Lecture 3-4: Maximum Likelihood Estimation

Econometric Methods

Andrzej Torój

SGH Warsaw School of Economics – Institute of Econometrics

Outline

Outline

Let's roll dice (1)

Let's roll dice once.

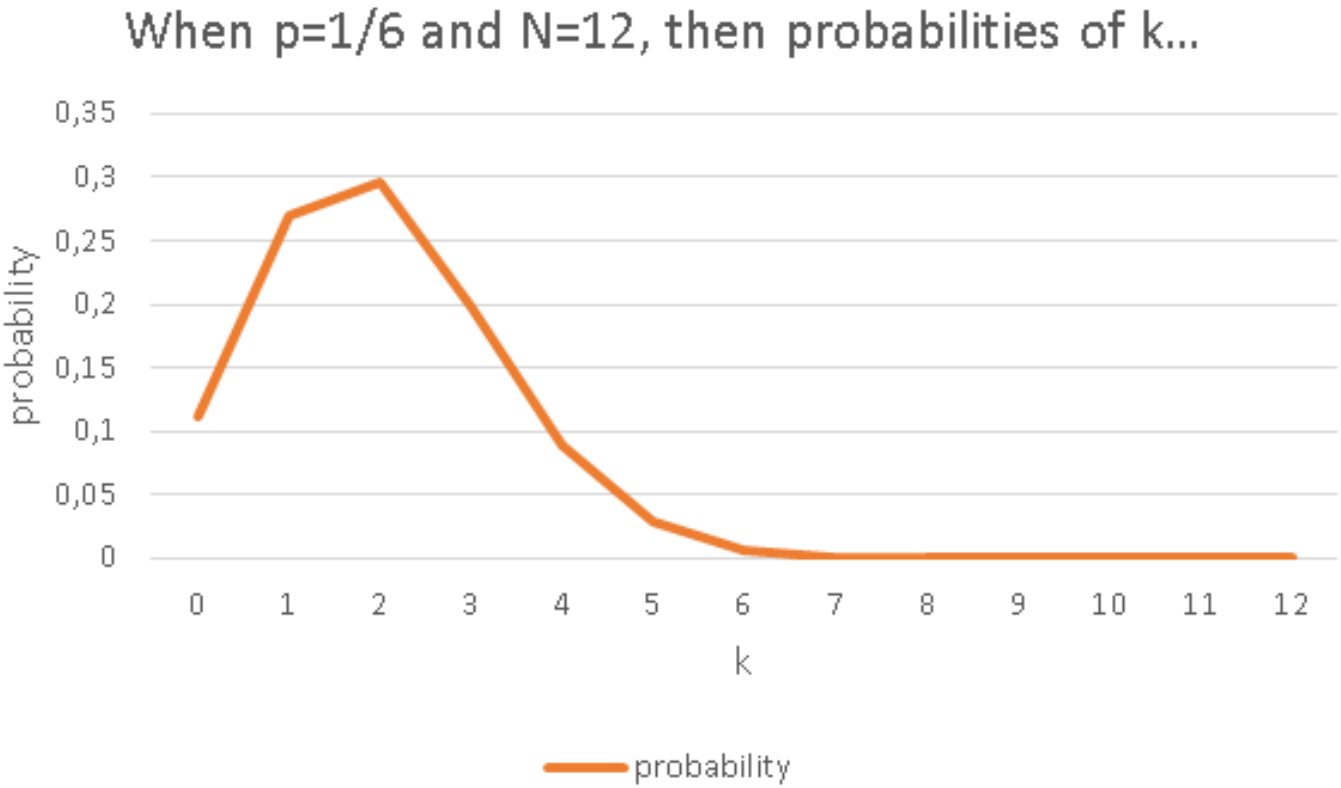
- Probability of getting ❻: $p = \frac{1}{6}$
- Probability of getting sth. else ✖: $1 - p = \frac{5}{6}$

Let's roll dice N times.

- Probability of getting ❻ k times (binomial / Bernoulli distribution):
  $P(k) = \binom{N}{k} p^k (1-p)^{N-k}$

Let's roll dice (2)

Knowing p (and, say, $N = 12$), we obtain a distribution of k.



