1

Maximum Likelihood Estimation (MLE)

Hüseyin Taştan[1]

[1]Department of Economics
Ph.D. Program - Advanced Econometrics
Yildiz Technical University

November 26, 2013

2

Joint Density Function

- Suppose that we have a random sample of n observations collected in the n-vector $\boldsymbol{y} = [y_1, y_2, \ldots, y_n]^\top$.
- Since we have a random sample each y_i will be identically distributed, coming from the same probability density function $f(y_i, \boldsymbol{\theta})$.
- Additionally they will be independently distributed. In short we can write $y_i \sim \ iid \ f(y_i, \boldsymbol{\theta})$, which is implied by the random sample assumption.
- Using the property of **statistical independence** the joint probability density function (pdf) of the random sample y is given by

$$f(y_1, y_2, \ldots, y_n; \boldsymbol{\theta}) = f(\boldsymbol{y}, \boldsymbol{\theta}) = \prod_{i=1}^{n} f(y_i, \boldsymbol{\theta}) \tag{1}$$

where the unknown parameter vector is collected in

$$\boldsymbol{\theta} = [\theta_1, \theta_1, \ldots, \theta_k]^\top, \ k \times 1$$

3

Loglikelihood Function

There are two interpretations of the joint pdf:

- Usual density interpretation: Given $\boldsymbol{\theta}$ what is the joint density of the random sample? We know (fix) the parameters but we do not know the sample (not observed yet).
- Likelihood interpretation: Given the random sample, what is the likelihood of the random sample from a particular population distribution. We know (observe) the random sample but we do not know the parameter vector. The joint pdf $f(\boldsymbol{y}, \boldsymbol{\theta})$ is evaluated at the data given by n-vector $\boldsymbol{y}$. Instead, it is referred to as the likelihood function of the model for the given data set.

Let us focus on the second interpretation:

$$Likelihood\ function \ = L(\boldsymbol{\theta}, \boldsymbol{y}) = f(\boldsymbol{y}, \boldsymbol{\theta}) = \prod_{i=1}^{n} f(y_i, \boldsymbol{\theta}) \tag{2}$$

ML estimation maximizes the likelihood function w.r.t. the parameters.

4

Maximum Likelihood Estimates

- The ML estimators are defined as the ones which maximizes the likelihood that the sample is chosen from the population distribution which is assumed to be known:

$$\hat{\theta} = \arg\max_{\theta} L(\boldsymbol{\theta}, \boldsymbol{y}) \tag{3}$$

- The loglikelihood function is

$$\begin{aligned} Q(\boldsymbol{\theta}) &\equiv \log L(\boldsymbol{\theta}, \boldsymbol{y}) \\ &= \log \left[\prod_{i=1}^{n} f(y_i, \boldsymbol{\theta}) \right] \\ &\equiv \ell(\boldsymbol{\theta}) \end{aligned} \tag{4}$$

$$\max_{\boldsymbol{\theta}} Q(\boldsymbol{\theta}) = \sum_{i=1}^{n} \log f(y_i, \boldsymbol{\theta}) \tag{5}$$

- Note that $\ell(y_i, \boldsymbol{\theta}) \equiv \log f(y_i, \boldsymbol{\theta})$ is the contribution to the loglikelihood function made by the observation i.

The Score Vector

- The gradient vector of the loglikelihood function is also called the score vector and given by

$$\begin{aligned} \frac{\partial Q(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} &= \frac{1}{L}\frac{\partial L}{\partial \boldsymbol{\theta}} && (6) \\ &\equiv s(\boldsymbol{\theta}; \boldsymbol{y}) && (7) \\ &= \begin{bmatrix} \frac{\partial Q}{\partial \theta_1} \\ \frac{\partial Q}{\partial \theta_2} \\ \vdots \\ \frac{\partial Q}{\partial \theta_k} \end{bmatrix} = \mathbf{0}_k && (8) \end{aligned}$$

- Note that each component of the gradient vector is a sum of n contributions from observations.
- If the model is correctly specified, then the expectations of the elements of the scores evaluated et the true $\boldsymbol{\theta}$ are zero.

Expectation of the Scores

- Let us derive the expected vaule of the score vector. By definiton,

$$\int \cdots \int f(y_1, y_2, \ldots, y_n; \boldsymbol{\theta}) dy_1, dy_2, \ldots, dy_n = \int \cdots \int L(\boldsymbol{\theta}; \boldsymbol{y}) \boldsymbol{dy} = 1$$

- Differentiating both sides:

$$\int \cdots \int \frac{\partial L(\boldsymbol{\theta}, \boldsymbol{y})}{\partial \boldsymbol{\theta}} dy = 0$$

$$\begin{aligned} E(s(\boldsymbol{\theta}; \boldsymbol{y})) &= \int \cdots \int s(\boldsymbol{\theta}; \boldsymbol{y}) L(\boldsymbol{\theta}; \boldsymbol{y}) dy \\ &= \int \cdots \int \frac{\partial \ell}{\partial \boldsymbol{\theta}} L(\boldsymbol{\theta}; \boldsymbol{y}) dy \\ &= \int \cdots \int \frac{\partial L(\boldsymbol{\theta}, \boldsymbol{y})}{\partial \boldsymbol{\theta}} dy \\ &= 0 \end{aligned}$$

