Boston University Questrom School of Business

MF840 – Spring 2021

Eric Jacquier

MAXIMUM LIKELIHOOD ESTIMATION

Definition, application to the mean/variance estimation

Application to the regression

Asymptotic Properties

Variance of a function of the parameters: Delta Method

1. MLE – Maximum Likelihood Estimator – Principle and Derivation

First example, the *mean / variance problem*, comparison with OLS (could be regression, GLS,..)

$$y_t = \mu + \varepsilon_t \qquad [1]$$

Collect t = 1, … , T observations
Need to estimate μ, σ

1.1 Recall OLS estimation of μ

Find μ to minimize ε'ε ⇔ Min $\Sigma_t (y_t-\mu)^2$ [2]
FOC: $0 = -2 \Sigma_t(y_t-\mu)$

- Solution: $\hat{\mu} = \Sigma_t y_t/T$ the *sample mean*.
- What assumptions did we need? Model equation [1] and optimization criterion [2] nothing else.
- **Then**, we discussed the properties of $\hat{\mu}$: is it unbiased? What is its variance?

We added assumptions: $\varepsilon_t \sim$ **i.i.d.**$(0, \sigma^2)$ still no distribution was needed

We obtained: Unbiasedness
$\text{Var}(\hat{\mu}) = \sigma^2/T$
"Sample mean is the BLUE of μ" the Gauss-Markov theorem.

- **Then only**, we asked about the distribution of $\hat{\mu}$, recall three possibilities

  1. $\varepsilon_t \sim$ Normal => $\hat{\mu} \sim$ Normal

  2. T large => $\hat{\mu}$ is approximately normal even if ε_t is not normal

  3. Neither 1 nor 2? we are in trouble!

- Finally, OLS did not tell us how to estimate the model error variance σ^2.

  We used an unbiased estimator as a separate additional, adhoc, step:

  $s^2 = \frac{1}{T-1}\sum_1^T (y_t - \hat{\mu})^2$

  We know how to prove that $E(s^2) = \sigma^2$

1.2 Introducing Maximum Likelihood Estimation (MLE)

Mean/variance model: $y_t = \mu + \varepsilon_t$ [1]

- MLE immediately requires a distributional assumption: $\varepsilon_t \sim$ i.i.d. **N**$(0, \sigma^2)$

  So we **know** the density of y_t, and the joint density of $Y = (y_1, y_2, \ldots, y_T)$

  $p(y_t|\mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(y_t-\mu)^2}{2\sigma^2}}, \forall t,$ identical assumption, normality

  $p(y_1, y_2, \ldots, y_T \mid \mu, \sigma) = \prod_t p(y_t \mid \mu, \sigma)$ independence assumption

- We call this joint density of the data, the ***likelihood function*** *of the parameters*

$$p(Y|\mu, \sigma) = \prod_{t=1}^{T} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(y_t-\mu)^2}{2\sigma^2}} = \frac{1}{(\sqrt{2\pi}\sigma)^T} e^{-\frac{\sum(y_t-\mu)^2}{2\sigma^2}} = \ell(\mu, \sigma \mid Y) \quad [2]$$

We can view [2] as the joint pdf of Y given the parameters (μ, σ)

or

We can view [2] as a mathematical function of the parameters (μ, σ) given the observed data Y

- Intuition:
    1. Data Y were randomly generated by this density with **unknown** parameters μ, σ.
    2. Idea: Find the value of the parameters (μ, σ) that makes the data Y the likeliest to have been observed.

- MLE method:
    - Write the joint density of the data as in [2]
    - Given our **known** sample Y, consider [2] as a function of the unknown parameters (μ, σ).
    - Find the value of (μ, σ) that maximizes [2]

- Likelihood function in [2] is a product of individual densities (… in our very simple model).

We **always** take the logarithm to make the optimization simpler: This is the **log-likelihood**

$$Log\ell \propto -T\ log\sigma - \frac{\sum(y_t-\mu)^2}{2\sigma^2}$$

We don't need to drag stuff irrelevant in the optimization (be careful with that!)