Let's roll dice like an econometrician

- But that's not the way we proceed in econometrics. As econometricians,
  - we don't know p (true parameter value)
  - we know k (observations in the N-element sample)
- Rolling $N = 12$ times, we obtain the following sequence of draws: ✖✖✖✖⑥✖✖⑥✖✖✖⑥
  - What's the probability of getting this particular sample?
  - knowing p: $\left(\frac{5}{6}\right)^4 \frac{1}{6} \left(\frac{5}{6}\right)^2 \frac{1}{6} \left(\frac{5}{6}\right)^3 \frac{1}{6} = 0.0224$
  - without knowing p:
    $(1-p)^4 p (1-p)^2 p (1-p)^3 p = p^3 (1-p)^9 = ...$

Dice example: how search for p (1)

One way to estimate the parameter (p) is to ask oneself the question...

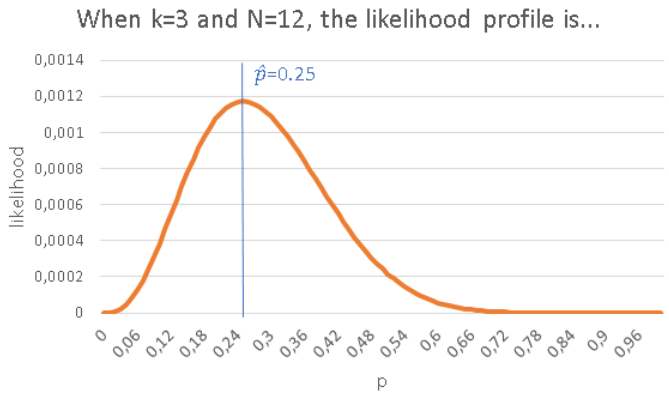
...what parameter value(s), $\hat{p}$, maximize the probability of observing the particular sample that we got?

- Underlying idea:
  - Since we got this particular sample, it was apparently the most likely one to obtain.
  - So let's look for parameter values that would really make it most likely.
- For example, if You play a game 1000 times and always lose, what's Your personal estimate of the probability of winning?

Dice example: how search for p (2)

Maximize, i.e. compute partial derivatives wrt. individual parameters and set them to 0:

- objective funciton: $L(p) = p^3 (1-p)^9$
- partial derivative: $\frac{\partial L}{\partial p} = 3p^2 (1-p)^9 - 9(1-p)^8 p^3$
- first-order condition: $\frac{\partial L}{\partial p} = 0 \Leftrightarrow 3\hat{p}^2 (1-\hat{p})^9 = 9(1-\hat{p})^8 \hat{p}^3$
- solution: $\hat{p} = \frac{1}{4}$



General formula for binomial cases

Note that, in this simple case, this result is super intuitive!

- 3 out of 12 sample elements were ❻
- This allows us to expect $\hat{p} = \frac{1}{4}$ even if we somehow know that $p = \frac{1}{6}$ (which we normally don't know as econometricians)
  $L(p, k, N) = p^k (1-p)^{N-k}$
  $\frac{\partial L}{\partial p} = kp^{k-1}(1-p)^{N-k} - (N-k)(1-p)^{N-k-1} p^k = 0$
  $k\hat{p}^{k-1}(1-\hat{p})^{N-k} = (N-k)(1-\hat{p})^{N-k-1}\hat{p}^k$
  $k(1-\hat{p}) = (N-k)\hat{p}$
  $k - k\hat{p} = N\hat{p} - k\hat{p}$
  $\hat{p} = \frac{k}{N}$
- ...the sample fraction of "wins" (i.e. rolling ❻, observing 1 rather than 0, etc.) is the estimate of p.

Sample Probability/Density vs Likelihood Function

- **Sample Probability/Density**: how likely it is to draw a given sample $\mathbf{y}$ from the data generating process, given its parameters $\boldsymbol{\theta}$?
  $P(\mathbf{y}|\boldsymbol{\theta})$

  We refer to this concept in **Bayesian Econometrics / Ekonometria Bayesowska**, treating the sample density as one of the inputs into the Bayes theorem formula: $P(\boldsymbol{\theta}|\mathbf{y}) = \frac{P(\mathbf{y}|\boldsymbol{\theta})P(\boldsymbol{\theta})}{P(\mathbf{y})}$. But even then, we actually mean...

- **Likelihood Function**:
  $L(\boldsymbol{\theta}|\mathbf{y}) = P(\mathbf{y}|\boldsymbol{\theta})$
  - i.e. it's mathematically the same, but interpreted more pragmatically as a function of parameters $\boldsymbol{\theta}$ (conditional on $\mathbf{y}$ – the only data sample available)

Likelihood Function

If the sample elements are ***independent*** (the standard case), the likelihood function is:

$$L(\theta|\mathbf{y}) = \prod_{i=1}^{N} p(\boldsymbol{\theta}|y_i)$$

i.e. the ***product*** of

- probability functions (for discrete variables) or
- density functions (for continuous variables)

of individual observations.

