

# Maximum Likelihood Estimation

After finish these sessions, you should understand:

## Maximum Likelihood Estimation (MLE)

- MLE Concept
- MLE Computation Procedure
  - Non-linear Optimization
- Hypothesis Testing in NLS
  - Overall Test – Chi-squares Test
  - Individual Test – z-test

# Maximum Likelihood Estimation

## Motivation

### Disadvantages of Least Squares Methods

$$\hat{\beta} = (X'X)^{-1} X'y = \beta + (X'X)^{-1} X'\varepsilon$$

If  $\varepsilon$  is large, it will have impacts on estimated coefficients.

To solve this problem,

- Transform the data.
- Apply another distribution – nonnormal.

# Maximum Likelihood Estimation

## Maximum Likelihood Estimation

### Idea of Maximum Likelihood

- MLE stands for Maximum Likelihood Estimation
- Thus, the method is to try to maximize the likelihood function of the model.
- What is the likelihood function of the model?
- Distinctions between Probability Density function (*pdf*) and Likelihood function

# Maximum Likelihood Estimation

## Maximum Likelihood Estimation

Probability density function describes probability density for  $(y, X)$  treating  $\theta$  as given

Likelihood function describes the situation when treating  $X$  and  $y$  as given and treating  $\theta$  as variables.

$$L(\theta, y, X) \equiv p(y, X, \theta)$$

# Example

$$X_1 = 4 \text{ and } X_2 = 6$$

Estimate  $\mu$

**Assume:** Normal distribution and  $\sigma = 1$

$$\text{Pdf. } f(X) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{\left[-\frac{(X-\mu)^2}{2\sigma^2}\right]} = \frac{1}{\sqrt{2\pi}} e^{\left[-\frac{(X-\mu)^2}{2}\right]}$$

Likelihood function

$$L = \left( \frac{1}{\sqrt{2\pi}} e^{\left[-\frac{(4-\mu)^2}{2}\right]} \right) \left( \frac{1}{\sqrt{2\pi}} e^{\left[-\frac{(6-\mu)^2}{2}\right]} \right)$$

# Example

$\mu$	$p(4 \mu)$	$p(6 \mu)$	L	$\log L$
3.5	0.3520	0.0175	0.0062	-5.0883
4.0	0.3989	0.0540	0.0215	-3.8383
4.5	0.3520	0.1295	0.0456	-3.0883
4.6	0.3332	0.1497	0.0499	-2.9983
4.7	0.3122	0.1713	0.0535	-2.9283
4.8	0.2896	0.1941	0.0562	-2.8783
4.9	0.2660	0.2178	0.0579	-2.8483
5.0	0.2419	0.2419	0.0585	-2.8383
5.1	0.2178	0.2660	0.0579	-2.8483
5.2	0.1941	0.2896	0.0562	-2.8783
5.3	0.1713	0.3122	0.0535	-2.9283
5.4	0.1497	0.3332	0.0499	-2.9983
5.5	0.1295	0.3520	0.0456	-3.0883
6.0	0.0540	0.3989	0.0215	-3.8383
6.5	0.0175	0.3520	0.0062	-5.0883

# Example

$$L = -\frac{n}{2} \log(2\pi) - \frac{n}{2} \log \sigma^2 - \frac{1}{2\sigma^2} \sum_i (Y_i - \beta_1 - \beta_2 X_i)^2$$

## MLE

$$\beta_2 = 0.5091$$
$$\sigma^2 = 42.1591$$

$$\beta_1 = 24.4545$$

$$\log \text{Likelihood} = -16.1162294$$

$$\text{Likelihood} = 0.0000001002$$

## OLS Result

$$\beta_2 = 0.50909$$

$$\beta_1 = 24.4545$$

$$r^2 = 0.96206156$$

$$\text{var}(\beta_2) = 0.00128$$

$$\text{var}(\beta_1) = 41.1371$$

$$\sigma^2 = 42.15909091$$

$$\text{se}(\beta_2) = 0.03574$$

$$\text{se}(\beta_1) = 6.41382$$

$$\text{se} = 6.493003227$$

$$\log \text{Likelihood} = -16.1162294$$

# Maximum Likelihood Estimation

## Maximum Likelihood Estimation

Consider all cases:  $i = 1, 2, 3, \dots, n$  and assume independent for all  $x_i$

$$L(y, X, \theta) = f(y_i, X_i | \theta) = \prod_{i=1}^n f(y_i, X_i, \theta)$$

Loglikelihood function can be defined as

$$l(\theta) = \log(L(\theta))$$

For example:

$$l(\beta, \sigma^2) = -\frac{n}{2} \log(2\pi) - \frac{n}{2} \log \sigma^2 - \frac{1}{2\sigma^2} (y - X\beta)'(y - X\beta)$$