Let's maximize

$$Log\ell \propto -T \log\sigma - \frac{\sum(y_t-\mu)^2}{2\sigma^2} \qquad [2']$$

First Order Conditions (FOC)

$$\partial \text{Log}\ell/\partial\mu = \frac{2\sum(y_t-\mu)}{2\sigma^2} = 0 \qquad [3]$$

$$\partial \text{Log}\ell/\partial\sigma = -\frac{T}{\sigma} + \frac{2\sum(y_t-\mu)^2}{2\sigma^3} = 0 \qquad [4]$$

[3] → $\widehat{\mu_{MLE}} = \Sigma\, y_t / T$ [5]

[4] → $\widehat{\sigma^2_{MLE}} = \frac{1}{T}\sum_1^T (y_t - \hat{\mu})^2$ [6]

- Note
  - We get **all** the parameters of the density of the data by the same logic
  - Divisor for σ is **T, not (T-1)**, **MLE estimator can be biased**
  - Why do we find the same result as OLS for the mean ?

- **MLE is invariant to transformations**

In [4] we differentiated on σ, so we should have found the estimate for σ.
But we just wrote the estimate of σ^2 in [6] ?

Is the estimate of the square of a parameter, just the square of the estimate ?

Is the estimate of a function of a parameter, just the function of the estimate?

Our FOC was with respect to σ.
Formally, what would be the MLE of σ^2?
Shouldn't we differentiate with respect to σ^2 to find it!?

Chain rule $\quad \partial \mathbf{Log\ell} / \partial \boldsymbol{\sigma} = \partial \mathbf{Log\ell} / \partial \mathbf{(f(\sigma))} \times \partial f(\sigma) / \partial \sigma$

$$\partial Log\ell / \partial \sigma = \partial Log\ell / \partial(\sigma^2) \times \underbrace{\partial \sigma^2 / \partial \sigma}_{\neq 0}$$

Both derivatives are zero together for any function of σ with non-zero derivative

=> MLE is invariant to transformations: $\widehat{f(\theta)}_{MLE} = f(\hat{\theta}_{MLE})$

Exercise: differentiate [2'] with respect to σ^2, what do you find?

This validates the *"plug-in approach"* for functions of the parameters

- Remember: Likelihood function is just another name for the joint density of the data

  We say: the joint density of the data but ...
  We also say: the likelihood function of the parameters

  Once we get the data, we know Y.

  Then we look at the joint density of Y as a function of the unknown parameters given Y.

- Remember: Crucial difference with Least Squares:

  - We need **all** the distributional assumptions **right away** (e.g. joint density of the noise, independence of the noise, etc…) to estimate the parameters.

  - We can estimate **all** the parameters of a density by writing the FOCs, one per parameter.

  - The concept generalizes to any density. It may not be computationally simple, it may require numerical optimization. But the principle is the same.

- Autocorrelated, heteroskedastic errors?
  …. Just another (more complicated) density

2. MLE for the Regression Model

- Model:

  $Y = X\beta + \varepsilon, \qquad \varepsilon \sim \mathbf{N}(0, \sigma^2 I)$ <- Need all the assumptions on the noise right away

- First write the density of one **noise** ε_t, then the joint density as the product by independence

  $p(\varepsilon \mid \sigma) = \prod_t p(\varepsilon_t \mid \sigma)$. i.i.d.

- But we need to write p(Y), not p(ε).

  $Y = f(\varepsilon) = X\beta + \varepsilon$

  Must use the Change of Variable rule (seen in 793) to get p(Y):

  $$\mathbf{p_Y(Y)\, dY = p_\varepsilon(\varepsilon)\, d\varepsilon \;\; |det(\, d\varepsilon / dY)|}$$

  Here f is simple: $Y = X\beta + \varepsilon$, The Jacobian matrix is identity.

  In more complex models, the Jacobian **may not be identity**. Do not forget this step.

$$p(Y \mid \beta,\sigma) \propto \frac{1}{\sigma^T}\exp\left\{-\frac{1}{2\sigma^2}(Y-X\beta)'(Y-X\beta)\right\} \equiv \ell(\beta,\sigma \mid Y)$$

$\text{Log}\ell = -\,T\,\text{Log}\,\sigma - (Y\text{-}X\beta)'(Y\text{-}X\beta) / 2\sigma^2$

First Order Conditions (FOC)

$\partial\,\text{Log}\ell/\partial\beta = -(-2\,X'Y + 2\,X'X\beta) / 2\sigma^2 = 0$ [1]

$\partial\,\text{Log}\ell/\partial\sigma = -\,T/\sigma + (Y\text{-}X\beta)'(Y\text{-}X\beta) / \sigma^3 = 0$ [2]

[1] → $\hat{\beta}_{MLE} = (X'X)^{-1}X'Y$

[2] → $\widehat{\sigma^2}_{MLE} = \frac{1}{T}\hat{\varepsilon}'\hat{\varepsilon}$

- s^2 divisor: again T, not (T-K)

- The MLE is biased. It this good or bad? What are its statistical properties?