The non-standard case of mutually dependent observations is handled in **Spatial Econometrics / Ekonometria Przestrzenna**. The focus shifts then from single observation's probability/density to joint, multivariate probability/density of all observations.

Maximum Likelihood Estimation

1. Formulate the likelihood function for a given problem:
$$L(\boldsymbol{\theta}|\mathbf{y}) = \prod_{i=1}^{N} p(y_i|\boldsymbol{\theta})$$
2. Maximize L with respect to $\boldsymbol{\theta}$:
   1. **analytically**, if possible:
   $\frac{\partial L}{\partial \theta_1} = 0 \qquad \frac{\partial L}{\partial \theta_2} = 0 \qquad \ldots$
   2. **numerically**, otherwise:
   e.g. gradient descent or other, more advanced local or global methods
3. $\hat{\boldsymbol{\theta}}^{ML} = \arg\max_{\boldsymbol{\theta}} L(\boldsymbol{\theta}|\mathbf{y})$

Gradient Descent method

1. Start with $\boldsymbol{\theta_0}$ (initial guess).
2. Compute gradient (vector of partial derivatives) at $\boldsymbol{\theta_0}$:
   $\boldsymbol{\nabla_0} = \left[\begin{array}{ccc} \frac{\partial L}{\partial \theta_1}(\boldsymbol{\theta_0}) & \frac{\partial L}{\partial \theta_2}(\boldsymbol{\theta_0}) & \ldots \end{array} \right]^T$
3. Update $\boldsymbol{\theta_1} = \boldsymbol{\theta_0} + \gamma \boldsymbol{\nabla_0}$, where γ is an arbitrary small positive value.
4. Repeat (2)-(3) until stop criterion hit, e.g. $\boldsymbol{\theta_n} - \boldsymbol{\theta_{n-1}} < 0.0001$.

Exercise 1: write an R script maximizing the L function with respect to p in the dice-rolling example.

Numerical optimization

- Gradient descent is very simple (and inefficient). Here just for illustrative purposes.
- There are different routines that can be used in a black-box mode, with a minimum user input of indicating the starting point $\boldsymbol{\theta_0}$.
  - Like R's **optim** function.

But black-box is never good. To develop additional awareness and skills here: **Optimization Methods / Metody Optymalizacji**

Likelihood versus log-likelihood

- The principle of maximizing the likelihood function extends (under reasonable regularity conditions) to monotonic transformations thereof.
- In particular, $\arg\max_{\boldsymbol{\theta}} L(\theta|\mathbf{y}) = \arg\max_{\boldsymbol{\theta}} \ln L(\theta|\mathbf{y})$.
- In practical applications (and off-the-shelf software), the log-likelihood function is strongly preferred.
- Reason? Numerical stability and numerical precision.

Exercise 2. To see this, consider a simple example.
Draw $N = 1000$ observations from $N(0, 1)$. Next, compute:
1) likelihood – by multiplying all densities, $L = \prod f(x_i)$
2) log-likelihood – by adding all log-densities, $\ln L = \sum \ln f(x_i)$
What happened?

Outline

Linear regression (1): likelihood of $\boldsymbol{\varepsilon}$

$$\underbrace{\mathbf{y}}_{N\times 1} = \underbrace{\mathbf{X}}_{N\times k}\ \underbrace{\boldsymbol{\beta}}_{k\times 1} + \underbrace{\boldsymbol{\varepsilon}}_{N\times 1}$$

$$\boldsymbol{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_N \end{bmatrix} \qquad \varepsilon_i \sim N\left(0, \sigma^2\right)\ i.i.d. \qquad \mathbf{X} - \text{exogenous; } \mathbf{X},\mathbf{y} - \text{observable}$$

Density of i-th residual: $f\left(\varepsilon_i\right) = \left(\frac{1}{\sigma^2 2\pi}\right)^{\frac{1}{2}} e^{-\frac{\varepsilon_i^2}{2\sigma^2}}$

Likelihood: $L\left(\sigma^2|\boldsymbol{\varepsilon}\right) = \prod_{i=1}^{N} f\left(\varepsilon_i\right) = \left(\frac{1}{\sigma^2 2\pi}\right)^{\frac{N}{2}} e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{N}\varepsilon_i^2}$

Linear regression (2): likelihood of $\mathbf{y}$

Note that:

- $\sum_{i=1}^{N} \varepsilon_i^2 = \boldsymbol{\varepsilon}^T \boldsymbol{\varepsilon}$
- $\boldsymbol{\varepsilon} = \mathbf{y} - \mathbf{X}\boldsymbol{\beta}$

