Maximum Likelihood Methods

2025/2026, Semester 1

Ralf A. Wilke
Copenhagen Business School

ML Methods

Main Texts:

Wooldridge, 2010,Chapter 13, parts of chapter 12 (12.6 hypothesis testing).

- ML estimation and inference should be known from other courses.

- We focus on derivation of ML objective on the grounds of the Kullback-Leibler (K-L) information inequality.
  - In particular pages 9-12.
  - Does not require independent observations. Important for many economic applications (e.g. panel data).

- Maximum likelihood estimation (MLE) makes full distributional assumptions about the dependent variable(s) conditional on the exogenous variables.
  - Generally the most efficient estimation procedure in this class of estimators.
  - OLS or related non linear least squares (NLS) in contrast does not require full distributional assumptions.
- Practical trade-off between efficiency and robustness.
  - Failure of distributional assumptions generally results in inconsistency of the estimator.

- There should be a practical gain from using MLE and not OLS/NLS.
  - If MLE and LS method have similar efficiency, little reason to apply MLE.

- Distributional assumptions should be ideally motivated by some economic theory to avoid arbitrariness.
  - Difficult in practice.

- Only apply MLE if …
  - there is a practical gain from using MLE and not LS method or
  - you are sure about distribution or
  - MLE is to some degree robust with respect to the distributional assumption.

Conditional Maximum Likelihood Estimation

- Independent identically distributed observations $\{y_i, \mathbf{x}_i : i = 1, 2, ..., N\}$ with *y* is a real number and ***x*** is K-dimensional.
  - Can be easily extended to multiple ***y*** (system of equations).
  - Focus on cross section data for simplicity.
- We want to specify the conditional density of *y* given ***x***.
  - Regression setting!
- We assume that this density is known up to a finite number of parameters.
  - Parametric model.

- $p_0(y|\mathbf{x})$ is the conditional density of y_i given $\mathbf{x}_i = \mathbf{x}$.
  - "True" density.
  - $\mathbf{x} \in \mathcal{X}$ and $y \in \mathcal{Y}$ (supports/all possible values of ***x*** and *y*, respectively).
- *y* can be continuous or discrete or have both continuous and discrete characteristics.
  - We introduce notation that is compatible with both.
  - We define $v(dy)$ as a σ- finite measure:
    - When y is discrete, $v(dy)$ turns integrals into sums.
    - When y is continuous, $v(dy)$ the integral is the usual integral (Riemann).

$$\int_{\mathcal{Y}} p_0(y|\mathbf{x}) v(dy) = 1$$

- OLS or nonlinear least squares (NLS) minimise the sum of squared residuals. In terms of the population model this is:

$$min_{\beta} E\{[y - m(\mathbf{x}, \beta)]^2\}$$

  - For linear models, OLS has closed form solution.
  - For non-linear models NLS typically does not have a closed form solution.
- What is an analogous motivation for conditional maximum likelihood?
  - Use conditional Kullback-Leibler (K-L) information inequality.

- As before $p_0(y|\mathbf{x})$ is the conditional density.
- $f(y|\mathbf{x})$ is another conditional density: $f \geq 0$ for all ***x*** and *y* and

$$\int_{\mathcal{Y}} f(y|\mathbf{x}) v(dy) = 1$$

- The conditional K-L information inequality is given by

$$\mathcal{K}(f;\mathbf{x}) = \int_{\mathcal{Y}} log[p_0(y|\mathbf{x})/f(y|\mathbf{x})] p_0(y|\mathbf{x}) v(dy) \geq 0$$

for all ***x***.

  - The inequality is only zero if p_0 and f are the same (almost) everywhere.
  - The greater $\mathcal{K}(f;\mathbf{x})$ the more dissimilar are the two densities.
- Another way to express the K-L information inequality is:

$$E\{log[p_0(y|\mathbf{x})]\} \geq E\{log[f(y|\mathbf{x})]\}$$

- MLE utilises the observation that $\mathcal{K}(f;\mathbf{x})$ is minimised at $f = p_0$.
- f is to be specified to serve as a model for p_0.
- We consider parametric models:

$$\{f(y|\mathbf{x};\theta), \theta \in \Theta\}$$

with $\Theta \subset \mathcal{R}^P$ is *Px1*.