Covariance Matrix of the Scores and the Information Matrix

- Covariance matrix of the score vector

$$\begin{aligned} \mathsf{Cov}(s(\boldsymbol{\theta})) &= \mathsf{E}\left[(s(\boldsymbol{\theta}) - \mathsf{E}(s(\boldsymbol{\theta})))(s(\boldsymbol{\theta}) - \mathsf{E}(s(\boldsymbol{\theta})))^\top\right] \\ &= \mathsf{E}\left[s(\boldsymbol{\theta}) s(\boldsymbol{\theta})^\top\right], \quad \text{since } \mathsf{E}\left(s(\boldsymbol{\theta})\right) = 0 \\ &= \mathsf{E}\left[\left(\frac{\partial \ell}{\partial \boldsymbol{\theta}}\right)\left(\frac{\partial \ell}{\partial \boldsymbol{\theta}}\right)^\top\right] \equiv \boldsymbol{I}(\boldsymbol{\theta}), \end{aligned}$$

where $\boldsymbol{I}(\boldsymbol{\theta})$ is called the (Fisher) **information matrix** (in outer product form).
- Information Matrix can also be defined as the negative of the expectation of the Hessian:

$$\boldsymbol{I}(\theta) = -\mathsf{E}\left[\frac{\partial^2 \ell}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^\top}\right].$$

which can be estimated from data.

Properties of MLEs

- **Invariance principle**: Let $\hat{\boldsymbol{\theta}}_{mle}$ be the MLE of $\boldsymbol{\theta}$. Also let $\gamma = g(\boldsymbol{\theta})$ be a function of $\boldsymbol{\theta}$. According to the invariance principle the MLE of γ is $\hat{\gamma}_{mle} = g(\hat{\boldsymbol{\theta}}_{mle})$.
- **Consistency**: under certain assumptions $\hat{\boldsymbol{\theta}}_{mle}$ is consistent:

$$\operatorname*{plim}_{n \to \infty} \hat{\boldsymbol{\theta}}_{mle} = \boldsymbol{\theta}$$

- **Asymptotic normality**:

$$as \;\; n \longrightarrow \infty, \qquad \sqrt{n}(\hat{\boldsymbol{\theta}}_{mle} - \boldsymbol{\theta}) \xrightarrow{d} N\left(0, \boldsymbol{\Sigma}\right)$$

where

$$\boldsymbol{\Sigma} = \boldsymbol{I}(\theta)^{-1}, \quad \boldsymbol{I}(\boldsymbol{\theta}) = -\mathsf{E}\left[\frac{\partial^2 \ell}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^\top}\right]$$

MLE Example

Let $\mathbf{x} = (x_1, x_2, \ldots, x_n)$ be random sample from a normal distribution with mean μ and variance σ^2. Find the MLEs of population parameters.
We have a random sample from $X \sim N(\mu, \sigma^2)$, thus marginal pdf is given by

$$f(x_i; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(x_i - \mu)^2\right), \quad -\infty < x_i < \infty, \ i = 1, 2, \ldots, n$$

Likelihood function:

$$\begin{aligned} L(\mu, \sigma^2 \mid \mathbf{x}) &= \prod_{i=1}^n f(x_i; \mu, \sigma^2) = \prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(x_i - \mu)^2\right) \\ &= \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n \prod_{i=1}^n \exp\left(-\frac{1}{2\sigma^2}(x_i - \mu)^2\right) \\ &= (2\pi\sigma^2)^{-n/2} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^n (x_i - \mu)^2\right) \end{aligned}$$

Example cont.

Loglikelihood function:

$$\log L(\mu, \sigma^2 \mid \mathbf{x}) = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^n (x_i - \mu)^2$$

FOC (scores):

$$\begin{aligned} \frac{\partial}{\partial \mu} \log L(\mu, \sigma^2 \mid \mathbf{x}) &= \frac{1}{\sigma^2}\sum_{i=1}^n (x_i - \mu) = 0 \\ \frac{\partial}{\partial \sigma^2} \log L(\mu, \sigma^2 \mid \mathbf{x}) &= -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_{i=1}^n (x_i - \mu)^2 = 0 \end{aligned}$$

Solution:

$$\hat{\mu}_{mle} = \overline{X}, \quad \hat{\sigma}^2_{mle} = \frac{1}{n}\sum_{i=1}^n (X_i - \overline{X})^2$$

Example cont.