# Maximum Likelihood Computation

Maximum Likelihood Estimation

Numerical Aspects of Optimization

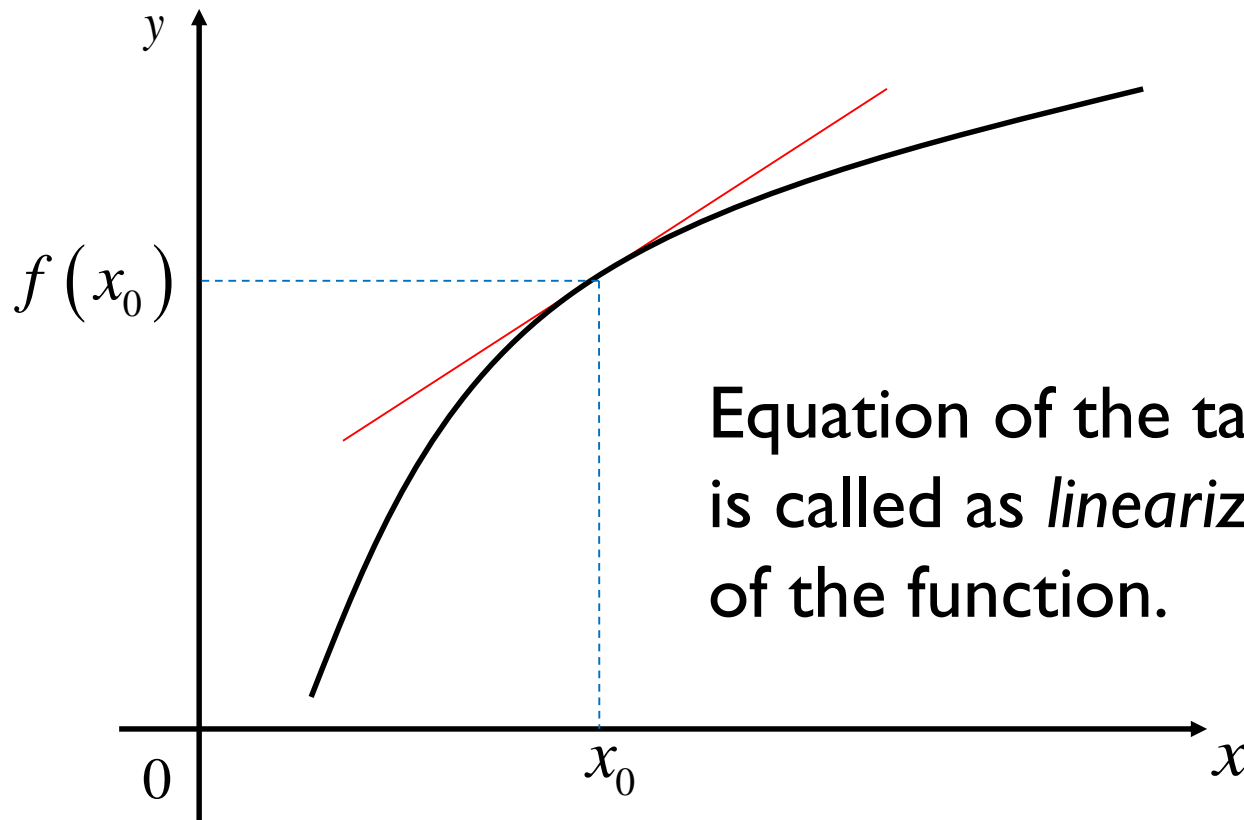
Non-linear Optimization:

1. Linearize non-linear function using Taylor Linear Approximation.
2. Min. or Max. the Linear-Approximation of the non-linear objective function → find the FOC.

# Maximum Likelihood Computation

## Linearize at Point $x_0$

For any function  $f(x)$ , the tangent is a close approximation of the function for some small distance from the tangent point.



# Maximum Likelihood Computation

## Linearize at Point $x_0$

Start with the point/slope equation:

$$y - y_0 = m(x - x_0) \quad y_0 = f(x_0) \quad m = f'(x_0)$$

$$y - f(x_0) = f'(x_0)(x - x_0)$$

$$y = f(x_0) + f'(x_0)(x - x_0)$$

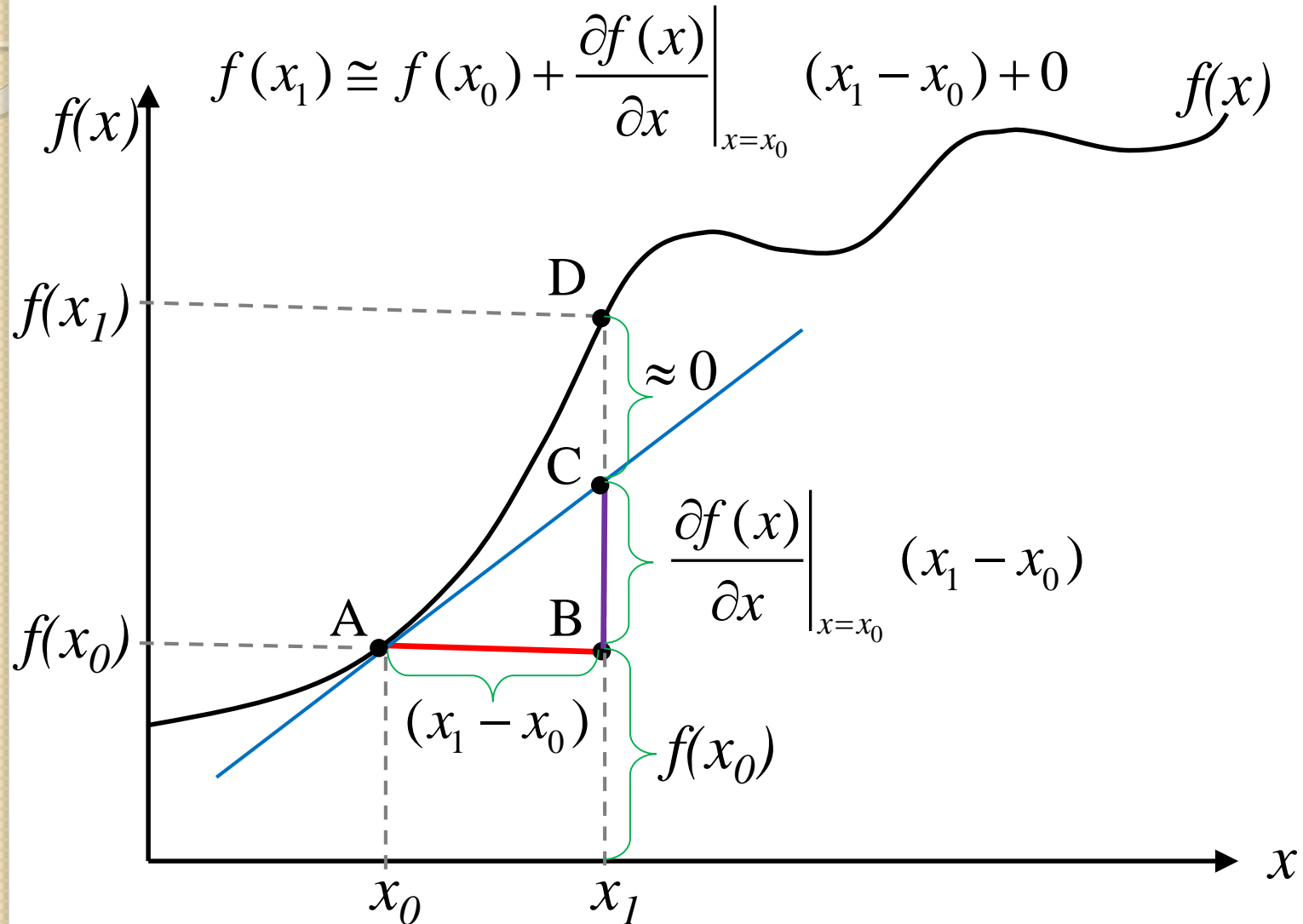
$$L(x) = f(x_0) + f'(x_0)(x - x_0) \text{ linearization of } f \text{ at } x_0$$

$f(x) \approx L(x)$  is standard linear approximation of  $f$  at  $x_0$

The linearization is the equation of the tangent line.

# Maximum Likelihood Computation

## Linear Approximation of $f(x)$ at Point $x_0$



# Maximum Likelihood Computation

## Maximum Likelihood Estimation

## Numerical Aspects of Optimization

## Non-linear Optimization:

Gradient:

$$G = \frac{\partial l(\theta)}{\partial \theta} = \begin{bmatrix} \frac{\partial l(\theta)}{\partial \theta_1} \\ \frac{\partial l(\theta)}{\partial \theta_2} \\ \vdots \\ \frac{\partial l(\theta)}{\partial \theta_k} \end{bmatrix}$$

Hessian:

$$H = \frac{\partial^2 l(\theta)}{\partial \theta \partial \theta'} = \begin{bmatrix} \frac{\partial^2 l(\theta)}{\partial \theta_1 \partial \theta_1} & \frac{\partial^2 l(\theta)}{\partial \theta_1 \partial \theta_2} & \dots & \frac{\partial^2 l(\theta)}{\partial \theta_1 \partial \theta_k} \\ \frac{\partial^2 l(\theta)}{\partial \theta_2 \partial \theta_1} & \frac{\partial^2 l(\theta)}{\partial \theta_2 \partial \theta_2} & \dots & \frac{\partial^2 l(\theta)}{\partial \theta_2 \partial \theta_k} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 l(\theta)}{\partial \theta_k \partial \theta_1} & \frac{\partial^2 l(\theta)}{\partial \theta_k \partial \theta_2} & \dots & \frac{\partial^2 l(\theta)}{\partial \theta_k \partial \theta_k} \end{bmatrix}$$