Then, likelihood: $L\left(\boldsymbol{\beta}, \sigma^2 | \boldsymbol{y}\right) = \left(\frac{1}{\sigma^2 2\pi}\right)^{\frac{N}{2}} e^{-\frac{1}{2\sigma^2}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})^T(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})}$

Caution! Change of variable

To involve $\boldsymbol{\beta}$ and observable data in the likelihood, one has to switch from $\boldsymbol{\varepsilon}$ to $\mathbf{y}$. Easy? Yes and no. In single-equation models and under independent information, this is "just" a substitution like above. In general, however, one has to apply the change of variable theorem: $L\left(\theta | y\right) = L\left(\theta | \varepsilon\right) \cdot |\det \frac{\partial \boldsymbol{\varepsilon}}{\partial \mathbf{y}}|$. Note that here $\frac{\partial \boldsymbol{\varepsilon}}{\partial \mathbf{y}} = 1$.

This will no longer be the case with simultaneous equations models (Time Series Econometrics / Ekonometria Szeregów Czasowych) or models with interdependent observations (Spatial Econometrics / Ekonometria Przestrzenna).

Linear regression (3): first-order conditions

Log-likelihood:

$$\begin{aligned}\ln L\left(\boldsymbol{\beta}, \sigma^2|\mathbf{y}\right) &= -\frac{N}{2}\ln\left(\sigma^2 2\pi\right) - \frac{1}{2\sigma^2}\left(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}\right)^T\left(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}\right) = \\ &= -\frac{N}{2}\ln\left(\sigma^2 2\pi\right) - \frac{1}{2\sigma^2}\left(\mathbf{y}^T\mathbf{y} - 2\boldsymbol{\beta}^T\mathbf{X}^T\mathbf{y} + \boldsymbol{\beta}^T\mathbf{X}^T\mathbf{X}\boldsymbol{\beta}\right)\end{aligned}$$

First-order conditions:

By matrix calculus principles

$\frac{\partial(\mathbf{x}'\mathbf{A}\mathbf{x})}{\partial\mathbf{x}} = (\mathbf{A}+\mathbf{A}')\,\mathbf{x}$
If $\mathbf{A}$ symmetric, this further collapses to $2\mathbf{A}\mathbf{x}$.

1. $\frac{\partial \ln L}{\partial\boldsymbol{\beta}} = -\frac{1}{2\sigma^2}\left(-2\mathbf{X}^T\mathbf{y} + 2\mathbf{X}^T\mathbf{X}\boldsymbol{\beta}\right) = \frac{1}{\sigma^2}\left(\mathbf{X}^T\mathbf{y} - \mathbf{X}^T\mathbf{X}\boldsymbol{\beta}\right) = 0$
   $\hat{\boldsymbol{\beta}}^{ML} = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{y}$
2. $\frac{\partial \ln L}{\partial\sigma^2} = -\frac{N}{2\sigma^2} + \frac{1}{2}\left(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}\right)^T\left(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}\right)\frac{1}{\sigma^4} = 0$
   $\hat{\sigma}^{2,\,ML} = \frac{(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})^T(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})}{N}$

Linear regression (4): Hessian at maximum

Second derivatives:

$$\begin{bmatrix} \frac{\partial^2 \ln L}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^T} & \frac{\partial^2 \ln L}{\partial \boldsymbol{\beta} \partial \sigma^2} \\ \frac{\partial^2 \ln L}{\partial \sigma^2 \partial \boldsymbol{\beta}^T} & \frac{\partial^2 \ln L}{\partial \sigma^2 \partial \sigma^2} \end{bmatrix} =$$

$$\begin{bmatrix} -\frac{1}{\sigma^2}\mathbf{X}^T\mathbf{X} & -\frac{1}{\sigma^4}\left(\overbrace{\mathbf{X}^T\mathbf{y} - \mathbf{X}^T\mathbf{X}\boldsymbol{\beta}}^{=0 \text{ at } \hat{\beta}^{ML}}\right) \\ -\frac{1}{\sigma^4}\left(\overbrace{\mathbf{X}^T\mathbf{y} - \mathbf{X}^T\mathbf{X}\boldsymbol{\beta}}^{=0 \text{ at } \hat{\beta}^{ML}}\right) & \frac{N}{2\sigma^4} - (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})\frac{1}{\sigma^6} \end{bmatrix}$$

Denote the $\boldsymbol{\beta}$-related part as $\mathbf{H}$ ("Hessian") and note that:

$$-\mathbf{H}^{-1} = \sigma^2 \left(\mathbf{X}^T\mathbf{X}\right)^{-1}$$

coincides with the variance-covariance formula for $\hat{\beta}^{OLS}$.

