The outline for Unit 2

UNIT 1. Introduction: The regression model. ✓

UNIT 2. Estimation principles.

2.1 Ordinary Least Squares. ✓
2.2 Maximum Likelihood.
2.3 Method of Moments.

Addendum A. A brief revision of asymptotic theory. ✓

Addendum B. Estimating the OLS estimator's distribution using bootstrap.

UNIT 3: Hypothesis testing principles.

UNIT 4: Heteroscedasticity in the regression model.

UNIT 5: Endogeneity of regressors.

The likelihood function

Suppose we have a sample of size n of the random vectors y and z. Suppose the joint density of $Y = \begin{pmatrix} y_1 & \dots & y_n \end{pmatrix}$ and $Z = \begin{pmatrix} z_1 & \dots & z_n \end{pmatrix}$ is characterized by a parameter vector ψ_0 :

$$f_{YZ}(Y, Z, \psi_0).$$

This density can be factored as

$$f_{YZ}(Y, Z, \psi_0) = f_{Y|Z}(Y|Z, \theta_0) f_Z(Z, \rho_0).$$

The *likelihood function* is just this density evaluated at other values ψ

$$L(Y, Z, \psi) = f(Y, Z, \psi), \psi \in \Psi,$$

where Ψ is a *parameter space.*

The likelihood estimator

The *maximum likelihood estimator* of ψ_0 is the value of ψ that maximizes the likelihood function, usually denoted by $\widehat{\psi}$:

$$\widehat{\psi} = \arg\max f_{YZ}(Y, Z, \psi_0) \;=\; \arg\max f_{Y|Z}(Y|Z, \theta_0) f_Z(Z, \rho_0).$$

Note that if θ_0 and ρ_0 share no elements, then the maximizer of the conditional likelihood function $f_{Y|Z}(Y|Z, \theta)$ with respect to θ is the same as the maximizer of the overall likelihood function $f_{YZ}(Y, Z, \psi) = f_{Y|Z}(Y|Z, \theta) f_Z(Z, \rho)$, for the elements of ψ that correspond to θ.

In this case, the variables Z are said to be *exogenous* for estimation of θ, and we may more conveniently work with the conditional likelihood function $f_{Y|Z}(Y|Z, \theta)$ for the purposes of estimating θ_0.

- If the n observations are independent, the likelihood function can be written as

$$L(Y|Z,\theta) = \prod_{t=1}^{n} f_t(y_t|z_t,\theta),$$

  where the f_t can be of different form.

- If this is not possible, we can always factor the likelihood into *contributions of observations*, by using the fact that a joint density can be factored into the product of a marginal and conditional (iteratively):

$$L(Y,\theta) = f(y_1|z_1,\theta)f(y_2|y_1,z_2,\theta)f(y_3|y_1,y_2,z_3,\theta)\cdots f(y_n|y_1,y_2,\ldots y_{t-n},z_n,\theta).$$

To simplify notation, define $x_t = \{y_1, y_2, ..., y_{t-1}, z_t\}$, i.e., it contains exogenous and predetermined endogenous variables.

Now the likelihood function can be written as

$$L(Y, \theta) = \prod_{t=1}^{n} f(y_t | x_t, \theta)$$

The criterion function can be defined as the average log-likelihood function:

$$s_n(\theta) = \frac{1}{n} \ln L(Y, \theta) = \frac{1}{n} \sum_{t=1}^{n} \ln f(y_t | x_t, \theta).$$

The *maximum likelihood estimator* may thus be defined equivalently as

$$\hat{\theta} = \arg\max s_n(\theta).$$

Example 1: Bernoulli trials.

Suppose that we are flipping a coin that may be biased, so the probability of a heads may not be 0.5.

May be we're interested in estimating the probability of heads. Let $y = 1(heads)$ be a binary variable that indicates whether or not a heads is observed.

The outcome of a toss is a Bernoulli random variable:

$$\begin{array}{ll} f_Y(y, p_0) & = p_0^y (1 - p_0)^{1-y} \quad \text{if } y \in \{0, 1\} \\ & = 0 \qquad\qquad\qquad\quad \text{if } y \notin \{0, 1\} \end{array}$$

So a representative term that enters to the likelihood function is:

$$f_Y(y, p) = p^y (1 - p)^{1-y}$$

and

$$\ln f_Y(y, p) = y \ln p + (1 - y) \ln (1 - p) .$$

The derivative of this is:

$$\begin{aligned}\frac{\partial \ln f_Y(y,p)}{\partial p} &= \frac{y}{p} - \frac{(1-y)}{(1-p)} \\ &= \frac{y-p}{p\,(1-p)}.\end{aligned}$$

Averaging this over a sample of size n gives:

$$\frac{\partial s_n(p)}{\partial p} = \frac{1}{n}\sum_{i=1}^{n} \frac{y_i - p}{p\,(1-p)}.$$

Setting to zero and solving gives:

$$\hat{p} = \bar{y}.$$

So it's easy to calculate the MLE of p_0 in this case.