# Maximum Likelihood Computation

## Maximum Likelihood Estimation

To maximize the log-likelihood function

FOC: 
$$G(\theta) = \frac{\partial l(\theta)}{\partial \theta} = 0$$

Linearized: 
$$f(x_1) \cong f(x_0) + \left. \frac{\partial f(x)}{\partial x} \right|_{x=x_0} (x_1 - x_0)$$

$$G(\theta) \cong G(\theta_0) + \left. \frac{\partial G(\theta)}{\partial \theta} \right|_{\theta=\theta_0} (\theta - \theta_0)$$

$$G(\theta) \cong G(\theta_0) + H(\theta_0)(\theta - \theta_0)$$

# Maximum Likelihood Computation

First-order condition:  $G(\hat{\theta}) = 0$

Linearized around given value  $\theta_0$

$$G(\hat{\theta}) \cong G(\hat{\theta}_0) + H(\hat{\theta}_0)(\hat{\theta} - \hat{\theta}_0) = 0$$

Then,  $G(\hat{\theta}_0) = H(\hat{\theta}_0)\hat{\theta} - H(\hat{\theta}_0)\hat{\theta}_0$

$$H(\hat{\theta}_0)\hat{\theta} = H(\hat{\theta}_0)\hat{\theta}_0 - G(\hat{\theta}_0)$$

$$\hat{\theta} = H(\hat{\theta}_0)^{-1} H(\hat{\theta}_0)\hat{\theta}_0 - H(\hat{\theta}_0)^{-1} G(\hat{\theta}_0)$$

$$\hat{\theta} = \hat{\theta}_0 - H(\hat{\theta}_0)^{-1} G(\hat{\theta}_0)$$

**Newton-Raphson Algorithm**

$$\hat{\theta}_{t+1} = \hat{\theta}_t - H(\hat{\theta}_t)^{-1} G(\hat{\theta}_t)$$

# ML Computation – Algorithm

$$\hat{\theta}_{t+1} = \hat{\theta}_t - \Delta_t$$

Methods that employ line searches differ according to the choice of  $\delta$  and methods of approximating  $\lambda^*$

Methods of choosing  $\delta$  mostly set up by letting  $\lambda \geq 0$

Newton-Raphson  $\Delta_t = H(\hat{\theta}_t)^{-1} G(\hat{\theta}_t)$

## Quadratic Hill-climbing (Goldfeld-Quandt) Methods

$$\Delta = \tilde{H}^{-1} G(\hat{\theta}_t) \quad \text{where} \quad \tilde{H} = H(\hat{\theta}_t) + \gamma I$$

Newton  $\Delta = K^{-1}(\hat{\theta}_t) G(\hat{\theta}_t)$  where  $K = (GG') + H(\hat{\theta}_t)$

Guass-Newton or BHHH  $\Delta = (GG')^{-1} G(\hat{\theta}_t)$

Marquardt  $\Delta = (GG' + \gamma I)^{-1} G(\hat{\theta}_t)$

# Maximum Likelihood Computation

## Issues of Concern:

1. Initial Value – Global max vs Local max
2. Algorithm
3. Convergence Value – Default STATA =  $10^{-13}$
4. Maximum Iterative Times
  - Convergence achieved vs not achieved

# ML Properties

## Maximum Likelihood Estimation

### ML in Linear Model

$$y = X\beta + \varepsilon, \quad \varepsilon \sim N(0, \sigma^2 I)$$

Loglikelihood function:

$$l(\beta, \sigma^2) = -\frac{n}{2} \log(2\pi) - \frac{n}{2} \log \sigma^2 - \frac{1}{2\sigma^2} (y - X\beta)'(y - X\beta)$$

Maximize loglikelihood function by:

$$\frac{\partial l}{\partial \beta} = \frac{1}{\sigma^2} X'(y - X\beta) = 0$$

$$\frac{\partial l}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4} (y - X\beta)'(y - X\beta) = 0$$

# ML Properties

## Maximum Likelihood Estimation

### ML in Linear Model

$$\hat{\beta}_{ML} = (X'X)^{-1} X'y = \hat{\beta}_{OLS}$$

$$\hat{s}_{ML}^2 = \frac{1}{n} (y - X \hat{\beta})'(y - X \hat{\beta}) = \frac{n-k}{n} \hat{s}_{OLS}^2$$

By assuming normality assumption ML and OLS provide the same results – unbiased  $\beta$ , but not  $s^2$  -- only when  $n \rightarrow \infty$

### ML in Non-linear Regression Models

Assuming normality – NLS = ML

# ML Properties

## Asymptotic Properties

### Asymptotic Distribution of ML Estimators

$$\sqrt{n}(\hat{\theta}_{ML} - \hat{\theta}_0) \xrightarrow{d} N(0, I_0^{-1})$$

### Information Matrix:

$$I_n(\hat{\theta}_0) = E \left[ \frac{\partial l}{\partial \theta} \frac{\partial l}{\partial \theta'} \right] = -E \left[ \frac{\partial^2 l}{\partial \theta \partial \theta'} \right]$$

# ML Properties

## Asymptotic Properties

### Approximate Distribution for Finite Samples

$$\hat{\theta}_{ML} \approx N\left(\theta_0, I_n^{-1}(\hat{\theta}_{ML})\right)$$