2nd derivatives:

$$\begin{aligned} \frac{\partial^2}{\partial \mu^2} \log L(\mu, \sigma^2 \mid \mathbf{x}) &= -\frac{n}{\sigma^2} \\ \frac{\partial^2}{\partial \mu \partial \sigma^2} \log L(\mu, \sigma^2 \mid \mathbf{x}) &= -\frac{1}{\sigma^4}\sum_{i=1}^n (x_i - \mu) \\ \frac{\partial^2}{\partial (\sigma^2)^2} \log L(\mu, \sigma^2 \mid \mathbf{x}) &= \frac{n}{2\sigma^4} - \frac{1}{\sigma^6}\sum_{i=1}^n (x_i - \mu)^2 \\ \frac{\partial^2}{\partial \sigma^2 \partial \mu} \log L(\mu, \sigma^2 \mid \mathbf{x}) &= -\frac{1}{\sigma^4}\sum_{i=1}^n (x_i - \mu) \end{aligned}$$

Example cont.

Hessian matrix evaluated at the MLE solution:

$$\begin{aligned} H|_{\hat{\mu}_{mle}, \hat{\sigma}^2_{mle}} &= \begin{bmatrix} \frac{\partial^2}{\partial \mu^2} \log L(\mu, \sigma^2) & \frac{\partial^2}{\partial \mu \partial \sigma^2} \log L(\mu, \sigma^2) \\ \frac{\partial^2}{\partial \sigma^2 \partial \mu} \log L(\mu, \sigma^2) & \frac{\partial^2}{\partial (\sigma^2)^2} \log L(\mu, \sigma^2) \end{bmatrix} \\ &= \begin{bmatrix} -\frac{n}{\hat{\sigma}^2_{mle}} & 0 \\ 0 & -\frac{n}{2\hat{\sigma}^4_{mle}} \end{bmatrix} \end{aligned}$$

Covariance matrix of MLE:

$$\widehat{\Sigma} = -\boldsymbol{H}^{-1} = \begin{bmatrix} \frac{\hat{\sigma}^2_{mle}}{n} & 0 \\ 0 & \frac{2\hat{\sigma}^4_{mle}}{n} \end{bmatrix}$$

ML Estimation of Linear Model

- Consider the linear regression model where the error terms follow multivariate normal distribution:

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{u}, \quad \boldsymbol{u} \sim N(\boldsymbol{0},\ \sigma^2 \boldsymbol{I}_n)$$

- Multivariate Normal Density for $\boldsymbol{u}$

$$f(\boldsymbol{u}) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left(-\frac{1}{2\sigma^2}\boldsymbol{u}^\top \boldsymbol{u}\right)$$

- Conditional density of $\boldsymbol{y}$

$$f(\boldsymbol{y}|\boldsymbol{X}) = f(\boldsymbol{u})\left|\frac{\partial \boldsymbol{u}}{\partial \boldsymbol{y}}\right| = f(\boldsymbol{u}), \quad \text{since } \left|\frac{\partial \boldsymbol{u}}{\partial \boldsymbol{y}}\right| = \boldsymbol{I}_n$$

where $\left|\frac{\partial \boldsymbol{u}}{\partial \boldsymbol{y}}\right|$ is the absolute value of the determinant formed from $n \times n$ matrix of partial derivatives of the elements of $\boldsymbol{u}$ with respect to $\boldsymbol{y}$.

ML Estimation of Linear Model

The loglikelihood is given by

$$\begin{aligned}
\log L(\boldsymbol{\theta}; \boldsymbol{y}) &= \log f(\boldsymbol{u}) \\
&= -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}\boldsymbol{u}^\top \boldsymbol{u} \\
&= c - \frac{n}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}(\boldsymbol{y} - \boldsymbol{X\beta})^\top(\boldsymbol{y} - \boldsymbol{X\beta})
\end{aligned} \tag{9}$$

where $c = -\frac{n}{2}\log(2\pi)$ and $\boldsymbol{\theta}^\top = (\boldsymbol{\beta}^\top, \sigma^2)$.
FOC (score vector):

$$\begin{aligned}
\frac{\partial \log L(\boldsymbol{\theta})}{\partial \boldsymbol{\beta}} &= -\frac{1}{\sigma^2}(-\boldsymbol{X}^\top \boldsymbol{y} + \boldsymbol{X}^\top \boldsymbol{X\beta}) = 0 \\
\frac{\partial \log L(\boldsymbol{\theta})}{\partial \boldsymbol{\sigma}^2} &= -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}(\boldsymbol{y} - \boldsymbol{X\beta})^\top(\boldsymbol{y} - \boldsymbol{X\beta}) = 0
\end{aligned} \tag{10}$$

ML Estimation of Linear Model

- Solving FOC:

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{X}^\top \boldsymbol{y} \equiv \hat{\boldsymbol{\beta}}_{OLS}$$

$$\hat{\sigma}^2 = \frac{1}{n}(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})^\top(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}) = \frac{1}{n}\hat{\boldsymbol{u}}^\top \hat{\boldsymbol{u}}$$

Note that,

$$\mathsf{E}\left(\frac{\hat{\boldsymbol{u}}^\top \hat{\boldsymbol{u}}}{n-k}\right) = \sigma^2 \ \Rightarrow\ \mathsf{E}(\hat{\sigma}^2) = \frac{\sigma^2(n-k)}{n}$$

ML estimator of the error variance, $\hat{\sigma}^2$, is biased but consistent.

