E}_\theta(G_t^T(y,\theta)G_t(y,\theta)).$$

Since by (10.27) each element of the gradient vector is the sum of the elements of one of the columns of the matrix of contributions to the gradient,

$$g_i(y,\theta) = \sum_{t=1}^{n} G_{ti}(y^t,\theta),$$

we can also write

$$g(y,\theta) = \sum_{t=1}^{n} G_t^T(y^t,\theta).$$

Then we easily obtain the required result by writing

$$\begin{aligned}\mathbb{E}_\theta(g(y,\theta)g^T(y,\theta)) &= \mathbb{E}_\theta\left(\left(\sum_{t=1}^{n} G_t^T(y^t,\theta)\right)\left(\sum_{s=1}^{n} G_s(y^t,\theta)\right)\right)\\ &= \mathbb{E}_\theta\left(\sum_{t=1}^{n}\sum_{s=1}^{n} G_t^T(y^t,\theta)G_s(y^t,\theta)\right)\\ &\overset{(*)}{=} \mathbb{E}_\theta\left(\sum_{t=1}^{n} G_t^T(y^t,\theta)G_t(y^t,\theta)\right)\\ &= \sum_{t=1}^{n} \mathbb{E}_\theta(G_t^T(y,\theta)G_t(y,\theta)),\end{aligned}$$

where in $(*)$ we used that by (10.30) $\forall t \neq s$

$$\mathbb{E}_\theta\left(\sum_{t=1}^{n}\sum_{s=1}^{n} G_t^T(y^t,\theta)G_s(y^t,\theta)\right) = 0.$$

Asymptotic Efficiency of the ML Estimator

Below we will prove the ML estimator asymptotically achieves the **Cramér-Rao lower bound** – which is one of many of its attractive features. Notice, however, that it happens, in general, only **asymptotically**, as, in general, ML estimators are **not unbiased** (but are asymptotically unbiased).

To start with, consider any *other* root-n consistent and asymptotically unbiased estimator, which we will denote $\tilde{\theta}$ ($\hat{\theta}$ will stand for the MLE). It can be shown that[1]

$$\text{plim}_{n\to\infty} \sqrt{n}\big(\tilde{\theta} - \theta_0\big) = \text{plim}_{n\to\infty} \sqrt{n}\big(\hat{\theta} - \theta_0\big) + v,$$

where v is a random, zero mean k-vector, uncorrelated with $\text{plim}_{n\to\infty} \sqrt{n}\big(\hat{\theta} - \theta_0\big)$. Hence, taking Var's on the both sides of the above expression gives us the following result

$$\text{Var}\Big(\text{plim}_{n\to\infty} \sqrt{n}\big(\tilde{\theta} - \theta_0\big)\Big) = \text{Var}\Big(\text{plim}_{n\to\infty} \sqrt{n}\big(\hat{\theta} - \theta_0\big)\Big) + \text{Var}\big(v\big).$$

Because $\text{Var}\big(v\big)$ is PSD, it follows that the asymptotic covariance matrix of the *other* estimator $\tilde{\theta}$ must be larger than the one of the MLE $\hat{\theta}$. This asymptotic efficiency result is an asymptotic version of Cramér-Rao lower bound (which applies to any **unbiased** estimator), stating that the covariance matrix of an unbiased estimator cannot be smaller that $\mathbf{I}^{-1}$. In case of the MLE the latter is asymptotically equal to its covariance matrix. So we can say that the MLE *attains* the ramér-Rao lower bound.

DM, 10.12

Let $\hat{\theta}$ denote any unbiased estimator of the k parameters of a parametric model fully specified by the loglikelihood function $\ell(\theta)$. The unbiasedness property can be expressed as the following identity:

$$\mathbb{E}_\theta \tilde{\theta} = \int L(y,\theta)\tilde{\theta} dy = \theta \tag{1}$$

By using the relationship between $L(y,\theta)$ and $\ell(y,\theta)$ and differentiating this identity with respect to the components of θ, show that

$$Cov_\theta\big(g(\theta), (\tilde{\theta} - \theta)\big) = \mathbb{I}$$

where $\mathbb{I}$ is a $k \times k$ identity matrix, and the notation Cov_θ indicates that the covariance is to be calculated under the DGP characterized by θ.
Let V denote the $2k \times 2k$ covariance matrix of the $2k$-vector obtained by stacking the k components of $g(\theta)$ above the k components of $\tilde{\theta} - \theta$. Partition this matrix into 4 $k \times k$ blocks as follows:

$$V = \begin{bmatrix} V_1 & C \\ C^T & V_2 \end{bmatrix} \tag{2}$$

where V_1 and V_2 are, respectively, the covariance matrices of the vectors $g(\theta)$ and $\tilde{\theta} - \theta$ under the DGP characterized by θ. Then use the fact that V is positive semidefinite to show that the difference between V_2 and $\mathbb{I}^{-1}(\theta)$, where $\mathbb{I}(\theta)$ is the (finite-sample) information matrix for the model, is a positive semidefinite matrix.