3 Properties (Bias, Variance, Consistency, Efficiency) of the MLE

- **Result 1:** Asymptotic unbiasedness

  Looking at σ for the regression, we saw that the MLE can be biased in small sample

  The bias goes to zero as T→ ∞ ; 1/T → 1/(T-k).

  It is in fact the case for any model:

  As T→∞, the bias of the MLE estimator always goes to zero

- **Result 2:** **Cramer-Rao Lower Bound** (no proof)

  Denote the (co)variance (matrix) of any unbiased estimator $\hat{\theta}$: $V(\hat{\theta})$.

  Then $V(\hat{\theta}) \geq I(\theta)^{-1}$ ← ***Fisher Information Matrix***

  The ***Fisher information matrix*** is (minus) the matrix of second derivatives of the likelihood function with respect to the parameters

  C-R Lower Bound theorem says that no unbiased estimator can have a variance smaller than the inverse of the Fisher information matrix.

- **Result 3:** **MLE reaches the C-R lower bound asymptotically.** (no proof)

$$V\left(\hat{\theta}_{MLE}\right) \doteq I(\theta)^{-1} \qquad [3]$$

No proof but … you must know this for the rest of your statistical life!

- Results 1 + 3:

MLE has zero asymptotic bias

MLE has the lowest variance … **in large sample.**

MLE is consistent and efficient

- Computational issues

Result 3 tells us how to compute the approximate (asymptotic) covariance matrix for $\hat{\theta}_{MLE}$

We can also show the second equality

$$\mathbf{V}(\hat{\boldsymbol{\theta}}_{\boldsymbol{MLE}}) \;\dot{\approx}\; -\left[\boldsymbol{E}\left[\frac{\partial^2 \boldsymbol{Log}\ell(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\,\partial\boldsymbol{\theta}'}\right]\right]^{-1} = \left[\boldsymbol{E}\left[\frac{\partial \boldsymbol{Log}\ell(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}} \times \left(\frac{\partial \boldsymbol{Log}\ell(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}\right)'\right]\right]^{-1} \qquad [4]$$

This second, alternate way, to compute the covariance matrix is called the **Hessian matrix.**

- In theory it does not matter which one we use since they are equal!

  In practice with data and complicated models, we may have to use sample means to compute the expectations inside the matrix ….

  …… and then the matrix of second derivatives computed with sample means may not be positive for a given data set

  But the Hessian computation **is always positive** since it is an outer product

- The computation in [4] results in a function of the unknown θ. We are not done!

  We **estimate [4]** at the MLE point by replacing the parameters in [4] with their MLE estimates. We do this last.

4 Computing the (co)variance matrix of the MLE

We compute the covariance matrix **at the MLE point estimate**, using [1] above.
But first we may have to take expectations to get the information matrix.

4.1 Normal μ, σ model $Log\ell \propto -T\ log\sigma - \frac{\sum(y_t-\mu)^2}{2\sigma^2}$

- $\frac{\partial^2 Log\ell}{\partial\mu\,\partial\mu} = \frac{\partial}{\partial\mu}[\qquad\qquad] = \frac{-T}{\sigma^2}$ [1]

- $\frac{\partial^2 Log\ell}{\partial\mu\,\partial\sigma} = \frac{\partial}{\partial\sigma}[\qquad\qquad] =$

  Now take the expectation: $=$ [2]

- $\frac{\partial^2 Log\ell}{\partial\sigma\partial\sigma} = \frac{\partial}{\partial\sigma}[-\qquad + \qquad\qquad] = \frac{T}{\sigma^2} - \frac{3\sum(y_t-\mu)^2}{\sigma^4}$