ML Estimation of Linear Model

2nd Order Conditions and Expectations:

$$\frac{\partial^2 \log L}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^\top} = -\frac{\boldsymbol{X}^\top \boldsymbol{X}}{\sigma^2}, \quad -\mathsf{E}\left(\frac{\partial^2 \log L}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^\top}\right) = \frac{\boldsymbol{X}^\top \boldsymbol{X}}{\sigma^2}$$

$$\frac{\partial^2 \log L}{\partial \boldsymbol{\beta} \partial \sigma^2} = -\frac{\boldsymbol{X}^\top \boldsymbol{u}}{\sigma^4}, \quad -\mathsf{E}\left(\frac{\partial^2 \log L}{\partial \boldsymbol{\beta} \partial \sigma^2}\right) = 0$$

$$\frac{\partial^2 \log L}{\partial (\sigma^2)^2} = \frac{n}{2\sigma^4} - \frac{\boldsymbol{u}^\top \boldsymbol{u}}{\sigma^6}, \quad -\mathsf{E}\left(\frac{\partial^2 \log L}{\partial (\sigma^2)^2}\right) = \frac{n}{2\sigma^4}$$

since $\mathsf{E}(\boldsymbol{u}^\top \boldsymbol{u}) = n\sigma^2$

ML Estimation of Linear Model

- Information matrix:

$$\boldsymbol{I}(\boldsymbol{\theta}) = \boldsymbol{I}(\boldsymbol{\beta}, \sigma^2) = \begin{bmatrix} \frac{1}{\sigma^2}\boldsymbol{X}^\top \boldsymbol{X} & 0 \\ 0 & \frac{n}{2\sigma^4} \end{bmatrix}$$

- Inverse of the information matrix is the covariance matrix of $\hat{\boldsymbol{\theta}}$

$$\hat{\boldsymbol{\Sigma}} = \boldsymbol{I}^{-1} = \begin{bmatrix} \sigma^2(\boldsymbol{X}^\top \boldsymbol{X})^{-1} & 0 \\ 0 & \frac{2\sigma^4}{n} \end{bmatrix}$$

Note that $\hat{\boldsymbol{\beta}}$ and $\hat{\sigma}^2$ are distributed independently.

ML Estimation of Linear Model

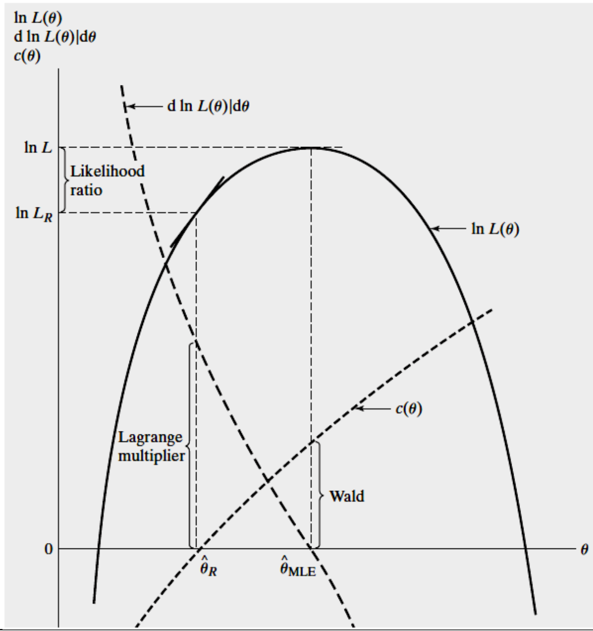
- Substituting $\hat{\boldsymbol{\beta}}$ and $\hat{\sigma}^2$ into the loglikelihood function and exponentiating we obtain

$$\begin{aligned} L(\hat{\boldsymbol{\beta}}, \hat{\sigma}^2) &= (2\pi e)^{-n/2}(\hat{\sigma}^2)^{-n/2} \\ &= \left(\frac{2\pi e}{n}\right)^{-n/2} (\hat{\boldsymbol{u}}^\top \hat{\boldsymbol{u}})^{-n/2} \\ &= constant \cdot (\hat{\boldsymbol{u}}^\top \hat{\boldsymbol{u}})^{-n/2} \end{aligned}$$

Hypothesis Testing in MLE (Greene, p.525)

- Let $H_0 : c(\theta) = 0$ be the restriction we want to test.
- **Likelihood Ratio Test**: If the restriction $c(\theta) = 0$ is valid, then imposing it should not lead to a large reduction in the log-likelihood function. Therefore, the test is based on the difference, $\log L_U - \log L_R$, where L_U is the value of the likelihood function at the unconstrained value and L_R is the value of the likelihood function at the restricted estimate.
- **Wald Test**: If the restriction is valid, then $c(\hat{\theta}_{mle})$ should be close to zero because the MLE is consistent. Therefore, the test is based on $c(\hat{\theta}_{mle})$. We reject the hypothesis if this value is significantly different from zero.
- **Lagrange Multiplier Test**: If the restriction is valid, then the restricted estimator should be near the point that maximizes the log-likelihood. Therefore, the slope of the log-likelihood function should be near zero at the restricted estimator. The test is based on the slope of the log-likelihood at the point where the function is maximized subject to the restriction.