Now imagine that we had a bag full of bent coins, each bent around a sphere of a different radius and we might suspect that the probability of a heads could depend upon the radius, v.g., $p_i \equiv p(x_i, \beta) = (1 + \exp(-x_i'\beta))^{-1}$ where $x_i = \begin{bmatrix} 1 & r_i \end{bmatrix}'$.

Now

$$\frac{\partial p_i(\beta)}{\partial \beta} = p_i \left(1 - p_i\right) x_i,$$

so

$$\begin{aligned} \frac{\partial \ln f_Y(y, \beta)}{\partial \beta} &= \frac{y - p_i}{p_i \left(1 - p_i\right)} p_i \left(1 - p_i\right) x_i \\ &= \left(y_i - p(x_i, \beta)\right) x_i. \end{aligned}$$

So the derivative of the average log likelihood function is now a set of nonlinear equations in the two unknown elements of β. There is no explicit solution for the two elements that set the equations to zero. This is common situation with ML estimators.

Consistency of MLE

Assumptions:

Compact parameter space: $\theta \in \Theta$, an open bounded subset of $\Re^K$. Maximization is over $\overline{\Theta}$, which is compact.

This implies that θ is an interior point of the *parameter space* $\overline{\Theta}$.

Uniform convergence: $s_n(\theta) \overset{u.a.s}{\rightarrow} \lim_{n\rightarrow\infty} \mathrm{E}_{\theta_0} s_n(\theta) \equiv s_\infty(\theta, \theta_0), \forall \theta \in \overline{\Theta}$.

This requires that almost sure convergence holds for all possible parameter values.

Continuity: $s_n(\theta)$ is continuous in θ, $\forall \theta \in \overline{\Theta}$. This implies that $s_\infty(\theta, \theta_0)$ is continuous in θ.

Identification: $s_\infty(\theta, \theta_0)$ has a unique maximum in its first argument.

Consistency of MLE - 2

First, $\hat{\theta}_n$ certainly exists, since a continuous function has a maximum on a compact set.

Second, for any $\theta \neq \theta_0$: $\mathrm{E}\left[\ln\left(\frac{L(\theta)}{L(\theta_0)}\right)\right] \leq \ln\left(\mathrm{E}\left[\frac{L(\theta)}{L(\theta_0)}\right]\right)$, by Jensen's inequality ($\ln(\cdot)$ is a concave function).

Jensen's inequality: If f is concave then $\mathrm{E}[f(X)] \leq f(\mathrm{E}[X])$ and if f is convex then $\mathrm{E}[f(X)] \geq f(\mathrm{E}[X])$.

Now, the expectation on the RHS is

$$\mathrm{E}\left[\frac{L(\theta)}{L(\theta_0)}\right] = \int \frac{L(\theta)}{L(\theta_0)} L(\theta_0) dy = 1,$$

since $L(\theta_0)$ is the density function of the observations.

Consistency of MLE - 3

Therefore, since $\ln(1) = 0$,

$$\mathrm{E}\left[\ln\left(\frac{L(\theta)}{L(\theta_0)}\right)\right] \leq 0,$$

or

$$\mathrm{E}\left[s_n\left(\theta\right)\right] - \mathrm{E}\left[s_n\left(\theta_0\right)\right] \leq 0.$$

Taking limits, we obtain

$$s_\infty(\theta, \theta_0) - s_\infty(\theta_0, \theta_0) \leq 0$$

except on a set of zero probability (by the uniform convergence assumption).

By the identification assumption there is a unique maximizer, then the inequality is strict if $\theta \neq \theta_0$:

$$s_\infty(\theta, \theta_0) - s_\infty(\theta_0, \theta_0) < 0, \forall \theta \neq \theta_0, \text{a.s.}$$

Consistency of MLE - 4

Suppose that θ^* is a limit point of $\hat{\theta}_n$ (any sequence from a compact set has at least one limit point). Since $\hat{\theta}_n$ is a maximizer, independent of n, we must have

$$s_\infty(\theta^*, \theta_0) - s_\infty(\theta_0, \theta_0) \geq 0.$$

These last two inequalities imply that

$$\theta^* = \theta_0, \text{a.s.}$$

Thus there is only one limit point, and it is equal to the true parameter value with probability one. In other words,

$$\lim_{n \to \infty} \hat{\theta} = \theta_0, \text{ a.s.}$$

The score function

Differentiability: Assume that $s_n(\theta)$ is twice continuously differentiable in a neighborhood of θ_0, at least when n is large enough.

To maximize the log-likelihood function, we take derivatives:

$$\begin{aligned} g_n(Y,\theta) &= D_\theta s_n(\theta) \\ &= \frac{1}{n}\sum_{t=1}^{n} D_\theta \ln f(y_t|x_x,\theta) \\ &\equiv \frac{1}{n}\sum_{t=1}^{n} g_t(\theta). \end{aligned}$$

This is the **score vector**. Note that the score function has Y as an argument, which implies that it is a random function.