### Information Matrix – Second order cond.:

$$\frac{\partial^2 l}{\partial \beta \partial \beta'} = -\frac{1}{\sigma^2} X'X$$

$$\frac{\partial^2 l}{\partial \beta \partial \sigma^2} = -\frac{1}{\sigma^4} X'(y - X\beta)$$

$$\frac{\partial^2 l}{\partial \sigma^2 \partial \sigma^2} = \frac{n}{2\sigma^4} - \frac{1}{\sigma^6} (y - X\beta)'(y - X\beta)$$

# ML Properties

## Asymptotic Properties

Approximate Distribution for Finite Samples

Follows independent assumption:

$$I_n(\hat{\theta}_0) = \begin{pmatrix} \frac{1}{\sigma^2} X'X & 0 \\ 0 & \frac{n}{2\sigma^4} \end{pmatrix}$$

Follows stability assumption:

$$I_n(\hat{\theta}_0) = \begin{pmatrix} \frac{1}{\sigma^2} Q & 0 \\ 0 & \frac{1}{2\sigma^4} \end{pmatrix}$$

# ML Properties

## Summary of Computations in ML

Step 1: Formulate the log-likelihood.

Step 2: Maximize the log-likelihood.

Step 3: Asymptotic tests.

# ML Inference

## Individual Test – Z-test

In OLS case, individual test is performed by using t-test because the estimated parameters are t-distributed.

In MLE case, distribution of the estimated parameters are not always t-distributed depending on distribution of the regression model. Thus, individual test can be performed by using **Z-test**:  $H_0: \beta_i = 0$

$$z\text{-test} \approx \frac{\hat{\beta}_i}{s_{\hat{\beta}_i}} \sim N(0,1)$$

# ML Inference

## Likelihood Ratio (LR) Test

Based on the loss of log-likelihood that results if the restrictions are imposed.

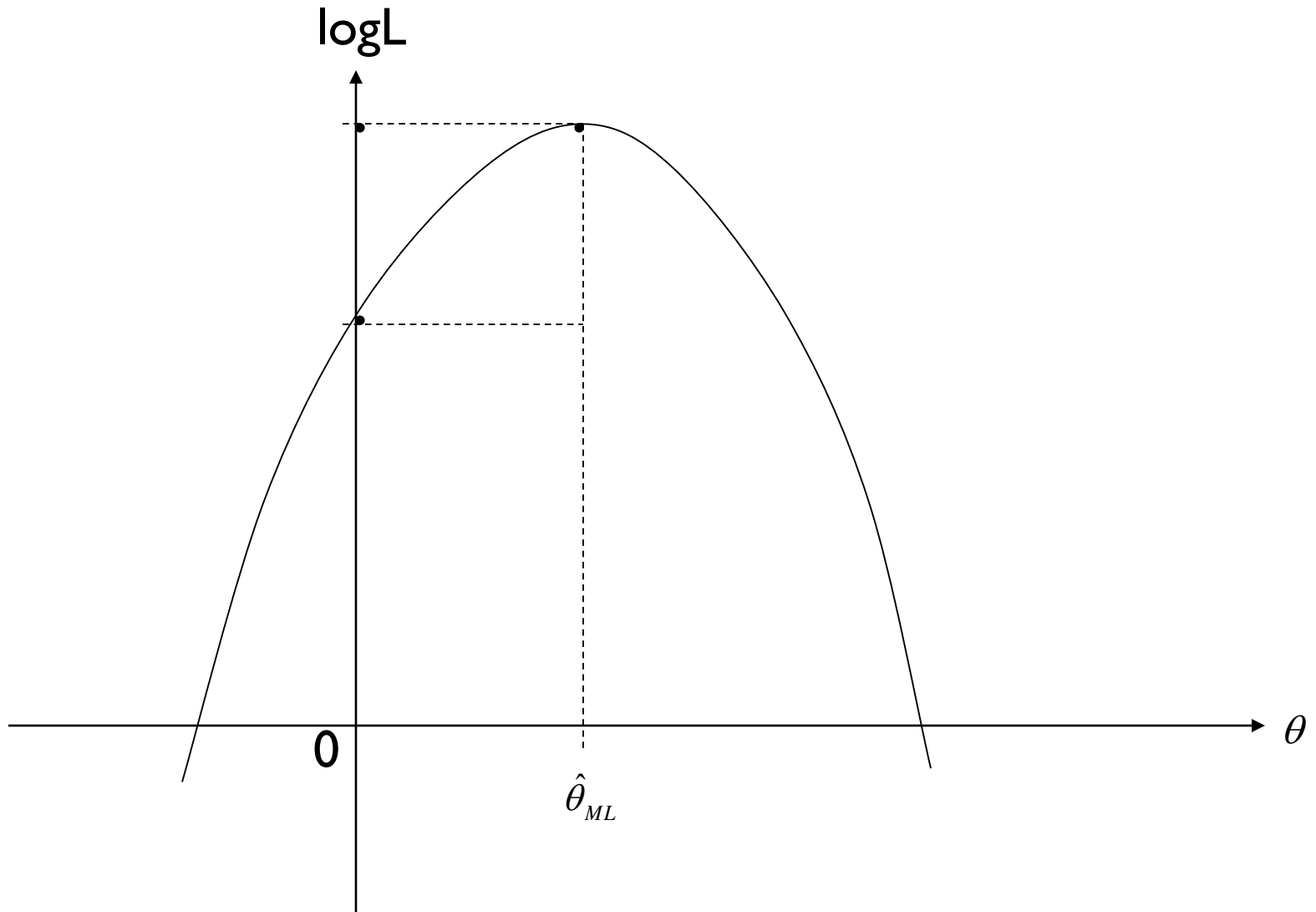
$$LR = 2\log(L(\hat{\theta}_1)) - 2\log(L(\hat{\theta}_0)) = 2l(\hat{\theta}_1) - 2l(\hat{\theta}_0)$$