Wald, LM and LR Tests in MLE



Likelihood Ratio (LR) Test

- Let $H_0 : R(\boldsymbol{\theta}) = r$ be the set of restrictions we want to test. Note that $R(\boldsymbol{\theta})$ can be nonlinear.
- Likelihood Ratio test relies on the estimation of both restricted and unrestricted models.
- L_u: unrestricted likelihood, $\hat{\boldsymbol{\theta}}_u$: unrestricted MLE
- L_r: restricted likelihood, $\hat{\boldsymbol{\theta}}_r$: restricted MLE
- $L_r \leq L_u$
- LR is defined as
$$\lambda = \frac{L_r(\hat{\boldsymbol{\theta}}_r)}{L_u(\hat{\boldsymbol{\theta}}_u)}$$
- Null hypothesis will be rejected if λ is small enough...but how small?

Likelihood Ratio (LR) Test

- Likelihood Ratio
$$\lambda = \frac{L_r(\hat{\boldsymbol{\theta}}_r)}{L_u(\hat{\boldsymbol{\theta}}_u)}$$
- For large samples
$$LR = -2\log\lambda = 2(\log L_u - \log L_r) \sim \ \chi^2_q$$
where q is the number of restrictions.
- H_0 will be rejected if LR is larger than the appropriate chi-squared critical value.
- Practical shortcoming of LR test: it requires the estimation of both restricted and unrestricted models. The restricted model may be difficult to estimate.

Wald Test

- Only unrestricted MLE is required.
- If $H_0 : R(\boldsymbol{\theta}) = \boldsymbol{r}$ is valid then the unrestricted $\hat{\boldsymbol{\theta}}_u$ should satisfy them. Otherwise, $R(\boldsymbol{\theta}) - \boldsymbol{r}$ significantly larger than zero. There are q restrictions in $R(\boldsymbol{\theta}) - \boldsymbol{r} = \mathbf{0}$.
- The Wald test is based on the normal quadratic form, i.e., if $\boldsymbol{x} \sim \ N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ then
$$(\boldsymbol{x} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{x} - \boldsymbol{\mu}) \sim \ \chi^2_q$$
- Note that, if the hypothesis $\mathsf{E}(\boldsymbol{x}) = \boldsymbol{\mu}$ is false the quadratic form will have a larger value than it would if it were true.

Wald Test

- Similarly, the Wald test statistic can be computed using
$$W = \left(R(\hat{\boldsymbol{\theta}}) - \boldsymbol{r}\right)^\top Avar\left(R(\hat{\boldsymbol{\theta}}) - \boldsymbol{r}\right)^{-1}\left(R(\hat{\boldsymbol{\theta}}) - \boldsymbol{r}\right) \sim \chi^2_q$$
where
$$Avar(R(\hat{\boldsymbol{\theta}}) - \boldsymbol{r}) = \left(\frac{\partial R(\hat{\boldsymbol{\theta}})}{\partial \hat{\boldsymbol{\theta}}^\top}\right) Avar(\hat{\boldsymbol{\theta}}) \left(\frac{\partial R(\hat{\boldsymbol{\theta}})}{\partial \hat{\boldsymbol{\theta}}^\top}\right)^\top$$
is the asymptotic covariance matrix. Note that $\frac{\partial R(\hat{\boldsymbol{\theta}})}{\partial \hat{\boldsymbol{\theta}}^\top}$ is $q \times k$ matrix of first derivatives of restrictions wrt parameters.

Wald Test

- Common form of restrictions is linear in parameters: $H_0 : \boldsymbol{R}\boldsymbol{\theta} = \boldsymbol{r}$.
- In this case
$$\frac{\partial R(\hat{\boldsymbol{\theta}})}{\partial \hat{\boldsymbol{\theta}}^\top} = \boldsymbol{R}$$
- We know
$$\hat{\boldsymbol{\theta}} \xrightarrow{d} N\left(\boldsymbol{\theta}, \boldsymbol{I}(\boldsymbol{\theta})^{-1}\right)$$
- This means that
$$(\boldsymbol{R}\hat{\boldsymbol{\theta}} - \boldsymbol{r}) \xrightarrow{d} N\left(\boldsymbol{0}, \boldsymbol{R}\boldsymbol{I}(\boldsymbol{\theta})^{-1}\boldsymbol{R}^\top\right)$$
- The Wald statistic becomes
$$W = (\boldsymbol{R}\hat{\boldsymbol{\theta}} - \boldsymbol{r})^\top \left(\boldsymbol{R}\boldsymbol{I}(\boldsymbol{\theta})^{-1}\boldsymbol{R}^\top\right)^{-1} (\boldsymbol{R}\hat{\boldsymbol{\theta}} - \boldsymbol{r}) \overset{a}{\sim} \chi^2_q$$