  - We assume that we know the functional form of f. The goal is to determine θ (which is a vector and should be denoted as $\boldsymbol{\theta}$).
  - A good model for f is such that it is a density for all values of *x* and θ; and for some $\theta_0 \in \Theta$: $f(y|\mathbf{x};\theta_0) = p_0(y|\mathbf{x})$

- Using the observation that $\mathcal{K}(f;\mathbf{x})$ can be written

$$E\{log[f(y_i|\mathbf{x}_i;\theta_0)]\} \geq E\{log[f(y_i|\mathbf{x}_i;\theta)]\}$$

or

$$E\{l_i(\theta_0|\mathbf{x}_i)\} \geq E\{l_i(\theta|\mathbf{x}_i)\}$$

with

$$\ell_i(\theta) = \ell(y_i, \mathbf{x}_i, \theta) = log f(y_i|\mathbf{x}_i;\theta)$$

- $\ell_i(\theta) = log f(y_i|\mathbf{x}_i;\theta)$ is called the conditional log-likelihood of observation *i*.
  - This is a random function because *y* and ***x*** are random.
- $E\{l_i(\theta_0|\mathbf{x}_i)\} \geq E\{l_i(\theta|\mathbf{x}_i)\}$ holds for all values of θ and ***x*** but the expectation is conditional on ***x***.
  - In our sample we can normally hardly condition on ***x***.
- But it is also true that

$$E_{\mathbf{x}}E\{l_i(\theta_0|\mathbf{x}_i)\} \geq E_{\mathbf{x}}E\{l_i(\theta|\mathbf{x}_i)\}$$

  and by using the law of iterated expectations we have

$$E\{l_i(\theta_0)\} \geq E\{l_i(\theta)\}.$$

  - The expectation is over the joint distribution of (*y*,***x***).
- Thus, θ_0 is the solution to $max_{\theta\in\Theta}E\{\ell_i(\theta)\}$.
  - Therefore maximum likelihood!

- In terms of the sample the analogue of $max_{\theta \in \Theta} E\{\ell_i(\theta)\}$ is $max_{\theta \in \Theta} N^{-1} \sum_{i=1}^{N} log f(y_i|\mathbf{x}_i;\theta) = max_{\theta \in \Theta} N^{-1} \mathcal{L}(\theta).$
  - This is the well-known sample log-likelihood to be maximised.
  - The solution $\hat{\theta}$ is called the (conditional) maximum likelihood estimator.
  - Also consistent when observations are dependent.
- Another way to write down the objective is

$$max_{\theta \in \Theta} \prod_{i=1}^{N} f(y_i|\mathbf{x}_i;\theta).$$

  - This is a model of the joint density $(y_1, \cdots, y_N)$ given fixed ***x*** and <u>additionally</u> assuming that observations are independent.
  - This problem then produces the same $\hat{\theta}$.
  - But it is more difficult to understand why this actually leads to a good estimate.

- Why using the K-L information inequality representation instead of assuming independent observations:
  - Independence assumptions are unrealistic in time-series and panel data applications. Too difficult to write the joint distribution under dependence (just special cases). Use partial likelihood.
  - The maximisation of the likelihood function becomes numerically challenging in more advanced statistical models with unobserved effects and is therefore too slow. Simplify the likelihood by using K-L information inequality arguments to speed up maximisation: Variational approximation (Christoffersen, 2021): partial likelihood.

- Example: Multiple regression with normal error.

$$y_i = \mathbf{x}_i\beta + u_i, \quad u_i \sim N(0, \sigma^2)$$

  - Then $y_i|\mathbf{x}_i \sim N(\mathbf{x}_i\beta, \sigma^2)$ and $\theta = (\beta, \sigma)$.
  - Thus

$$f(y_i|\mathbf{x}_i;\theta) = (2\pi\sigma^2)^{-1/2} exp[(-1/2)\left(\frac{y_i - \mathbf{x}_i\beta}{\sigma}\right)^2]$$

    and

$$N^{-1}\sum_i \ell_i(\theta) = -\tfrac{1}{2}log(2\pi) - \frac{log\sigma^2}{2} - \frac{1}{2\sigma^2 N}\sum_i (y_i - \mathbf{x}_i\beta)^2$$

  - Take $N^{-1}\nabla_\theta \sum_i \ell_i(\theta) = N^{-1}\frac{\partial \sum_i \ell_i(\theta)}{\partial\theta} = \mathbf{0}$
  - It turns out that the ML estimator for β is the OLS estimator!
  - The ML estimator for σ^2 is the sum of squared residuals divided by *N*.