The score function - 2

The ML estimator $\hat{\theta}$ sets the derivatives to zero:

$$g_n(\hat{\theta}) = \frac{1}{n}\sum_{t=1}^{n} g_t(\hat{\theta}) \equiv 0.$$

We will show that $\mathrm{E}_\theta\left[g_t(\theta)\right] = 0$, $\forall t$:

$$\begin{aligned}
\mathrm{E}_\theta\left[g_t(\theta)\right] &= \int \left[D_\theta \ln f(y_t|x_t,\theta)\right] f(y_t|x,\theta)dy_t \\
&= \int \frac{1}{f(y_t|x_t,\theta)}\left[D_\theta f(y_t|x_t,\theta)\right] f(y_t|x_t,\theta)dy_t \\
&= \int D_\theta f(y_t|x_t,\theta)dy_t.
\end{aligned}$$

The score function - 3

Under some regularity conditions on the boundedness of $D_\theta f$, we can switch the order of integration and differentiation, by the dominated convergence theorem:

$$\begin{aligned} \mathrm{E}_\theta \left[g_t(\theta) \right] &= D_\theta \int f(y_t|x_t, \theta) dy_t \\ &= D_\theta \, 1 \; = \; 0 \end{aligned}$$

where we use the fact that the integral of the density is 1.

- So $\mathrm{E}_\theta[g_t(\theta)] = 0$: *the expectation of the score vector is zero*.
- This hold for all t, so it implies that $\mathrm{E}_\theta[g_n(Y, \theta)] = 0$.

Asymptotic normality of MLE

Again, assuming that $s_n(\theta)$ is twice continuously differentiable.

Taking a first order Taylor's series expansion of $g(Y, \hat{\theta})$ about the true value θ_0 :

$$0 \equiv g(\hat{\theta}) = g(\theta_0) + (D_{\theta'} g(\theta^*)) \left(\hat{\theta} - \theta_0\right)$$

or with appropriate definitions

$$H(\theta^*) \left(\hat{\theta} - \theta_0\right) = -g(\theta_0),$$

where $\theta^* = \lambda\hat{\theta} + (1 - \lambda)\theta_0, 0 < \lambda < 1.$

Assume $H(\theta^*)$ is invertible. So

$$\sqrt{n} \left(\hat{\theta} - \theta_0\right) = -H(\theta^*)^{-1} \sqrt{n} g(\theta_0)$$

Asymptotic normality of MLE - 2

Now consider $H(\theta^*)$. This is

$$\begin{aligned} H(\theta^*) &= D_{\theta'} g(\theta^*) \\ &= D_\theta^2 s_n(\theta^*) \\ &= \frac{1}{n}\sum_{t=1}^{n} D_\theta^2 \ln f_t(\theta^*), \end{aligned}$$

where the notation

$$D_\theta^2 s_n(\theta) \equiv \frac{\partial^2 s_n(\theta)}{\partial\theta\partial\theta'}.$$

Given that this is an average of terms, it should usually be the case that this satisfies a strong law of large numbers (SLLN).

The usual wording is: Under some regularity conditions ...

Asymptotic normality of MLE - 3

Since we know that $\hat{\theta}$ is consistent, and since $\theta^* = \lambda\hat{\theta} + (1-\lambda)\theta_0$, we have that $\theta^* \overset{a.s.}{\to} \theta_0$.

Also, by the above differentiability assumption, $H(\theta)$ is continuous in θ. Given this, $H(\theta^*)$ converges to the limit of it's expectation:

$$H(\theta^*) \overset{a.s.}{\to} \lim_{n\to\infty} \mathrm{E}\left(D_\theta^2 s_n(\theta_0)\right) = H_\infty(\theta_0) < \infty$$

This matrix converges to a finite limit.

Re-arranging orders of limits and differentiation, which is legitimate given regularity conditions, we get

$$\begin{aligned} H_\infty(\theta_0) &= D_\theta^2 \lim_{n\to\infty} \mathrm{E}\left(s_n(\theta_0)\right) \\ &= D_\theta^2 s_\infty(\theta_0, \theta_0). \end{aligned}$$

Asymptotic normality of MLE - 4

We've already seen that

$$s_\infty(\theta, \theta_0) < s_\infty(\theta_0, \theta_0)$$

i.e., θ_0 maximizes the limiting objective function.

Since there is a unique maximizer, and by the assumption that $s_n(\theta)$ is twice continuously differentiable (which holds in the limit), then $H_\infty(\theta_0)$ must be negative definite, and therefore of full rank.

Therefore the previous inversion is justified, asymptotically, and we have

$$\sqrt{n}\left(\hat{\theta} - \theta_0\right) + H_\infty(\theta_0)^{-1}\sqrt{n}g(\theta_0) \xrightarrow{a.s.} 0.$$

Asymptotic normality of MLE - 5

Now consider $\sqrt{n}g(\theta_0)$. This is

$$\begin{aligned}\sqrt{n}g_n(\theta_0) &= \sqrt{n}D_\theta s_n(\theta) \\ &= \frac{\sqrt{n}}{n}\sum_{t=1}^{n} D_\theta \ln f_t(y_t|x_t,\theta_0) \\ &= \frac{1}{\sqrt{n}}\sum_{t=1}^{n} g_t(\theta_0).\end{aligned}$$

Supposing that the CLT applies:

See assumptions in previous class.