First, notice that in (1) the RHS is simply θ, which differentiated wrt θ will be simply a $k \times k$ identity matrix $\mathbb{I}$. For the LHS, we had

$$\ell(y,\theta) = \log L(y,\theta) \quad \Leftrightarrow \quad L(y,\theta) = \exp\big(\ell(y,\theta)\big),$$

so

$$\frac{\partial L(y,\theta)}{\partial \theta} = L(y,\theta) \underbrace{\frac{\partial \ell(y,\theta)}{\partial \theta}}_{g(y,\theta)}.$$

Hence, differentiation of (1) wrt θ gives

$$\int L(y,\theta)\tilde{\theta} g(y,\theta) dy = \mathbb{I}. \tag{3}$$

[1] Cf. Section 10.4 in DM.

Next, notice that $L(\cdot,\theta)$, i.e. a function of y with θ fixed, is the PDF of y, which means that the integral of two functions wrt $dL(y,\theta)$ (with θ fixed) is just the covariance of these functions wrt the distribution characterized by θ. Hence, (3) describes the covariance matrix of $\tilde{\theta}$ and $g(y,\theta) := g(\theta)$, i.e.

$$\mathrm{Cov}_\theta\big(g(\theta),\tilde{\theta}\big) = \int g(y,\theta)\tilde{\theta}L(y,\theta)dy = \mathbb{I}.$$

However, we know that[2] $\mathbb{E}_\theta g(\theta) = 0$, and by the unbiasedness assumption $\mathbb{E}_\theta\tilde{\theta} = \theta$, so

$$\begin{aligned}
\mathbb{I} &= \mathrm{Cov}_\theta\big(g(\theta),\tilde{\theta}\big)\\
&\overset{\text{def}}{=} \mathbb{E}_\theta\Big(\big(g(\theta) - \mathbb{E}_\theta g(\theta)\big)\big(\tilde{\theta} - \mathbb{E}_\theta\tilde{\theta}\big)\Big)\\
&= \mathbb{E}_\theta\Big(g(\theta)\big(\tilde{\theta} - \theta\big)\Big)\\
&= \int g(y,\theta)(\tilde{\theta} - \theta)L(y,\theta)dy\\
&= \mathrm{Cov}_\theta\big(g(\theta),(\tilde{\theta} - \theta)\big),
\end{aligned}$$

which is the first of the required results.

Second, consider V as in (2). By definition of V, its off-diagonal blocks C and C^T are given by the covariance matrix of $g(\theta)$ and $\tilde{\theta} - \theta$, which we have just shown is simply the identity matrix. As far as V_1 is concerned, it is the covariance matrix of the gradient vector $g(\theta)$ and we know from the previous exercise that it is equal to $\mathbf{I}(\theta)$, the information matrix. Finally, V as a covariance matrix needs to be PSD. Hence, we arrive at

$$V = \begin{bmatrix}\mathbf{I}(\theta) & \mathbb{I}\\ \mathbb{I} & V_2\end{bmatrix} \geq 0,$$

which implies that also

$$V^{-1} \geq 0$$

as well as each of the diagonal blocks of V^{-1} must also be PSD. To complete the exercise we need to show that

$$V_2 - \mathbf{I}^{-1}(\theta) \geq 0.$$

Recall the discussion about the asymptotic distribution of the Wald statistics, where we needed the inverse of the covariance matrix of the IV estimator corresponding to part of this vector[3]. We used there the following fact

$$A := \begin{bmatrix}A_1 & A_{12}\\ A_{21} & A_2\end{bmatrix} \qquad \Rightarrow \qquad A^{-1} = \begin{bmatrix}\cdot & \cdot\\ \cdot & \big(A_{22} - A_{21}A_{11}A12\big)^{-1}\end{bmatrix},$$

which applied to the current case yields

$$V^{-1} = \begin{bmatrix}\cdot & \cdot\\ \cdot & \big(V_2 - \mathbb{I}\,\mathbf{I}^{-1}(\theta)\,\mathbb{I}\big)^{-1}\end{bmatrix}.$$

So the lower diagonal block of the inverse of V, which, as we stated, is PSD, has the form

$$\big(V_2 - \mathbb{I}\,\mathbf{I}^{-1}(\theta)\,\mathbb{I}\big)^{-1} = \big(V_2 - \mathbf{I}^{-1}(\theta)\big)^{-1}.$$

This means that its inverse,

$$V_2 - \mathbf{I}^{-1}(\theta) \geq 0,$$

i.e. is PSD – which it the second of the required results.

[2]Recall that a crucial property of the matrix $G(y,\theta)$ is that **if y is generated by the GDP characterized by θ**, then the expectations of all the elements of the matrix, **evaluated at** θ, are zero – which is a consequence of the fact that all densities integrate to 1. Then, summing the expectations of the elements in each column of $G(y,\theta)$ yields that $\mathbb{E}_\theta g(y,\theta) = 0$.

[3]Week 1, slide 48.