Consistency of CML Estimation

- Reminder:

Definition: Consistency

Let $\hat{\beta}_j$ be an estimator of β_j based on a sample $Z_1, Z_2, \ldots, Z_n$ of size *n*. Then, $\hat{\beta}_j$ is a consistent estimator of β_j if for every $\epsilon > 0$,

$$P(|\hat{\beta}_j - \beta_j| > \epsilon) \to 0 \text{ as } n \to \infty.$$

$$E(\hat{\beta}_j) = \beta_j \text{ AND } \text{var}(\hat{\beta}_j) \to 0 \text{ as } n \to \infty \Rightarrow \text{plim}\hat{\beta}_j = \beta_j$$

- For the ML estimator $\hat{\theta}$ we only consider the required assumptions and resulting theorem.

Theorem 13.1 W2010 (Consistency of CMLE)

Under the Assumptions below, the ML estimator $\hat{\boldsymbol{\theta}}$ is the unique solution to

$$\max_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} N^{-1}\sum_{i=1}^{N} \log f(y_i|\mathbf{x}_i;\boldsymbol{\theta})$$

and is consistent for $\boldsymbol{\theta}$, i.e. $\operatorname{plim} \hat{\boldsymbol{\theta}} = \boldsymbol{\theta}$.

Using the notation from the previous slides:

a) $f(y_i|\mathbf{x}_i;\boldsymbol{\theta})$ is a density with respect to the measure $v(dy)$ (integrates/sums up to 1).

b) For some $\boldsymbol{\theta_0}\in\Theta$, $p_0(y|\mathbf{x}) = f(y|\mathbf{x};\boldsymbol{\theta_0})$ for all $\boldsymbol{x}$ and $\boldsymbol{\theta_0}$ is the unique solution to $\max_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} E\{\ell_i(\boldsymbol{\theta})\}$.

c) Θ is a compact set.

d) For each $\boldsymbol{\theta}$, $\ell_i(\boldsymbol{\theta})$ is a Borel measurable function.

e) For each (y,$\boldsymbol{x}$), $\ell_i(\boldsymbol{\theta})$ is a continuous function on Θ.

f) For all $\boldsymbol{\theta}$, $|\ell_i(\boldsymbol{\theta})|$ is bounded by something that is $< \infty$ in expectation.

Asymptotic distribution of the ML estimator

- More general theory for extremum or M-estimators suggests that the ML estimator has an asymptotic normal distribution.
- The first and second derivatives of the objective function determine the features of the normal distribution.
  - The first derivatives have to be zero in expectation to ensure unbiasedness.
  - The second derivatives determine the variance matrix.
- Thus, before we consider the distribution we focus on some properties of the derivatives.

- We assume that
  - $\boldsymbol{\theta_0}$ is in the interior of Θ.
  - $\ell_i(\boldsymbol{\theta})$ is twice continuously differentiable in $\boldsymbol{\theta}$ (derivatives are also continuous).
- Reminder: $\boldsymbol{\theta}$ is *Px1*.
  - This typically includes the *K* parameters on the regressors plus some nuisance parameters that describe features of the conditional distribution of *y given* ***x*** (e.g. the variance).
- We say that the *Px1* vector of partial derivatives of $\ell_i(\boldsymbol{\theta})$ is the **score** of the log likelihood:

$$\boldsymbol{s}_i(\boldsymbol{\theta}) = \nabla_\theta \ell_i(\boldsymbol{\theta})' = \left(\frac{\partial \ell_i(\theta)}{\partial \theta_1}, \ldots, \frac{\partial \ell_i(\theta)}{\partial \theta_P} \right)'$$

  - In order to maximise the sample log-likelihood, the averages (over *i*) of the first derivatives are set equal (or very close) to *0*.

- Most econometric models imply the following property:

$$E[\boldsymbol{s}_i(\boldsymbol{\theta}_0)|\mathbf{x}_i] = \mathbf{0}$$

  - The expectation is with respect to $f(y_i|\mathbf{x}_i;\theta)$, i.e. conditional on ***x***, and evaluated at $\boldsymbol{\theta_0}$.
  - This then implies $E[\boldsymbol{s}_i(\boldsymbol{\theta}_0)] = \mathbf{0}$.
- This means our sample log-likelihood converges asymptotically to a function that has a *0* gradient at the true parameter values (necessary condition).