  Again take the expectation over the data:

$$E\frac{\partial^2 Log\ell}{\partial\sigma^2} = \frac{T}{\sigma^2} - \frac{3E\sum(y_t-\mu)^2}{\sigma^4} = \frac{T}{\sigma^2} - \frac{3T\sigma^2}{\sigma^4} = \frac{-2T}{\sigma^2}$$ [3]

We have our information matrix to be inverted:

$$\mathbf{V}(\hat{\mu}_{MLE}, \hat{\sigma}_{MLE}) \approx -\left[E\left[\frac{\partial^2 Log\ell(\theta)}{\partial\theta\,\partial\theta'}\right]\right]^{-1}$$

$$= -\begin{bmatrix} \frac{-T}{\sigma^2} & 0 \\ 0 & \frac{-2T}{\sigma^2} \end{bmatrix}^{-1}$$

$$= \begin{pmatrix} \frac{T}{\sigma^2} & 0 \\ 0 & \frac{2T}{\sigma^2} \end{pmatrix}^{-1} = \begin{pmatrix} \mathbf{0} & \\ \mathbf{0} & \end{pmatrix} \qquad [4]$$

- Now prove by the same method that $\boldsymbol{V(\widehat{\sigma^2}_{MLE}) = \frac{2\sigma^4}{T}}$

- Use the Hessian matrix method and (you better ☺ !) find the same result as [4]

4.2 Normal Regression Model: β, σ

- $\text{Log}\ell = -\,T\,\text{Log}\,\sigma - (Y-X\beta)'(Y-X\beta) / 2\sigma^2$

- 1st derivatives:
  $\partial\,\text{Log}\ell/\partial\beta = -(-2X'Y+2X'X\beta) / 2\sigma^2$
  $\partial\,\text{Log}\ell/\partial\sigma = -\,T/\sigma + (Y-X\beta)'(Y-X\beta) / \sigma^3$

- 2nd derivatives:
  - $\frac{\partial^2 Log\ell(\beta,\sigma)}{\partial\beta^2} = -\frac{X'X}{\sigma^2}$ [1]
  - $E\,\frac{\partial^2 Log\ell(\beta,\sigma)}{\partial\beta\,\partial\sigma} = E\,\frac{\partial}{\partial\sigma}\left[-\frac{(-2X'Y+2X'X\beta)}{2\sigma^2}\right] = 0$ Why E = 0 ? [2]
  - $\frac{\partial^2 Log\ell(\beta,\sigma)}{\partial\sigma^2} = \frac{\partial}{\partial\sigma}\left[-\frac{T}{\sigma} + \frac{\varepsilon'\varepsilon}{\sigma^3}\right] = \frac{T}{\sigma^2} - \frac{3\varepsilon'\varepsilon}{\sigma^4}$

$$E\,\frac{\partial^2 Log\ell(\beta,\sigma)}{\partial\sigma^2} = \frac{T}{\sigma^2} - \frac{3E\varepsilon'\varepsilon}{\sigma^4} = \frac{T}{\sigma^2} - \frac{3T\sigma^2}{\sigma^4} = \frac{-2T}{\sigma^2} \quad [3]$$

$$V(\hat{\beta},\hat{\sigma}) = I(\beta,\sigma)^{-1} = -\begin{pmatrix} -\frac{X'X}{\sigma^2} & 0 \\ 0 & -\frac{2T}{\sigma^2} \end{pmatrix}^{-1} = \begin{pmatrix} \sigma^2(X'X)^{-1} & 0 \\ 0 & \frac{\sigma^2}{2T} \end{pmatrix}$$

4.3 Asymptotic Distribution of the MLE and small-sample checks

Result: The ML estimator is **asymptotically** normally distributed …. (no proof)

$\hat{\sigma}_{MLE} \dot{\sim} N(\sigma, \frac{\sigma^2}{2T})$. $\widehat{\sigma^2}_{MLE} \dot{\sim} N(\sigma^2, \frac{2\sigma^4}{T})$. Both are normally distributed ?