$$H_0: \theta = 0$$

$$LR \xrightarrow{d} \chi^2(k-1)$$

# ML Inference

## Likelihood Ratio (LR) Test



# ML Inference

## Wald Test

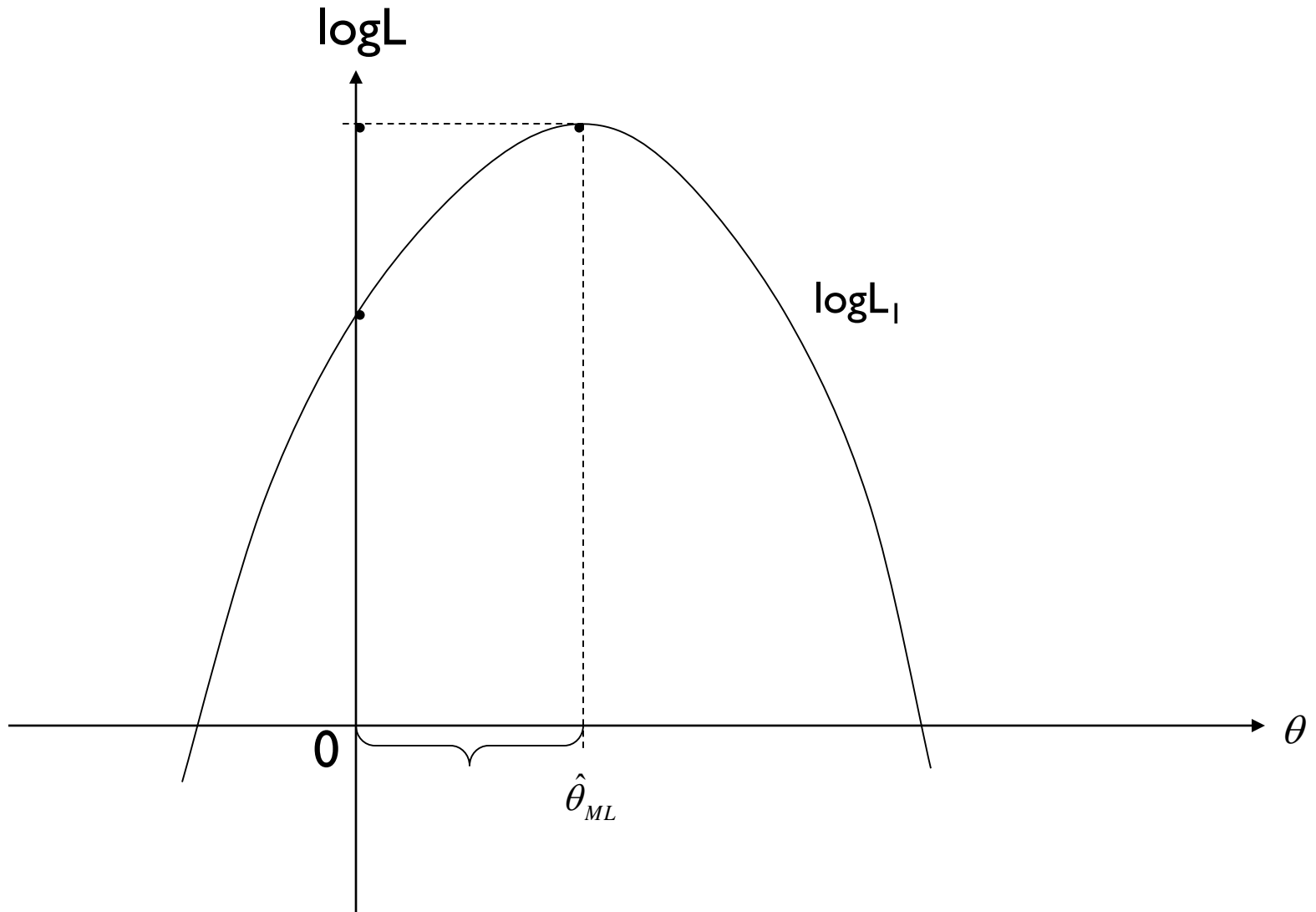
Based on unrestricted model alone.