- We denote

$$\boldsymbol{B}_0 = \text{Var}[\boldsymbol{s}_i(\boldsymbol{\theta}_0)] = E[\boldsymbol{s}_i(\boldsymbol{\theta}_0)\boldsymbol{s}_i(\boldsymbol{\theta}_0)']$$

  - Is the expected variance matrix of the score evaluated at $\boldsymbol{\theta_0}$.
  - $\boldsymbol{B}_0$ is called the information matrix.

- The *PxP* matrix of the second partial derivatives of $\ell_i(\boldsymbol{\theta})$ is called the **Hessian**:

$$\boldsymbol{H}_i(\boldsymbol{\theta}) = \nabla_\theta \boldsymbol{s}_i(\boldsymbol{\theta}) = \nabla^2_\theta \ell_i(\boldsymbol{\theta})$$

  - Symmetric
  - The expected value of $\boldsymbol{H}_i(\boldsymbol{\theta})$ is negative definite (sufficient condition, not shown here).
- We define: $\boldsymbol{A}_0 = -E[\boldsymbol{H}_i(\boldsymbol{\theta}_0)]$
  - Positive definite (if the expected log likelihood has a maximum at $\boldsymbol{\theta_0}$ = $\boldsymbol{\theta_0}$ is identified).

- One nice feature of the ML estimator is that $\boldsymbol{A}_0 = \boldsymbol{B}_0$.
  - So called Information matrix equality.
  - This implies that the ML estimator has the minimum variance property (its variance attains the Cramer-Rao bound; there is no other estimator in this class with a smaller variance).

Theorem 13.2 W2010 (Asymptotic Normality of CMLE)

Under the Assumptions of Theorem 13.1 and the assumptions below, the ML estimator $\hat{\boldsymbol{\theta}}$ has the following properties:

$$\sqrt{N}(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0) \overset{a}{\sim} N(\mathbf{0},\boldsymbol{A}_0^{-1})$$

And therefore the asymptotic variance of $\hat{\boldsymbol{\theta}}$ is $\boldsymbol{A}_0^{-1}/N$.

Using the notation from the previous slides:

a) $\boldsymbol{\theta_0}$ is in the interior of Θ.

b) $\ell(y, \mathbf{x};\boldsymbol{\theta})$ is twice differentiable in $\boldsymbol{\theta}$ in the interior of Θ for all values of *y* and ***x***.

c) Integration with respect to *y* and differentiation with respect to $\boldsymbol{\theta}$ of $f(y|\mathbf{x}_i;\boldsymbol{\theta})$ and $\mathbf{s}_i(\boldsymbol{\theta})f(y|\mathbf{x}_i;\boldsymbol{\theta})$ are interchangeable for all $\boldsymbol{\theta}$ in the interior of Θ.

d) $E[|\nabla_\theta^2 \ell_i(\boldsymbol{\theta})|] < \infty$

e) $\boldsymbol{A}_0$ is positive definite.

- Estimating the asymptotic variance of $\hat{\theta}$requires estimating $\boldsymbol{A}_0$.
- There are several possible estimators.
- Most commonly $-E[\boldsymbol{H}_i(\boldsymbol{\theta}_0)]$ is estimated using sample analogues:

$$N^{-1}\sum_{i=1}^{N} -\boldsymbol{H}_i(\hat{\boldsymbol{\theta}}) \quad \text{or} \quad N^{-1}\sum_{i=1}^{N} \boldsymbol{s}_i(\hat{\boldsymbol{\theta}})\boldsymbol{s}_i(\hat{\boldsymbol{\theta}})'$$

  - The first requires the Hessian, while the second only the score.
  - While simpler, the second has worse finite sample properties.
  - There are other estimators with even better properties (compare W2010, p.480).
  - Asymptotic standard errors are square roots of the diagonal elements of e.g. $\left[-N^{-1}\sum_{i=1}^{N}\boldsymbol{H}_i(\hat{\boldsymbol{\theta}})\right]^{-1}$

- Example: Multiple regression with normal error (cont.).