Using Hessian as variance-covariance estimate (1)

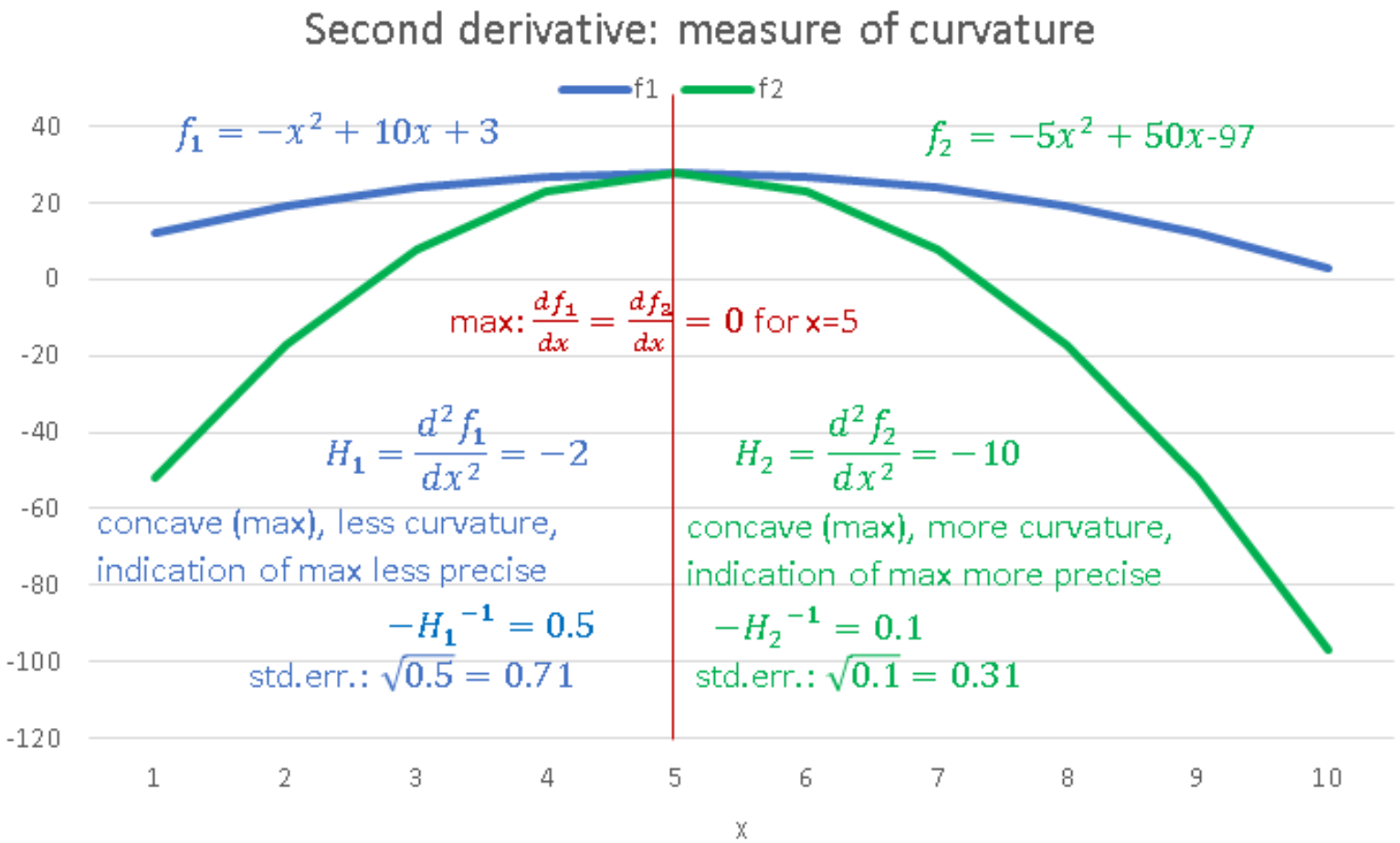
- This property goes beyond the linear regression model!

Variance-covariance of estimation under ML

The negative inverted Hessian matrix of the likelihood function, evaluated at its maximum, is generally used to approximate the variance-covariance matrix of estimation.

- One can read-off the square roots of the diagonal elements as standard errors of ML estimation.
- If impossible to derive analytically, the Hessian can also be approximated numerically.