Wald Test

- In the classical normal linear regression model the inverse of the information matrix is given by
$$\boldsymbol{I}^{-1}(\boldsymbol{\beta}) = \begin{bmatrix} \sigma^2(\boldsymbol{X}^\top\boldsymbol{X})^{-1} & 0 \\ 0 & \frac{2\sigma^4}{n} \end{bmatrix}$$
- In order to test the linear restrictions of the form $H_0 : \boldsymbol{R}\boldsymbol{\beta} = \boldsymbol{r}$, the Wald test becomes
$$W = (\boldsymbol{R}\hat{\boldsymbol{\beta}} - \boldsymbol{r})^\top \left(\boldsymbol{R}\sigma^2(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{R}^\top\right)^{-1} (\boldsymbol{R}\hat{\boldsymbol{\beta}} - \boldsymbol{r}) \overset{a}{\sim} \chi^2_q$$
- Substituting $\hat{\sigma}^2 = \hat{\boldsymbol{u}}^\top\hat{\boldsymbol{u}}/n$ we obtain
$$W = \frac{1}{\hat{\sigma}^2}(\boldsymbol{R}\hat{\boldsymbol{\beta}} - \boldsymbol{r})^\top \left(\boldsymbol{R}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{R}^\top\right)^{-1} (\boldsymbol{R}\hat{\boldsymbol{\beta}} - \boldsymbol{r}) \overset{a}{\sim} \chi^2_q$$

Lagrange Multiplier (LM) Test

- Relies only on the restricted MLE
- Imposing restrictions, constrained maximization problem can be written as a Lagrangean function
$$\log L^*(\boldsymbol{\theta}) = \log L(\boldsymbol{\theta}) + \boldsymbol{\lambda}^\top(R(\boldsymbol{\theta}) - \boldsymbol{r})$$
where $\boldsymbol{\lambda}$ is $q \times 1$ vector of Lagrange multipliers.
- FOC:
$$\begin{aligned} \frac{\partial \log L^*}{\partial \boldsymbol{\theta}} &= \frac{\partial \log L}{\partial \boldsymbol{\theta}} + \left(\frac{\partial R(\hat{\boldsymbol{\theta}})}{\partial \hat{\boldsymbol{\theta}}^\top}\right)^\top \boldsymbol{\lambda} = \boldsymbol{0} \\ \frac{\partial \log L^*}{\partial \boldsymbol{\lambda}} &= R(\boldsymbol{\theta}) - \boldsymbol{r} = \boldsymbol{0} \end{aligned}$$
- If imposing restrictions do not lead to significant difference in the maximized value of the loglikelihood then $\left(\frac{\partial R(\hat{\boldsymbol{\theta}})}{\partial \hat{\boldsymbol{\theta}}^\top}\right)^\top \boldsymbol{\lambda}$ should be close to zero.

Lagrange Multiplier (LM) Test

- At the restricted maximum, derivatives (scores) are
$$\frac{\partial \log L_r}{\partial \hat{\boldsymbol{\theta}}_r} = s(\hat{\boldsymbol{\theta}}_r) = -\left(\frac{\partial R(\hat{\boldsymbol{\theta}})}{\partial \hat{\boldsymbol{\theta}}^\top}\right)^\top \boldsymbol{\lambda}$$
- This implies that if the restrictions are valid the derivatives will be (approximately) zero.
- The LM test is also called the score test since it is based on the first derivatives.
- Since the covariance matrix of the scores is the information matrix the LM test statistic is
$$LM = s(\hat{\boldsymbol{\theta}}_r)^\top \boldsymbol{I}(\hat{\boldsymbol{\theta}}_r)^{-1} s(\hat{\boldsymbol{\theta}}_r) \overset{a}{\sim} \chi^2_q$$
- Note that both score vector and information matrix are evaluated at the restricted parameters.