$$\mathcal{I}_\infty(\theta_0)^{-1/2}\sqrt{n}g_n(\theta_0) \xrightarrow{d} \mathcal{N}\left[0, I_K\right]$$

where

$$\mathcal{I}_\infty(\theta_0) = \lim_{n\to\infty} \mathrm{E}_{\theta_0}\left(n\left[g_n(\theta_0)\right]\left[g_n(\theta_0)\right]'\right) = \lim_{n\to\infty} \mathrm{Var}_{\theta_0}\left(\sqrt{n}g_n(\theta_0)\right).$$

Asymptotic normality of MLE - 6

This can also be written as

$$\sqrt{n} g_n(\theta_0) \xrightarrow{d} \mathcal{N}\left[0, \mathcal{I}_\infty(\theta_0)\right].$$

- $\mathcal{I}_\infty(\theta_0)$ is known as the *information matrix*.
- Finally, we get

$$\sqrt{n}\left(\hat{\theta} - \theta_0\right) \overset{a}{\sim} \mathcal{N}\left[0, H_\infty(\theta_0)^{-1} \mathcal{I}_\infty(\theta_0) H_\infty(\theta_0)^{-1}\right].$$

The MLE estimator is asymptotically normally distributed.

Definition 1: An estimator $\hat{\theta}$ of a parameter θ_0 is $\sqrt{n}$-consistent and asymptotically normally distributed if

$$\sqrt{n}\left(\hat{\theta} - \theta_0\right) \xrightarrow{d} \mathcal{N}\left(0, V_\infty\right)$$

where V_∞ is a finite positive definite matrix.

Definition 2: An estimator $\hat{\theta}$ of a parameter θ_0 is asymptotically unbiased if

$$\lim_{n\to\infty} \mathrm{E}_\theta(\hat{\theta}) = \theta.$$

Estimators that are CAN are asymptotically unbiased, though not all consistent estimators are asymptotically unbiased.

Example 2: Consider the estimator $\hat{\theta} = \begin{cases} \theta_0 & \text{with probability } 1 - \frac{1}{n} \\ n & \text{with probability } \frac{1}{n} \end{cases}$.

The information matrix equality

We will show that $H_\infty(\theta) = -I_\infty(\theta)$. Let $f_t(\theta)$ be short for $f(y_t|x_t,\theta)$

$$\begin{aligned} 1 &= \int f_t(\theta)dy, \text{ so} \\ 0 &= \int D_\theta f_t(\theta)dy = \int \left(D_\theta \ln f_t(\theta)\right) f_t(\theta)dy \end{aligned}$$

Now differentiate again:

$$\begin{aligned} 0 &= \int \left[D_\theta^2 \ln f_t(\theta)\right] f_t(\theta)dy + \int \left[D_\theta \ln f_t(\theta)\right] D_{\theta'} f_t(\theta)dy \\ &= \mathrm{E}_\theta\left[D_\theta^2 \ln f_t(\theta)\right] + \int \left[D_\theta \ln f_t(\theta)\right]\left[D_{\theta'} \ln f_t(\theta)\right] f_t(\theta)dy \\ &= \mathrm{E}_\theta\left[D_\theta^2 \ln f_t(\theta)\right] + \mathrm{E}_\theta\left[D_\theta \ln f_t(\theta) D_{\theta'} \ln f_t(\theta)\right] \\ &= \mathrm{E}_\theta\left[H_t(\theta)\right] + \mathrm{E}_\theta\left[g_t(\theta) g_t(\theta)'\right] \end{aligned}$$

The information matrix equality - 2

Now sum over n and multiply by $\frac{1}{n}$

$$\mathrm{E}_\theta \frac{1}{n} \sum_{t=1}^{n} \left[H_t(\theta) \right] = -\mathrm{E}_\theta \left[\frac{1}{n} \sum_{t=1}^{n} \left[g_t(\theta) \right] \left[g_t(\theta)' \right] \right]$$

The scores g_t and g_s are uncorrelated for $t \neq s$, since for $t > s$, $f_t(y_t|y_1, ..., y_{t-1}, \theta)$ has conditioned on prior information, so what was random in s is fixed in t.

This allows us to write

$$\mathrm{E}_\theta \left[H(\theta) \right] = -\mathrm{E}_\theta \left[n g(\theta) g(\theta)' \right]$$

since all cross products between different periods expect to be zero.

Finally take limits, we get

$$H_\infty(\theta) = -\mathcal{I}_\infty(\theta).$$

The information matrix equality - 3

This holds for all θ, in particular, for θ_0. Using this,

$$\sqrt{n}\left(\hat{\theta}-\theta_0\right) \stackrel{a.s.}{\to} N\left[0, H_\infty(\theta_0)^{-1}\mathcal{I}_\infty(\theta_0)H_\infty(\theta_0)^{-1}\right]$$

simplifies to

$$\sqrt{n}\left(\hat{\theta}-\theta_0\right) \stackrel{a.s.}{\to} N\left[0, \mathcal{I}_\infty(\theta_0)^{-1}\right]$$