$$y_i = \mathbf{x}_i\beta + u_i, \quad u_i \sim N(0, \sigma^2)$$

  - Then $y_i|\mathbf{x}_i \sim N(\mathbf{x}_i\beta, \sigma^2)$ and $\theta = (\beta, \sigma)$, *P=K+1*.
  - Thus

$$f(y_i|\mathbf{x}_i;\theta) = (2\pi\sigma^2)^{-1/2} exp[(-1/2)\left(\frac{y_i - \mathbf{x}_i\beta}{\sigma}\right)^2]$$

$$N^{-1}\sum_i \ell_i(\theta) = -\tfrac{1}{2}log(2\pi) - \frac{log\sigma^2}{2} - \frac{1}{2\sigma^2 N}\sum_i (y_i - \mathbf{x}_i\beta)^2$$

  - The score is:

$$\boldsymbol{s}_i(\boldsymbol{\theta}) = \nabla_\theta \ell_i(\boldsymbol{\theta})' = \left(\frac{\partial \ell_i(\theta)}{\partial \theta_1}, ..., \frac{\partial \ell_i(\theta)}{\partial \theta_P}\right)'$$

    which is:

$$\nabla_\beta \ell_i(\boldsymbol{\theta})' = (\tfrac{1}{2\sigma^2}\nabla_\beta (y_i - \mathbf{x}_i\beta)^2)' = -(\sigma^2)^{-1}(-\mathbf{x}_i' y_i + \mathbf{x}_i'\mathbf{x}_i\beta)$$
$$= (\sigma^2)^{-1}(\mathbf{x}_i'(y_i - \mathbf{x}_i\beta))$$

$$\nabla_{\sigma^2} \ell_i(\boldsymbol{\theta}) = -\tfrac{1}{2}(\sigma^2)^{-1} + \tfrac{1}{2}(\sigma^2)^{-2}(y_i - \mathbf{x}_i\beta)^2$$

- Why does $E[\boldsymbol{s}_i(\boldsymbol{\theta}_0)|\mathbf{x}_i] = 0$ hold?
- Because: $E[\nabla_\beta \ell_i(\boldsymbol{\theta}_0)'|\mathbf{x}_i] = E[(\sigma_0^2)^{-1}(\mathbf{x}_i' u_i)|\mathbf{x}_i] = 0$

$$E[\nabla_{\sigma^2} \ell_i(\boldsymbol{\theta}_0)|\mathbf{x}_i] = E[-\tfrac{1}{2}(\sigma_0^2)^{-1} + \tfrac{1}{2}(\sigma_0^2)^{-2}(u_i)^2|\mathbf{x}_i]$$
$$= -\tfrac{1}{2}(\sigma_0^2)^{-1} + \tfrac{1}{2}(\sigma_0^2)^{-2}\sigma_0^2 = 0$$

  with $E(u^2|\mathbf{x}) = \sigma^2$.
- The Hessian is: $\boldsymbol{H}_i(\boldsymbol{\theta}) = \nabla_\theta \boldsymbol{s}_i(\boldsymbol{\theta}) = \nabla_\theta^2 \ell_i(\boldsymbol{\theta})$

  with components:

$$\boldsymbol{H}_i(\boldsymbol{\theta}) = \begin{pmatrix} \nabla_\beta^2 \ell_i(\theta)' & \nabla_{\sigma^2}\nabla_\beta \ell_i(\theta)' \\ \nabla_\beta \nabla_{\sigma^2} \ell_i(\theta)' & \nabla_{\sigma^2}^2 \ell_i(\theta)' \end{pmatrix}$$

- More specifically:

$$\nabla_{\beta}^{2}\ell_i(\boldsymbol{\theta})' = \nabla_{\beta}\left((\sigma^2)^{-1}(\mathbf{x}_i'(y_i - \mathbf{x}_i\beta))\right) = -(\sigma^2)^{-1}\mathbf{x}_i'\mathbf{x}_i$$

$$\nabla_{\sigma^2}\nabla_{\beta}\ell_i(\boldsymbol{\theta})' = \nabla_{\sigma^2}\left((\sigma^2)^{-1}(\mathbf{x}_i'(y_i - \mathbf{x}_i\beta))\right) = -(\sigma^2)^{-2}(\mathbf{x}_i'(y_i - \mathbf{x}_i\beta))$$