Lagrange Multiplier (LM) Test

- It can be shown that in the classical linear regression model the LM test statistic can be computed using
$$LM = \frac{n\hat{\boldsymbol{u}}_r^\top \boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\hat{\boldsymbol{u}}_r}{\hat{\boldsymbol{u}}_r^\top\hat{\boldsymbol{u}}_r}$$
where $\hat{\boldsymbol{u}}_r$ is the vector of residuals computed from the restricted model, and $\hat{\boldsymbol{u}}_r^\top\hat{\boldsymbol{u}}_r$ is the restricted sum of squared residuals (SSR).
- In the classical regression model, the LM test statistic for linear restrictions can easily be tested in two steps:
  1. Compute restricted $\hat{\boldsymbol{\beta}}_r$ and $\hat{\boldsymbol{u}}_r$
  2. Regress $\hat{\boldsymbol{u}}_r$ on all of the variables in $\boldsymbol{X}$ and compute
  $$LM = nR^2_{\hat{\boldsymbol{u}}_r} \overset{a}{\sim} \chi^2_q$$
  where $R^2_{\hat{\boldsymbol{u}}_r}$ is the coefficient of determination from the second step.
- As an example, see heteroscedasticity tests (Breusch-Pagan, White)

Tests in MLE framework

- All tests are asymptotically equivalent.
- But in the linear model they can give different results in small samples:
$$W \geq LR \geq LM$$
- See numerical example in Greene, p. 531.

Numerical Computation of MLE

- Statistical packages (STATA, Eviews, SAS, etc.) may have predefined MLE routines.
- However, in some cases MLE requires special programming.
- One can use MATLAB or STATA to carry out computations. We can use `fminunc` or `fmincon` in MATLAB after we coded the loglikelihood.
- As we saw in NLS framework essential knowledge of numerical optimization methods is inevitable (Quasi-Newton, BHHH, etc.).
- In STATA, we can use special ml routine for MLE whose general syntax is given below:

```
ml model    method progname eq [eq ...] [if] [in] [weight] [,
model_options svy diparm_options]
```

- See STATA manual.

MLE in STATA: Example

MLE of λ in the Poisson distribution:
$$f(y_i, \lambda) = \frac{e^{-\lambda}\lambda^{y_i}}{y_i!}, \quad y_i = 1, 2, 3, \ldots$$
The loglikelihood is given by
$$\log L(\lambda) = -n\lambda + \log(\lambda)\sum_{i=1}^{n} y_i - \sum_{i=1}^{n}\log(y_i!)$$

```
program drop poisson
program define poisson
args lnf lambda
quietly replace ‘lnf’ =  - ‘lambda’ + $ML_y*ln(‘lambda’)
                              - lnfactorial($ML y1)
end
```

Note that individual likelihood at each observation is coded.

MLE in STATA: Poisson Example

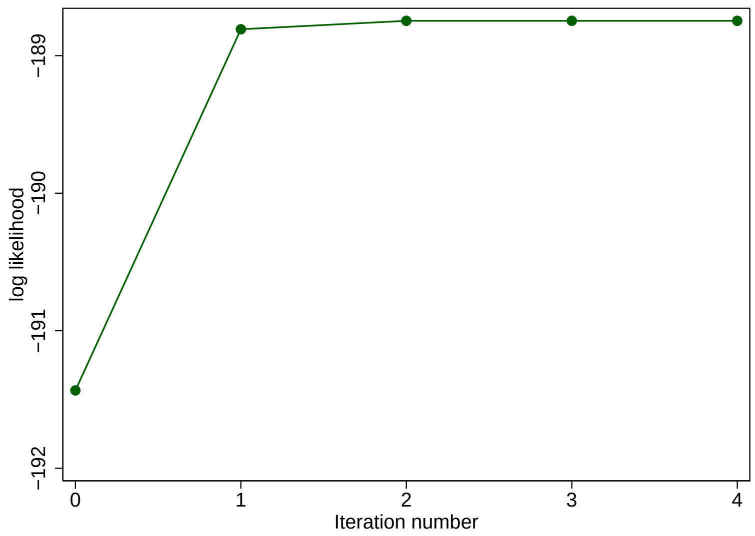
```
ml model lf poisson (y1=)
ml check
ml maximize
ml graph

initial:       log likelihood = -228.87426
rescale:       log likelihood = -191.43388
Iteration 0:   log likelihood = -191.43388
Iteration 1:   log likelihood =   -188.808
Iteration 2:   log likelihood = -188.74652
Iteration 3:   log likelihood =  -188.7465
Iteration 4:   log likelihood =  -188.7465

                                                  Number of obs   =        100
                                                  Wald chi2(0)    =          .
Log likelihood =  -188.7465                       Prob > chi2     =          .

------------------------------------------------------------------------------
          y1 |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       _cons |       3.03    .174069    17.41   0.000     2.688831    3.371169
------------------------------------------------------------------------------
```