$$H_0: \theta = 0$$

$$W = \hat{\theta}_1^2 \left( -\frac{d^2l}{d\theta^2} \right) \approx \left( \frac{\hat{\theta}_1}{s_{\hat{\theta}_1}} \right)^2 \approx \chi^2(k-1)$$

# ML Inference

## Wald Test



# ML Inference

## Lagrange Multiplier (LM) Test

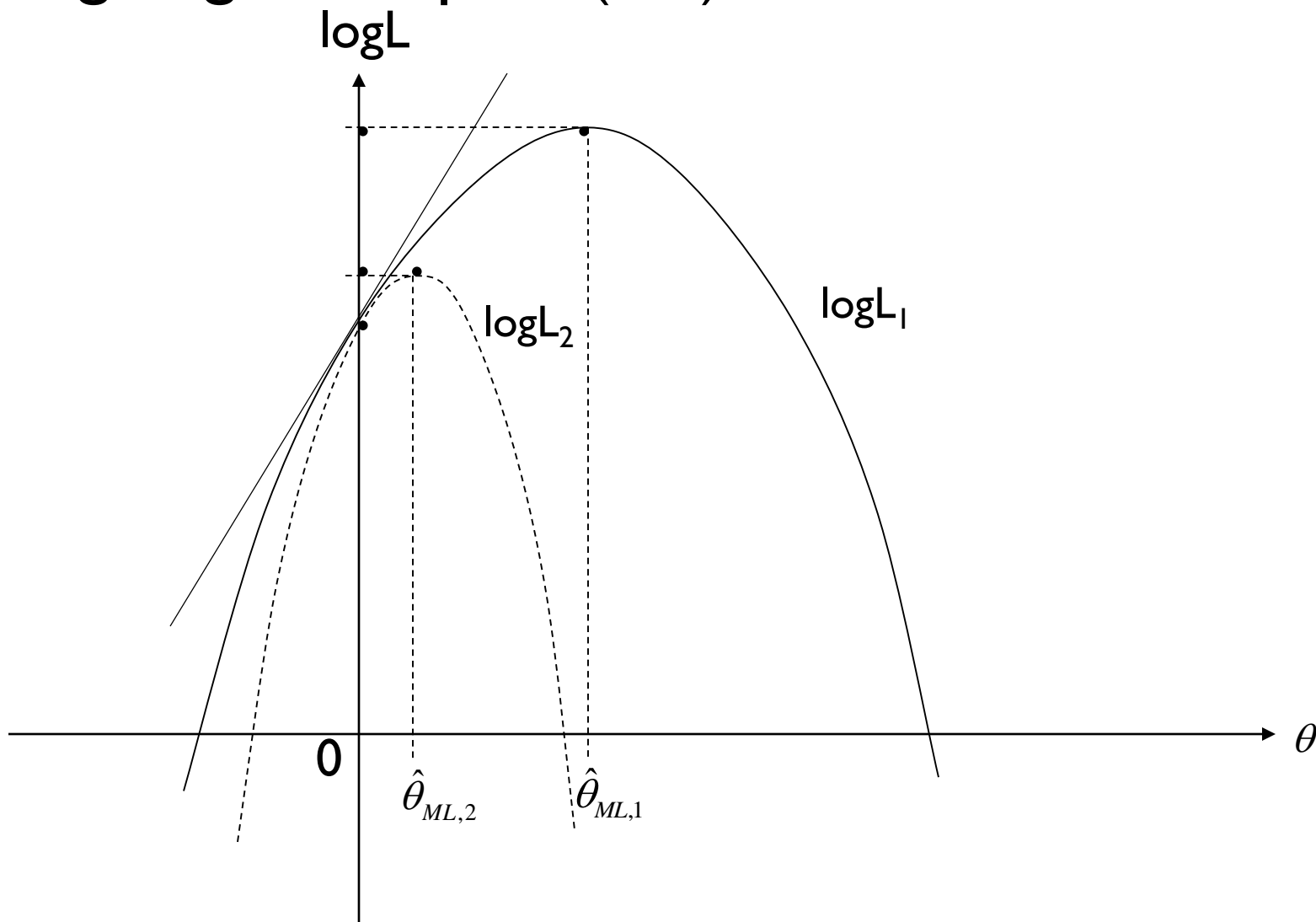
Score test considers whether the gradient (also called the 'score') of the unrestricted likelihood function is sufficiently close to zero at the restricted estimate  $\theta$ .

$$H_0: \theta = 0$$

$$LM = \frac{(\partial l / \partial \theta)^2}{-\partial^2 l / \partial \theta^2} = \left( \frac{\partial l}{\partial \theta} \right)' \left( -E \left[ \frac{\partial^2 l}{\partial \theta \partial \theta'} \right] \right)^{-1} \left( \frac{\partial l}{\partial \theta} \right) \approx \chi^2(k-1)$$