… How can x and x^2 both have the same distribution ?
… By MLE, any estimator and any function of the estimator is asymptotically normally distributed:

$\widehat{\boldsymbol{\theta}}_{\boldsymbol{MLE}} \dot{\sim} \boldsymbol{N}$ and whatever the function f: $\boldsymbol{f}(\widehat{\boldsymbol{\theta}}_{\boldsymbol{MLE}}) \dot{\sim} \boldsymbol{N}$

For σ, other problem: A normal distribution is unbounded, σ is positive

Is this a Problem? This is the power of approximation in very large sample

Check for variance, we have: $\widehat{\sigma^2}_{MLE} \dot{\sim} N(\sigma^2, \frac{2\sigma^4}{T})$ **asymptotically**

- Recall the unbiased estimator: $s^2 = \varepsilon'\varepsilon/(T-k)$ => $\widehat{\sigma^2}_{MLE} = (T-k)\, s^2 / T$

  We had proven: $(T-k)s^2/\sigma^2 \sim \chi^2(T-k)$ …. **exactly** with normal noise, so:

$$\widehat{\sigma^2}_{MLE} \sim \sigma^2 \frac{\chi^2(T-k)}{T} \quad \textbf{exactly}$$

- How bad is the contradiction?
  How large must T be for the approximation to be good?

  Regression model is simple, this should work well.
  For more complex models, the large sample approximation may not be good.

- If not sure, must check! Easy to check:

```
# TT observations, k parameters, σ=0.2.
tt<- 20; k<- 4; truvar<-0.2^2
varhat<- truvar*rchisq(100000, tt-4) / tt      # exact distribution
hist(varhat,nclass=100,freq=F,main=paste("Sample Size: ",tt));box()
abline(v=mean(varhat),lwd=2,col="blue")
vrange<-seq(min(varhat),max(varhat),length=200)
lines(vrange,dnorm(vrange,truvar, truvar*sqrt(2/tt)))   # MLE approximate
```

- Always think whether you may have a small sample problem. Potential problems more severe for bounded parameters like correlations

5 MLE of a transformations of parameters

Have $\hat{\theta}_{MLE}$, but need $\widehat{f(\theta)}_{MLE}$ (have estimate of σ but need estimate of σ^2 or $1/\sigma$)

5.1 Point estimate of $f(\theta)$

- Invariance => Use $f(\hat{\theta}_{MLE})$ …. Known as the **plug-in approach**
- How can this be justified?
  If $E(\hat{\theta})_{MLE} = \theta$, do we have: $E(f(\hat{\theta})_{MLE}) = f(\theta)$?

  Do we have: $E(x^2) = [E(x)]^2$?

  Of course not because x is a random variable, and we have a Jensen effect
- But with very large sample size, $V(\hat{\theta}) \ll E(\hat{\theta})$ and the Jensen effect becomes negligible. This is why it works as $T \to \infty$.

  Also why you must always check whether this approximation is realistic for **your** sample size.
- So for MLE: **Uses $f(\hat{\theta}_{MLE})$ for $\widehat{f(\theta)}_{MLE}$**

5.2 Variance of $f(\hat{\theta})_{MLE}$

1. Can try to rewrite the likelihood with respect to f(θ) and then compute the information matrix. Most often not convenient or feasible.

2. A more practical approach: the **Delta Method**

The Delta Method uses the expansion of a random variable x around a fixed value x_0.

$f(x) \approx f(x_0) + (x - x_0)' \ \partial f/\partial x|_{x0}$

$f(\hat{\theta}) \approx f(\theta) + (\hat{\theta} - \theta)' \ \partial f/\partial \theta|_{\theta}$ expand $\hat{\theta}$ around θ the true parameter

Square it: $[\ f(\hat{\theta}) - f(\theta)\ ]^2 \approx [\partial f/\partial \theta]' \ (\hat{\theta} - \theta)\ (\hat{\theta} - \theta)' \ \partial f/\partial \theta$

Take expectation: $E[\ f(\hat{\theta}) - f(\theta)]^2 \approx [\partial f/\partial \theta\ ]' \ V(\hat{\theta}_{MLE}) \ \partial f/\partial \theta$

$$V[\widehat{\mathbf{f(\theta)}}] \approx [\partial f/\partial \theta\ ]' \ V(\hat{\boldsymbol{\theta}}_{\mathbf{MLE}}) \ \partial f/\partial \theta \qquad [1]$$

- That was nice ! Now we need numbers, at what point do we compute the derivative ?
  At the MLE … it's the best we have

- Quality of this approximation again depends on how good the large sample approximation is.