MLE in STATA: Poisson Example



MLE in STATA: Normal Distribution Example

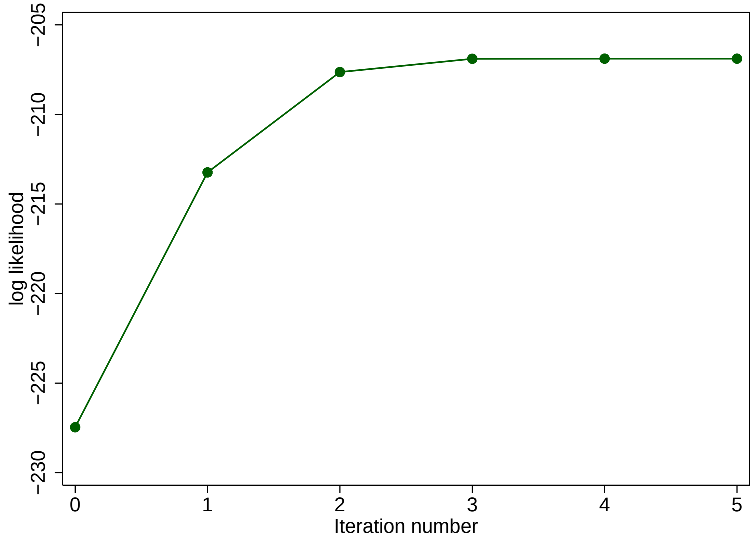
```
program drop _all
program define normal1
args lfn mu sigmasq
qui replace `lfn' = -0.5*log(2*_pi) - 0.5*log(`sigmasq')
                        -0.5*(($ML_y -`mu')^2)/`sigmasq'
end

. ml model lf normal1 (mu: y2= ) (sigma:)
. ml check
. ml maximize
. ml graph

                                                  Number of obs   =        100
                                                  Wald chi2(0)    =          .
Log likelihood = -206.88747                       Prob > chi2     =          .

------------------------------------------------------------------------------
          y2 |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
mu           |
       _cons |   5.086833   .1915419    26.56   0.000     4.711418    5.462248
-------------+----------------------------------------------------------------
sigma        |
       _cons |   3.668829   .5188507     7.07   0.000       2.6519    4.685757
------------------------------------------------------------------------------
```

MLE in STATA: Normal Example

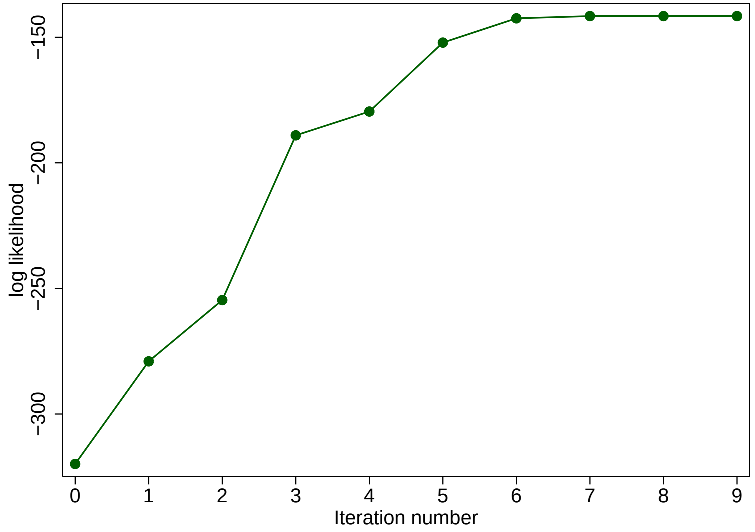


MLE in STATA: Normal Linear Regression Example

```
program drop _all
program define normal1
args lfn mu sigmasq
qui replace `lfn' = -0.5*log(2*_pi) - 0.5*log(`sigmasq')
                    -0.5*(($ML_y -`mu')^2)/`sigmasq'
end
. ml model lf normal1 (mu: y3 = x1 x2) (sigma:)
. ml check
. ml maximize
. ml graph
                                         Number of obs   =        100
                                         Wald chi2(2)    =    2883.18
Log likelihood = -141.57589              Prob > chi2     =     0.0000

------------------------------------------------------------------------------
          y3 |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
mu           |
          x1 |   2.172863   .3345717     6.49   0.000     1.517115    2.828612
          x2 |  -3.884038   .0731368   -53.11   0.000    -4.027383   -3.740692
       _cons |   .7908333   .2429265     3.26   0.001     .3147061     1.26696
-------------+----------------------------------------------------------------
sigma        |
       _cons |    .993661   .1405249     7.07   0.000     .7182373    1.269085
------------------------------------------------------------------------------
```

MLE in STATA: Normal Example



MLE in STATA: Compare MLE vs. OLS

Note that OLS and MLE estimates for σ^2 are different.

```
. reg y3 x1 x2

      Source |       SS       df       MS              Number of obs =     100
-------------+------------------------------           F(  2,    97) = 1398.34
       Model |  2864.90198     2  1432.45099           Prob > F      =  0.0000
    Residual |  99.3660943    97  1.02439272           R-squared     =  0.9665
-------------+------------------------------           Adj R-squared =  0.9658
       Total |  2964.26808    99  29.9421018           Root MSE      =  1.0121

------------------------------------------------------------------------------
          y3 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          x1 |   2.172863   .3397061     6.40   0.000     1.498641    2.847086
          x2 |  -3.884038   .0742591   -52.30   0.000    -4.031422   -3.736654
       _cons |   .7908333   .2466545     3.21   0.002     .3012925    1.280374
------------------------------------------------------------------------------
```