Using Hessian as variance-covariance estimate (2)



Linear regression with ML: exercise

Exercise 3. Use the dataset on students' satisfaction (**Lecture 1**).

1. Recall the OLS estimates with student satisfaction as dependent variable; grade and sex as regressors.
2. Define the log-likelihood function for the same regression problem and use `optim` to find the ML estimates. Compare to OLS.
3. Use Hessian at maximum to approximate the standard errors of estimation. Compare to OLS. Where does the difference come from? (Note that $\hat{\sigma}^{2,OLS} \neq \hat{\sigma}^{2,ML}$.)

Logit regression model (1)

$$y_i = \begin{cases} 1 & \text{with prob. } p_i = \frac{e^{\mathbf{x_i}\beta}}{1+e^{\mathbf{x_i}\beta}} \\ 0 & \text{with prob. } 1-p_i = \frac{1}{1+e^{\mathbf{x_i}\beta}} \end{cases}$$

Observing N instances of y_i $(i = 1, 2, ..., N)$ yields the following sample probability / likelihood:

$$\begin{aligned} L_{logit}(\beta|\mathbf{y},\mathbf{X}) &= \prod_{i=1}^{N} [y_i \cdot p_i + (1-y_i)\cdot(1-p_i)] = \\ &= \prod_{i=1}^{N} \left[y_i \cdot \frac{e^{\mathbf{x_i}\beta}}{1+e^{\mathbf{x_i}\beta}} + (1-y_i)\cdot \frac{1}{1+e^{\mathbf{x_i}\beta}} \right] \end{aligned}$$

$$\begin{aligned} \ln L_{logit}(\beta|\mathbf{y},\mathbf{X}) &= \sum_{i=1}^{N} \ln [y_i \cdot p_i + (1-y_i)\cdot(1-p_i)] = \\ &= \sum_{i=1}^{N} \left[y_i \cdot \ln \frac{e^{\mathbf{x_i}\beta}}{1+e^{\mathbf{x_i}\beta}} + (1-y_i)\cdot \ln \frac{1}{1+e^{\mathbf{x_i}\beta}} \right] \end{aligned}$$

Logit regression model (2)

$$\hat{\beta}_{logit}^{ML} = \arg\max_{\beta} \ln L_{logit}(\beta|\mathbf{y}, \mathbf{X})$$

Exercise 4. Use the **Titanic** dataset to estimate the logit model with passenger's *Survival* as dependent variable, and the following variables as regressors: *Pclass*, *Sex*, *SibSp*, *embark_q*, *embark_p* (explanations in the code).
1) estimate the logit model with the `glm(..., family = binomial(link = "logit"))` command;
2) then replicate this result by explicitly maximizing the above log-lik function.

C-D function with additive errors

$$y_i = \beta_o L_i^{\beta_1} K_i^{\beta_2} \varepsilon_i \;\rightarrow\; \ln y_i = \ln \beta_o + \beta_1 \ln L_i + \beta_2 \ln K_i + \ln \varepsilon_i$$

- Transformed model can be estimated with OLS. Zero-mean $\ln \varepsilon_i$ assumed.
- But multiplicative errors ε_i imply that the noise is proportional to the predicted level, $\beta_o L_i^{\beta_1} K_i^{\beta_2}$. That may, or may not, fit well with the data.

ALTERNATIVE:

$y_i = \beta_o L_i^{\beta_1} K_i^{\beta_2} + \varepsilon_i \;\rightarrow\; ?$

Assume $\varepsilon_i \sim N\left(0, \sigma^2\right)$ i.i.d.

Exercise 5. Derive and numerically maximize the likelihood function. Use **this** dataset.

Outline

MLE properties

- consistent: $p\lim \hat{\boldsymbol{\theta}}^{ML} = \boldsymbol{\theta}$;
- asymptotically normal: $\hat{\boldsymbol{\theta}}^{ML} \xrightarrow{a} N\left(\boldsymbol{\theta}, I\left(\boldsymbol{\theta}\right)^{-1}\right)$
  - Important result fo deriving the distributions of test statistics (soon to follow).
  - Fisher's information matrix: $I\left(\boldsymbol{\theta}\right) = -E\left(\frac{\partial^2 \ln L}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T}\right)$.
  - Note the expected value operator and the dependence on unknown true $\boldsymbol{\theta}$. If numerical evaluation is necessary, approximated with $-\mathbf{H}$ as demonstrated before.
- asymptotically efficient;
- invariant, i.e. the ML estimate of $g\left(\theta\right)$ is $g\left(\hat{\theta}^{ML}\right)$ if $g\left(.\right)$ is continuous and continuously differentiable

Sketch of proof: ***Greene*** (chapter "Maximum Likelihood Estimation").