To estimate the asymptotic variance, we need estimators of $H_\infty(\theta_0)$ and $\mathcal{I}_\infty(\theta_0)$. We can use

$$\begin{aligned}\widehat{\mathcal{I}_\infty(\theta_0)} &= n\sum_{t=1}^{n} g_t(\hat{\theta})g_t(\hat{\theta})' \\ \widehat{H_\infty(\theta_0)} &= H(\hat{\theta}).\end{aligned}$$

The information matrix equality - 4

From this we see that there are alternative ways to estimate $V_\infty(\theta_0)$ that are all valid. These include

$$\begin{aligned}
\widehat{V_\infty(\theta_0)} &= -\widehat{H_\infty(\theta_0)}^{-1} \\
\widehat{V_\infty(\theta_0)} &= \widehat{\mathcal{I}_\infty(\theta_0)}^{-1} \\
\widehat{V_\infty(\theta_0)} &= \widehat{H_\infty(\theta_0)}^{-1}\widehat{\mathcal{I}_\infty(\theta_0)}\widehat{H_\infty(\theta_0)}^{-1}
\end{aligned}$$

These are known as the *inverse Hessian*, *outer product of the gradient* (OPG) and *sandwich* estimators, respectively.

The sandwich form coincides with the covariance estimator of the *quasi*-ML estimator.

The Cramér-Rao lower bound

Theorem 1: The limiting variance of a CAN estimator of θ_0, say $\tilde{\theta}$, minus the inverse of the information matrix is a positive semidefinite matrix.

Proof: Since the estimator is CAN, it is asymptotically unbiased, so

$$\lim_{n\to\infty} \mathrm{E}_\theta(\tilde{\theta} - \theta) = 0$$

Differentiate wrt θ' :

$$D_{\theta'} \lim_{n\to\infty} \mathrm{E}_\theta(\tilde{\theta} - \theta) = \lim_{n\to\infty} \int D_{\theta'} \left[f(Y,\theta) \left(\tilde{\theta} - \theta \right) \right] dy = 0.$$

Noting that $D_{\theta'} f(Y,\theta) = f(\theta) D_{\theta'} \ln f(\theta)$, we can write

$$\lim_{n\to\infty} \int \left(\tilde{\theta} - \theta \right) f(\theta) D_{\theta'} \ln f(\theta) dy + \lim_{n\to\infty} \int f(Y,\theta) D_{\theta'} \left(\tilde{\theta} - \theta \right) dy = 0.$$

The Cramér-Rao lower bound - 2

Now note that $D_{\theta'}\left(\tilde{\theta}-\theta\right)=-I_K$, and $\int f(Y,\theta)(-I_K)dy=-I_K$. With this we have

$$\lim_{n\to\infty}\int\left(\tilde{\theta}-\theta\right)f(\theta)D_{\theta'}\ln f(\theta)dy=I_K.$$

We get

$$\lim_{n\to\infty}\int\sqrt{n}\left(\tilde{\theta}-\theta\right)\sqrt{n}\underbrace{\frac{1}{n}\left[D_{\theta'}\ln f(\theta)\right]}f(\theta)dy \quad = \quad I_K$$

Note that the bracketed part is just the transpose of the score vector, $g(\theta)$, so we can write

$$\lim_{n\to\infty}\mathrm{E}_\theta\left[\sqrt{n}\left(\tilde{\theta}-\theta\right)\sqrt{n}g(\theta)'\right]=I_K$$

The Cramér-Rao lower bound - 3

This means that the covariance of the score function with $\sqrt{n}\left(\tilde{\theta}-\theta\right)$, for $\tilde{\theta}$ any CAN estimator, is an identity matrix. Using this, suppose the variance of $\sqrt{n}\left(\tilde{\theta}-\theta\right)$ tends to $V_\infty(\tilde{\theta})$. Therefore,

$$V_\infty\left(\begin{array}{c}\sqrt{n}\left(\tilde{\theta}-\theta\right)\\ \sqrt{n}g(\theta)\end{array}\right)=\left[\begin{array}{cc}V_\infty(\tilde{\theta}) & I_K\\ I_K & \mathcal{I}_\infty(\theta)\end{array}\right].$$

Since this is a covariance matrix, it is positive semi-definite. Therefore, for any K -vector α,

$$\left[\begin{array}{cc}\alpha' & -\alpha'\mathcal{I}_\infty^{-1}(\theta)\end{array}\right]\left[\begin{array}{cc}V_\infty(\tilde{\theta}) & I_K\\ I_K & \mathcal{I}_\infty(\theta)\end{array}\right]\left[\begin{array}{c}\alpha\\ -\mathcal{I}_\infty(\theta)^{-1}\alpha\end{array}\right]\geq 0.$$

The Cramér-Rao lower bound - 4

This simplifies to

$$\alpha' \left(V_\infty(\tilde{\theta}) - \mathcal{I}_\infty^{-1}(\theta) \right) \alpha \geq 0.$$

Since α is arbitrary, $V_\infty(\tilde{\theta}) - \mathcal{I}_\infty(\theta)$ is positive semi-definite. ■

This means that $\mathcal{I}_\infty^{-1}(\theta)$ is a *lower bound* for the asymptotic variance of a CAN estimator.

