# Lecture 1: Maximum Likelihood Estimator


Professor: Mauricio Sarrias

Universidad de Talca

2020

1. Introduction
   - Motivation
   - Maximum Likelihood Estimator
   - Identification
   - The Score Function
   - The Information Matrix
2. Asymptotic Properties
   - Consistency
   - Asymptotic Normality
3. Estimation of Variance
4. Testing
   - Intuition
   - The Trinity
   - Proof, Proof and More Proof

# Reading

Reading (Mandatory):

- (Ruud)- Chapters 14 and 15.
- Buse, A. (1982). The likelihood ratio, Wald, and Lagrange multiplier tests: An expository note. The American Statistician, 36(3a), 153-157.

Suggested:

- (Winkelmann & Boes)- Chapters 2 and 3

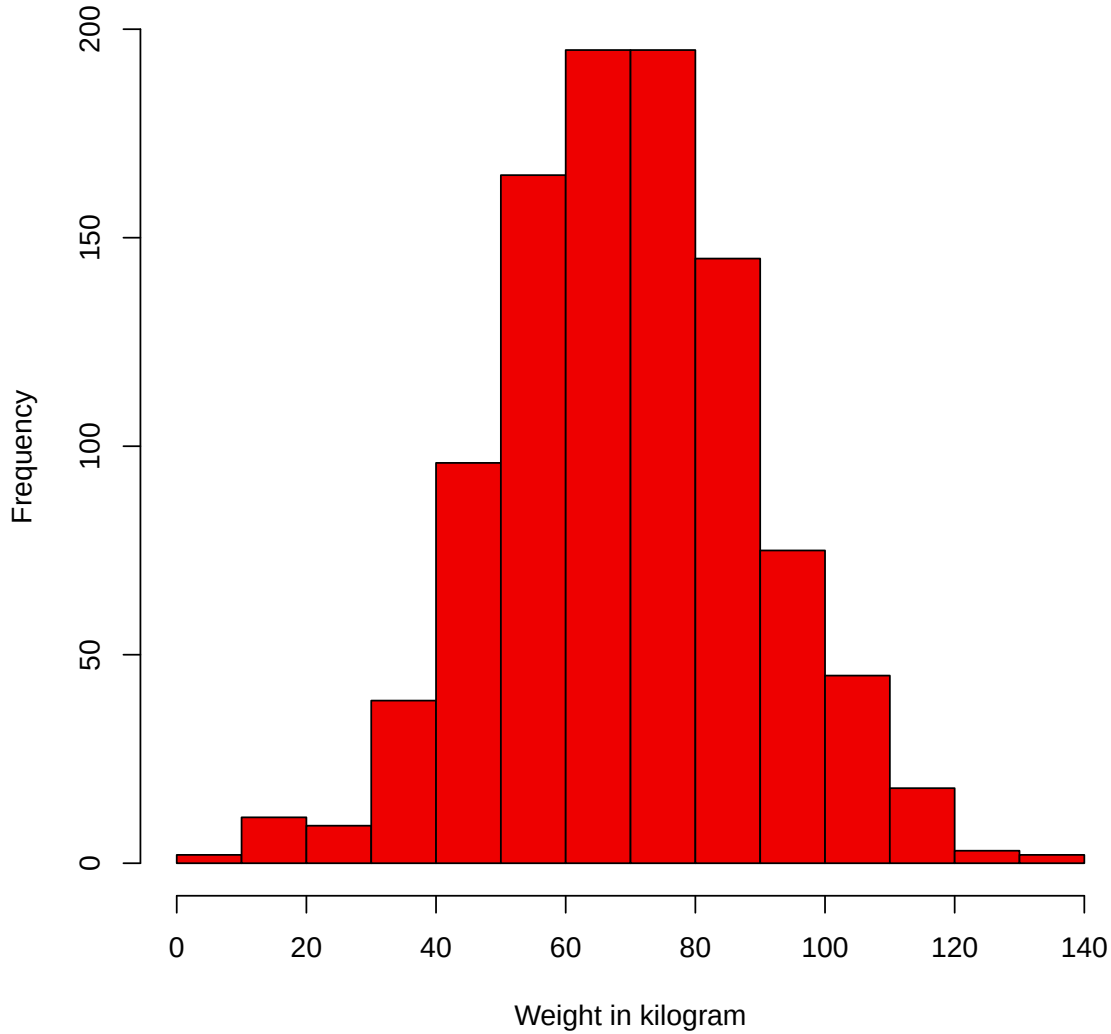
# Goals

- Understand the logic behind the Maximum Likelihood Estimator.
- Derive the asymptotic properties of the MLE.
- Understand and derive the basic test for the MLE.