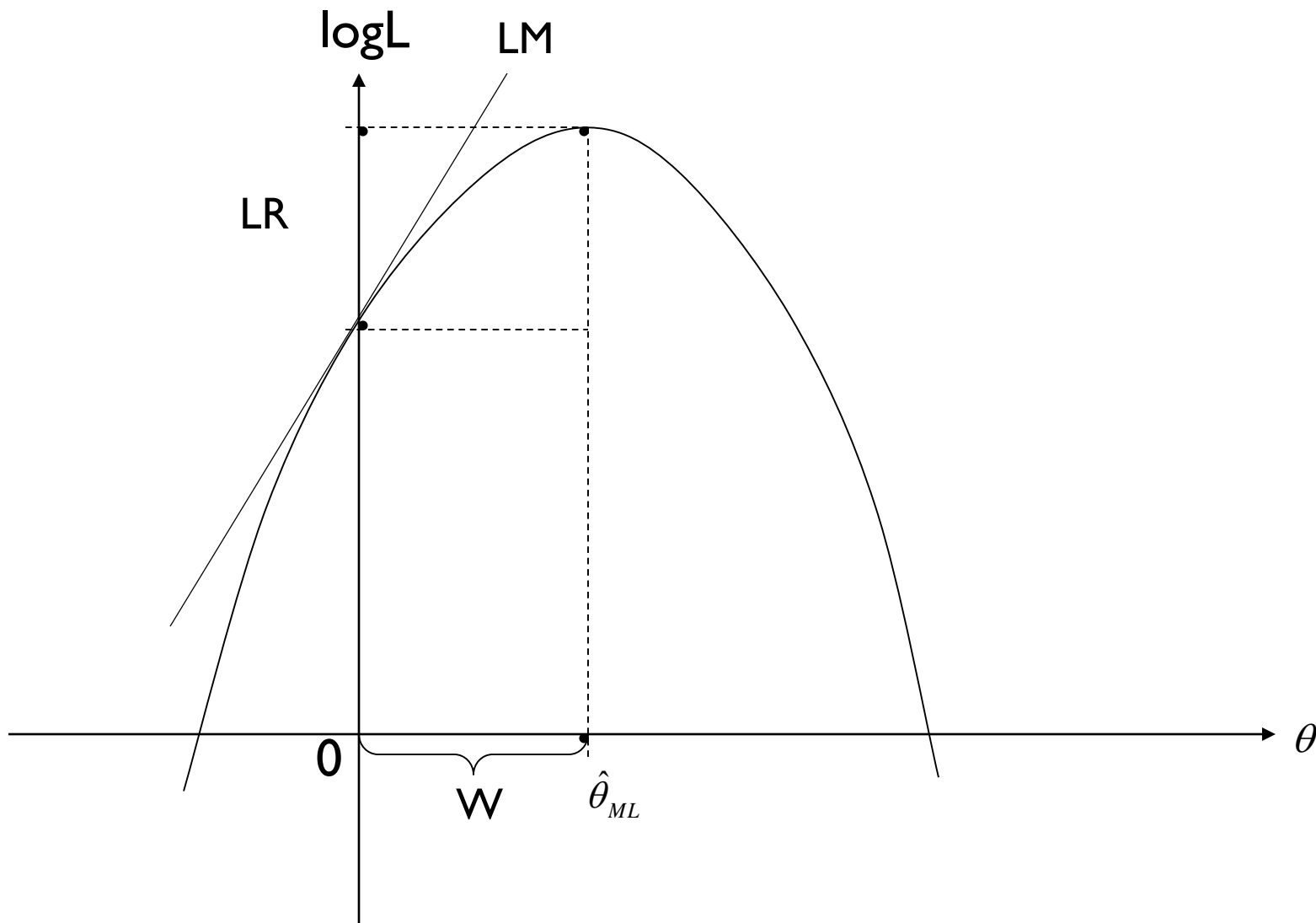
# ML Inference

## Lagrange Multiplier (LM) Test



# ML Inference

## Comparison of Three Tests



# ML Inference

## Comparison of Three Tests

	LR	Wald	LM
Estimated models	2 models	1 Unrestricted	1 Restricted
Advantage	Optimal power	Restricted model is complicated	Simple computations
Disadvantage	Needs 2 optimizations	Test depends on parameterization	Power may be small

$$LM \leq LR \leq W$$

# ML Evaluation Criteria

1. Sign and meaning of the Coefficients.
  - Whether the estimated results are according to the theory.
2. Overall Test – Chi-squares-test.
  - Whether all explanatory variables can be used in explaining the dependent variable.
3. Log-likelihood Value.
  - How well does the estimated results contribute to the likelihood of the model?
4. Individual Test – z-test.
  - Whether each explanatory variables can explain the dependent variable.

# ML Evaluation Criteria

OLS	MLE
1. Sign & Meaning	1. Sign & Meaning
2. Overall Test - F-test	2. Overall Test - Chi-square-test
3. R-squares	3. log-likelihood value
4. Individual Test - t-test	4. Individual Test - z-test