Definition 3: Given two CAN estimators of a parameter θ_0, say $\tilde{\theta}$ and $\hat{\theta}$, $\hat{\theta}$ is asymptotically efficient with respect to $\tilde{\theta}$ if $V_\infty(\tilde{\theta}) - V_\infty(\hat{\theta})$ is a positive semi-definite matrix.

A direct proof of asymptotic efficiency of an estimator is infeasible, but if one can show that the asymptotic variance is equal to the inverse of the information matrix, then the estimator is asymptotically efficient.

In particular, *the MLE is asymptotically efficient.*

MLE - Summary

- Consistent.
- Asymptotically normal (CAN).
- Asymptotically efficient.
- Asymptotically unbiased.

This is for general MLE: we haven't specified the distribution or the linearity/nonlinearity of the estimator.

Method of Moments

The OLS estimator can be thought as a method of moments estimators.

Assuming weak exogeneity, $\mathrm{E}[\boldsymbol{x}_t \epsilon_t] = 0$, we have

$$\mathrm{E}\left[\frac{1}{n}\sum_{t=1}^{n} \boldsymbol{x}_t \epsilon_t\right] = \mathrm{E}\left[\frac{\mathbf{X}'\boldsymbol{\varepsilon}}{n}\right] = 0.$$

The idea of the *MM estimator* is to choose the estimator such that the empirical counterpart hold:

$$\begin{aligned} \frac{\mathbf{X}'\widehat{\boldsymbol{\varepsilon}}}{n} &= 0 \\ \frac{\mathbf{X}'(\boldsymbol{y}-\mathbf{X}\widehat{\boldsymbol{\beta}})}{n} &= 0 \\ \widehat{\boldsymbol{\beta}} &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{y} \end{aligned}$$

MM provides another interpretation to OLS estimator.

Method of Moments - 2

Imposing other (or more) moment restrictions. Why not?

Example 3: Suppose that X follows a $\chi^2_{\theta_0}$ distribution and θ_0 is the parameter of interest.

- Using the first moment equation $\mathrm{E}[X] = \theta_0$, we obtain: $\widehat{\theta}_1 = n^{-1} \sum_{i=1}^{n} x_i$.
- Using the second moment equation $\mathrm{E}[X^2] = \theta_0^2 + 2\theta_0$, we obtain:

$$\widehat{\theta}_2 = -1 + \sqrt{1 + n^{-1} \sum_{i=1}^{n} x_i^2}.$$

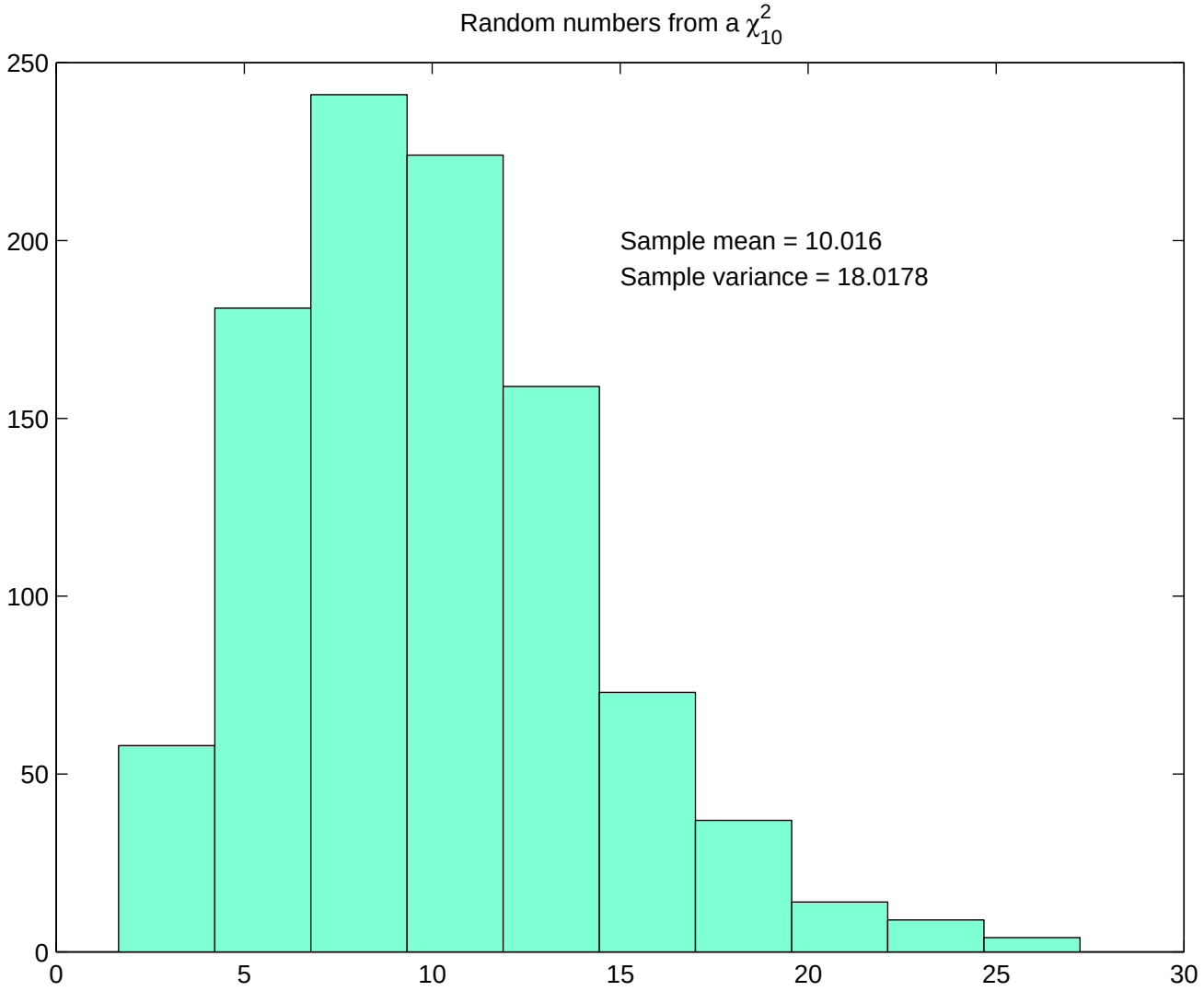
Notice that $n^{-1} \sum_{i=1}^{n} x_i$ and $n^{-1} \sum_{i=1}^{n} x_i^2$ are (by LLN) consistent estimators of θ_0 and $\theta_0^2 + 2\theta_0$, so $\widehat{\theta}_1$ and $\widehat{\theta}_2$ are consistent estimators of θ_0.