$$\begin{aligned}\nabla_{\beta}\nabla_{\sigma^2}\ell_i(\boldsymbol{\theta})' &= \nabla_{\beta}\left(-\tfrac{1}{2}(\sigma^2)^{-1} + \tfrac{1}{2}(\sigma^2)^{-2}(y_i - \mathbf{x}_i\beta)^2\right) \\ &= -(\sigma^2)^{-2}(\mathbf{x}_i(y_i - \mathbf{x}_i\beta)\end{aligned}$$

$$\begin{aligned}\nabla_{\sigma^2}^{2}\ell_i(\boldsymbol{\theta})' &= \nabla_{\sigma^2}\left(-\tfrac{1}{2}(\sigma^2)^{-1} + \tfrac{1}{2}(\sigma^2)^{-2}(y_i - \mathbf{x}_i\beta)^2\right) \\ &= \tfrac{1}{2}(\sigma^2)^{-2} - (\sigma^2)^{-3}(y_i - \mathbf{x}_i\beta)^2\end{aligned}$$

- To find $\boldsymbol{A}_0 = -E[\boldsymbol{H}_i(\boldsymbol{\theta}_0)]$ we use again $E(u^2|\mathbf{x}) = \sigma^2$

$$-E[\boldsymbol{H}_i(\boldsymbol{\theta}_0)] = \begin{pmatrix} E\left[\mathbf{x}_i'\mathbf{x}_i\right]/\sigma_0^2 & 0 \\ 0 & \frac{1}{2\sigma_0^4} \end{pmatrix}$$

- Since $\boldsymbol{A}_0 = \boldsymbol{B}_0$ it is the (Fischer) Information Matrix.
- $\boldsymbol{A}_0$ is block diagonal.

Inference Methods

- Testing an unrestricted model versus a model under Q restrictions.
- There are three tests that are commonly applied:
  1) LR test
  2) LM/score test
  3) Wald test

- All are asymptotically equivalent but differ in finite sample performance and practicability.

- Wald test
  - Allows for non-linear restrictions to be tested.
  - Based on the unrestricted model.
  - Default test in packages

  - Suppose there are *Q* (non-)linear continuously differentiable restrictions: $$H_0 : \mathbf{c}(\boldsymbol{\theta}_0) = \mathbf{0}$$
    - Well known special case: $H_0 : \theta_{j0} - c = 0$ (one linear restriction).
    - *c* maps from dimension *P* to dimension *Q* .
    - $\nabla_\theta \mathbf{c}(\boldsymbol{\theta}_0) = \mathbf{C}(\boldsymbol{\theta}_0)$ is *QxP* and has rank *Q* (full row rank).
      - In the case of linear restrictions this means that the restrictions are linearly independent.

- The Wald statistic is

$$W = \mathbf{c}(\hat{\boldsymbol{\theta}})'(\hat{\mathbf{C}}\hat{\mathbf{V}}\hat{\mathbf{C}}')^{-1}\mathbf{c}(\hat{\boldsymbol{\theta}})$$

with $\hat{V}$ an appropriate estimate of $\mathbf{A}_0$.

- In the special case of $H_0 : \theta_{j0} - c = 0$, the numerator reduces to $(\hat{\theta}_j - c)^2$.
- The "numerator becomes larger the further away are the estimates from the hypotised values.
- This is "weighted" by the reliability of the estimates and by the marginal increase in the numerator. This is the denominator.
  - In the case of a single linear restriction ***C***=*(0,…0,1,0,…,0)*, (*1xP*).
- One can show that $W \overset{a}{\sim} \chi^2_Q$ if H_0 is true.
  - This follows from that fact that the "numerator" is a quadratic form (product of asymptotically normally distributed r.v.).
- The "denominator" is invertible due to the rank condition on ***C***.

- Issues:
  - The statistic is not invariant to how the nonlinear restrictions are imposed.
  - This means *W* changes its values depending on how the nonlinear hypothesis is stated (for example doing a nonlinear log-transformation).
  - θ_0 needs to be in the interior of Θ.