Note!

$\mathbf{H}$, $\mathbf{H}\left(\hat{\theta}\right)$	Hessian	matrix of log-likelihood's second derivatives, evaluated at maximum
$\mathbf{I}$, $\mathbf{I}(\theta)$	Fisher's information matrix	minus *expected value* of second derivatives matrix, evaluated at *true parameter values*; *in practice*, approximated as $I(\theta) \simeq -H\left(\hat{\theta}\right)$; measure of log-likelihood's curvature at maximum
$\mathbf{V} = \begin{bmatrix} v_{1,1} & v_{1,2} & & \\ \cdot & v_{2,2} & & \\ & & \ddots & \\ & & & \cdot \end{bmatrix}$	variance-covariance matrix of ML estimator	$\mathbf{V} = \mathbf{I}(\theta)^{-1} \simeq -\mathbf{H}\left(\hat{\theta}\right)^{-1}$
std.err. of estimation		$SE\left(\widehat{\theta_1}\right) = \sqrt{v_{1,1}}$, $SE\left(\widehat{\theta_2}\right) = \sqrt{v_{2,2}}$, ...

Formulating the null hypothesis (1)

$H_0 : c(\boldsymbol{\theta}) = \mathbf{q}$

This is a set of m ($\leq$ length of $\boldsymbol{\theta}$) linear or nonlinear equations.
For example:

- $\theta_1 = 0; c(\boldsymbol{\theta}) = [\theta_1], \mathbf{q} = 0$ – **most frequent case**
- $\theta_1 = 0 \wedge \theta_2 = 0;\ c(\boldsymbol{\theta}) = \begin{bmatrix} \theta_1 \\ \theta_2 \end{bmatrix}, \mathbf{q} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$
- $\theta_1 + \theta_2 = 1 \wedge \theta_3 = \theta_4;\ c(\boldsymbol{\theta}) = \begin{bmatrix} \theta_1 + \theta_2 \\ \theta_3 - \theta_4 \end{bmatrix}, \mathbf{q} = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$
- $\frac{\theta_1}{\theta_1+\theta_2} = 0.5;\ c(\boldsymbol{\theta}) = \left[\frac{\theta_1}{\theta_1+\theta_2}\right], \mathbf{q} = 0.5$

Formulating the null hypothesis (2)

Let $\boldsymbol{\theta_0}$ be a vector of restricted estimates/parameters, according to H_0. For example:

- if $H_0 : \theta_1 = 0$ and $\boldsymbol{\theta} = [\theta_1]$, then $\boldsymbol{\theta_0} = [0]$
- if $H_0 : \theta_1 = 0$ and $\boldsymbol{\theta} = \begin{bmatrix} \theta_1 \\ \theta_2 \\ \vdots \end{bmatrix}$, then $\boldsymbol{\theta_0} = \begin{bmatrix} 0 \\ \hat{\theta}_{0,2} \\ \vdots \end{bmatrix}$

  – most frequent case
- if $H_0 : \theta_1 + \theta_2 = 1 \wedge \theta_3 = \theta_4$ and $\boldsymbol{\theta} = \begin{bmatrix} \theta_1 \\ \theta_2 \\ \theta_3 \\ \theta_4 \end{bmatrix}$, then $\boldsymbol{\theta_0} = \begin{bmatrix} \hat{\theta}_{0,1} \\ 1 - \hat{\theta}_{0,1} \\ \hat{\theta}_{0,3} \\ \hat{\theta}_{0,3} \end{bmatrix}$