Example 3: Compare by simulation the efficiency of estimators $\widehat{\theta}_1$ and $\widehat{\theta}_2$.

Lets MATLAB works ...

1. Generate m samples of size n of a χ^2_{10}.
2. Estimate the degree of freedom using $\widehat{\theta}_1$ and $\widehat{\theta}_2$.
3. Estimate the bias of these estimators.
4. Estimate the variance of these estimators.
5. Estimate the mean squared errors.

Example 3: A histogram of $n = 1000$ observations:



Example 3: Bias, variance and MSE. estimation.

Bias:

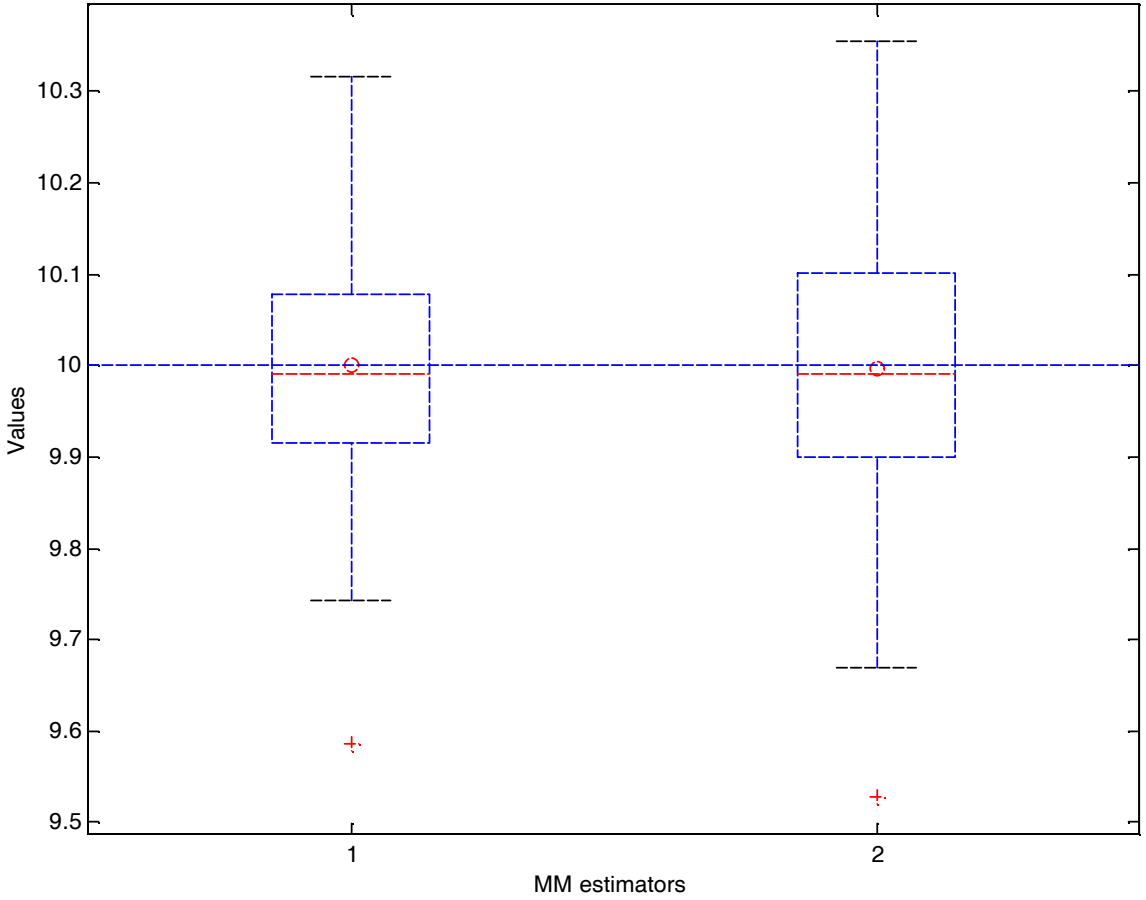
```
0.0005    -0.0038
```

Variance:

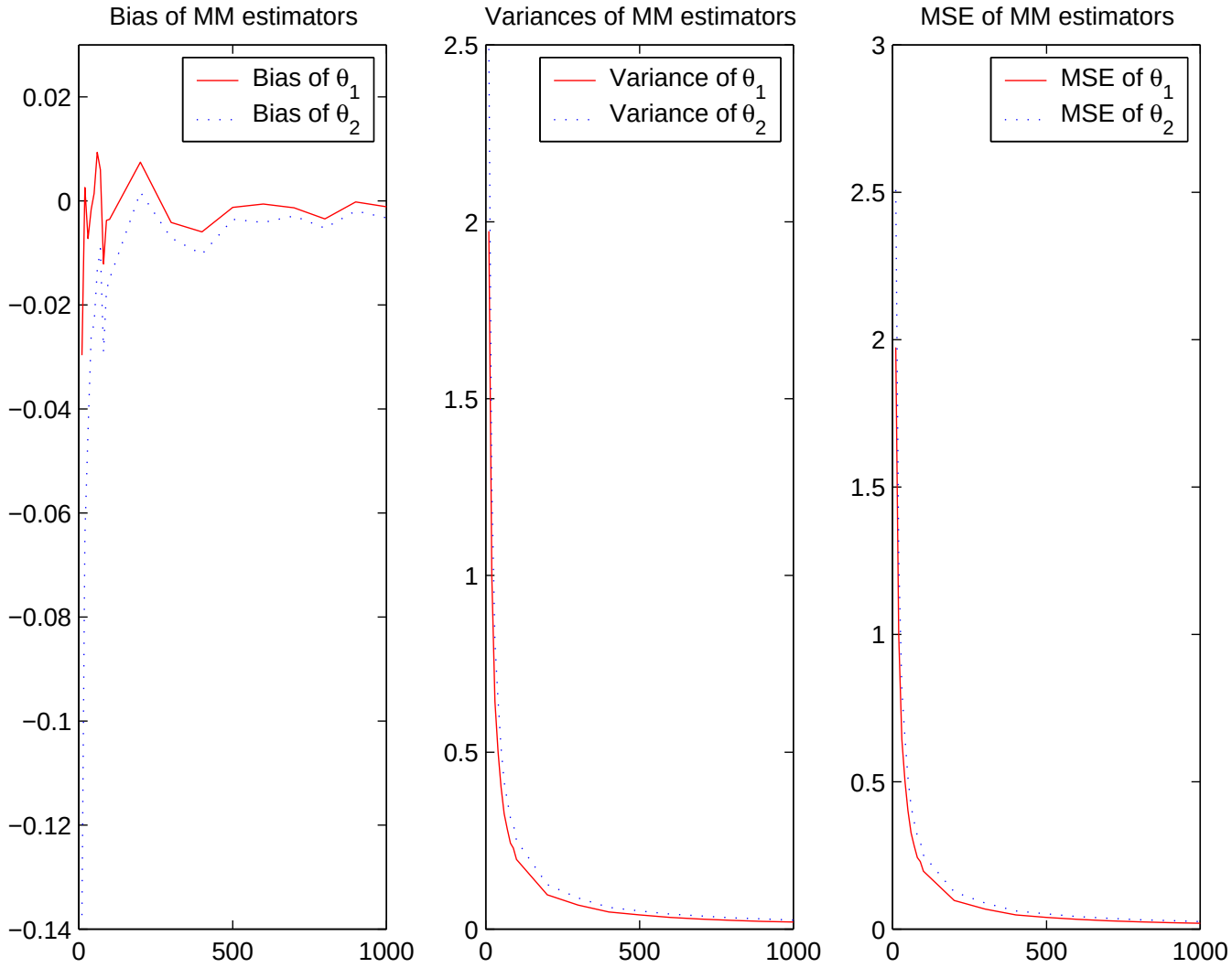
```
0.0160     0.0212
```

MSE:

```
0.0160     0.0212
```



Example 3: Bias, variance and MSE as a function of n.



Generalized Method of Moments

Definition 4: The GMM estimator of the K -dimensional parameter vector θ_0, is defined by

$$\hat{\theta} \equiv \arg\min_{\Theta} s_n(\theta) \equiv m_n(\theta)' W_n m_n(\theta),$$

where

$$m_n(\theta) = \frac{1}{n} \sum_{t=1}^{n} m_t(\theta)$$

is a g-vector, $g \geq K$, with $\mathrm{E}_\theta m(\theta) = 0$ (the moment restriction), and W_n converges almost surely to a finite $g \times g$ symmetric positive definite matrix W_∞.

Example 3 (bis): Combine θ_1 and θ_2 restrictions to get a GMM estimator.

Lets MATLAB works ...

Example 3: Bias, variance and MSE as a function of n.