- Likelihood Ratio (LR) test
  - Testing for a change in the objective function.
  - The hypotheses are about Q restrictions: $H_0 : \theta_0 = c$

- The idea is that the ML estimator $\hat{\boldsymbol{\theta}}$ is the solution to
$$max_{\theta \in \Theta} \mathcal{L}(\theta).$$
- Thus, the sample log-likelihood for any other value of $\boldsymbol{\theta}$ cannot be larger: $\mathcal{L}(\hat{\theta}) \geq \mathcal{L}(\theta)$
- Let $\tilde{\theta}$ be the ML estimator obtained under the restrictions of H_0.
  - If $H_0 : \theta_{0j} = 0$ (*Q=1*), $\tilde{\theta}$ is the ML estimator of the restricted model with *P-1* parameters.
- The LR test involves estimation of the unrestricted and the restricted model.
- If H_0 is true, we should expect $\mathcal{L}(\hat{\theta}) \approx \mathcal{L}(\tilde{\theta})$.
- If H_0 is not true, we should expect $\mathcal{L}(\tilde{\theta})$ to be smaller.
- The *LR* statistic is based on this observation:
$$LR = 2[\mathcal{L}(\hat{\theta}) - \mathcal{L}(\tilde{\theta})]$$
- It can be shown that $LR \overset{a}{\sim} \chi^2_Q$ if H_0 is true.
  - Reject H_0 if *LR* is too big.

- Issues with the LR test:
  - Might be cumbersome to estimate the unrestricted model.
  - $\boldsymbol{\theta_0}$ needs to be on the interior of Θ.

- Lagrange Multiplier (LM) or Score test
  - This test is based on the restricted estimates only.
  - The idea of the test is based on the observation that the ML estimates $\hat{\boldsymbol{\theta}}$ are such that $\sum_{i=1}^{N} \boldsymbol{s}_i(\hat{\boldsymbol{\theta}}) = 0$.
  - Suppose there are *Q* (non-)linear continuously differentiable restrictions: $H_0 : \mathbf{c}(\boldsymbol{\theta}_0) = \mathbf{0}$
  - The idea is to estimate the model by ML under the restrictions of H_0 and plug these estimates and the restrictions into the unrestricted model to obtain $\tilde{\theta}$ and $\sum_{i=1}^{N} \boldsymbol{s}_i(\tilde{\boldsymbol{\theta}})$.
  - If H_0 is true $\sum_{i=1}^{N} \boldsymbol{s}_i(\tilde{\boldsymbol{\theta}}) \approx 0$.

- The LM statistic is based on $\sum_{i=1}^{N} \boldsymbol{s}_i(\tilde{\boldsymbol{\theta}})$ and has the limiting distribution of $N^{-1/2} \sum_{i=1}^{N} \boldsymbol{s}_i(\tilde{\boldsymbol{\theta}})$ under H_0:

$$LM = \left(\sum_{i=1}^{N} \boldsymbol{s}_i(\tilde{\boldsymbol{\theta}})\right)' \left(\sum_{i=1}^{N} \mathbf{s}_i(\tilde{\boldsymbol{\theta}}) \mathbf{s}_i(\tilde{\boldsymbol{\theta}})'\right)^{-1} \left(\sum_{i=1}^{N} \boldsymbol{s}_i(\tilde{\boldsymbol{\theta}})\right) \overset{a}{\sim} \chi^2_Q$$

- Reject H_0 whenever the numerator is too big.
- There are different variants of the LM statistic which differ in the weighting matrix.
- All have the same limit distribution under H_0 but may differ in finite sample performance and practical aspects. (Compare W2010, p.482 for a discussion).
- LM statistic may be used with $\boldsymbol{\theta}_0$ on the boundary of Θ. (Compare W2010, p.424 for a discussion).

- Testing non-nested models
  - For example Logit vs. Probit model: Which is more appropriate?
  - This is more difficult and not covered in this course.
  - Hint: A general likelihood ratio framework has been developed by Vuong (1989).

Reference: Vuong, Q. H. (1989). Likelihood ratio tests for model selection and non-nested hypotheses. Econometrica, 307-333.

See also Section 13.11.2 “Model Selection Tests” in Wooldridge (2010).

Summary

- We have seen a formal presentation of the idea of maximum likelihood methods:
  - Estimation
  - Inference
- ML estimation is used very often in empirical economics and finance. Thus a good working knowledge is essential.
- ML estimation has the desired efficiency property (smallest variance among linear and non-linear estimators) but this is bought at the expenses of full specifying the conditional distribution of *y* given ***x***.