Three approaches to testing under ML

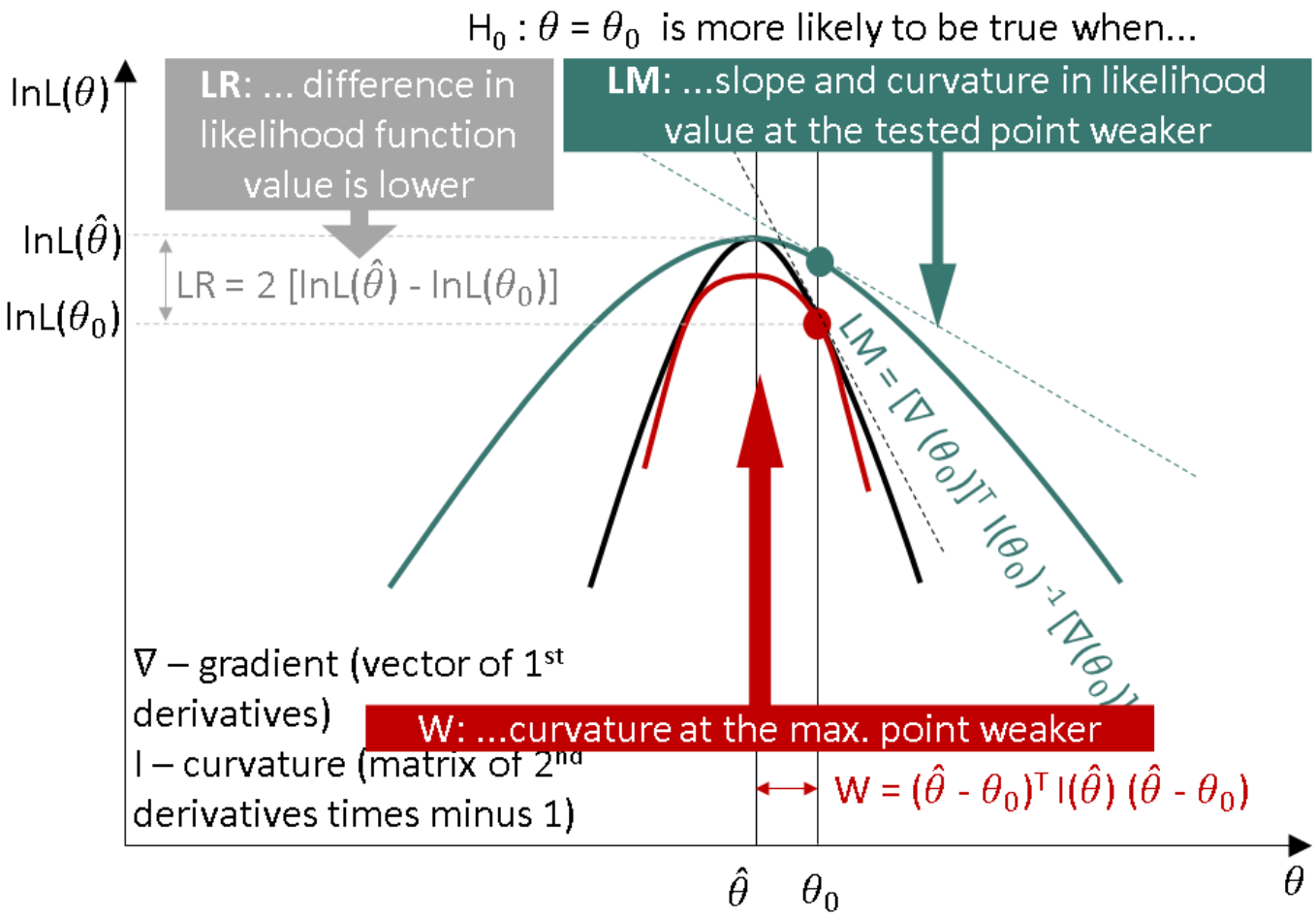
1. **Wald** test: $W = (c(\boldsymbol{\theta}) - \mathbf{q})^T \cdot I\left(\hat{\boldsymbol{\theta}}\right) \cdot (c(\boldsymbol{\theta}) - \mathbf{q})$,
   only the unrestricted model must be estimated
2. **Lagrange Multiplier** test:
   $LM = \frac{\partial \ln L}{\partial \boldsymbol{\theta}}(\boldsymbol{\theta_0})^T \cdot I(\boldsymbol{\theta_0})^{-1} \cdot \frac{\partial \ln L}{\partial \boldsymbol{\theta}}(\boldsymbol{\theta_0})$,
   only the restricted model must be estimated
3. **Likelihood Ratio** test: $LR = 2\left[\ln L\left(\hat{\boldsymbol{\theta}}\right) - \ln L(\boldsymbol{\theta_0})\right]$,
   both models must be estimated

These procedures are asymptotically:

- equivalent
- chi-squared distributed (df = number of restrictions).

A simple illustration of a single-parameter problem follows.

Single-parameter illustration (θ – scalar)



Testing hypotheses under ML: exercise

Exercise 6. In the logit model from Exercise 4, test the hypothesis that two variables should be dropped – those indicating the port where the passengers embarked.

1. Formulate the restrictions as a null hypothesis equation.
2. Install the package `{numDeriv}`. Get familiar with the functions `grad` and `hessian`.
3. Use the Wald test, referring to the existing unrestricted estimates.
4. Use the LM test, after obtaining the restricted estimates.
3. Use the LR test, referring to both restricted and unrestricted estimates.