1. Introduction
   - Motivation
   - Maximum Likelihood Estimator
   - Identification
   - The Score Function
   - The Information Matrix
2. Asymptotic Properties
   - Consistency
   - Asymptotic Normality
3. Estimation of Variance
4. Testing
   - Intuition
   - The Trinity
   - Proof, Proof and More Proof

# Motivating Example I

Let's assume we weighed 1000 people from Talca



# Motivating Example I

The goal of maximum likelihood is to find the optimal way to fit a distribution to the data.

## Remark I

Generally, we can writhe the probability or density function of $y_i = 1, ..., n$ as $f(y_i; \theta)$, we $y_i$ is the $i$th draw from the population and $\theta$ is the parameter of the distribution.

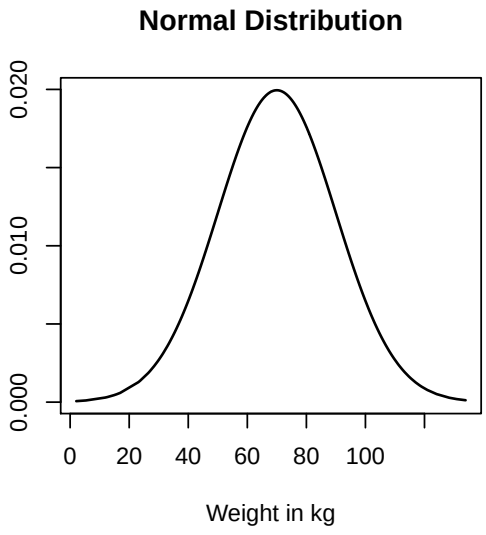
## Remark II

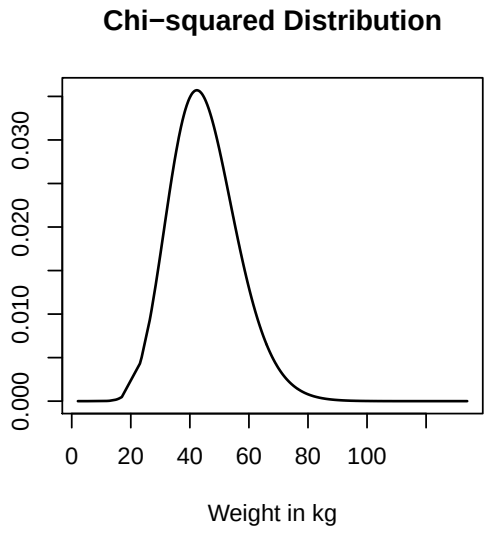
We usually assume independent sampling, i.e., the $i$th draw from the population is independent from all other draws $i' \neq i$

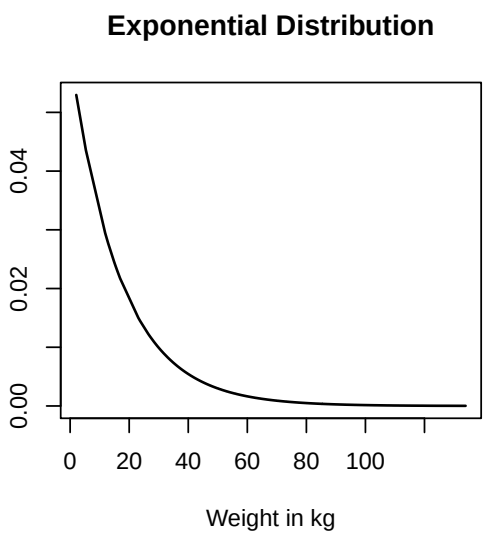
## So the question is:

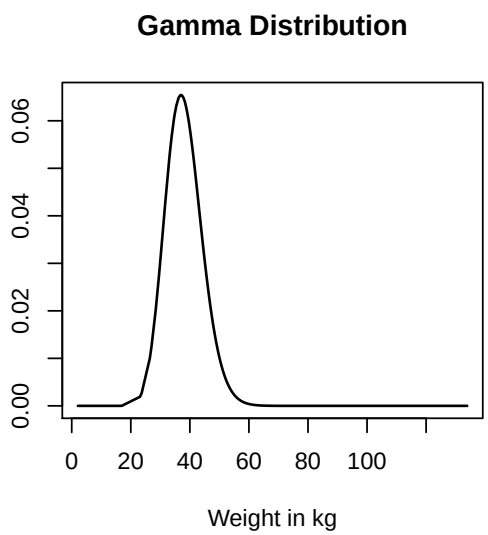
Which distribution does fit the previous weight data?

# Motivating Example I



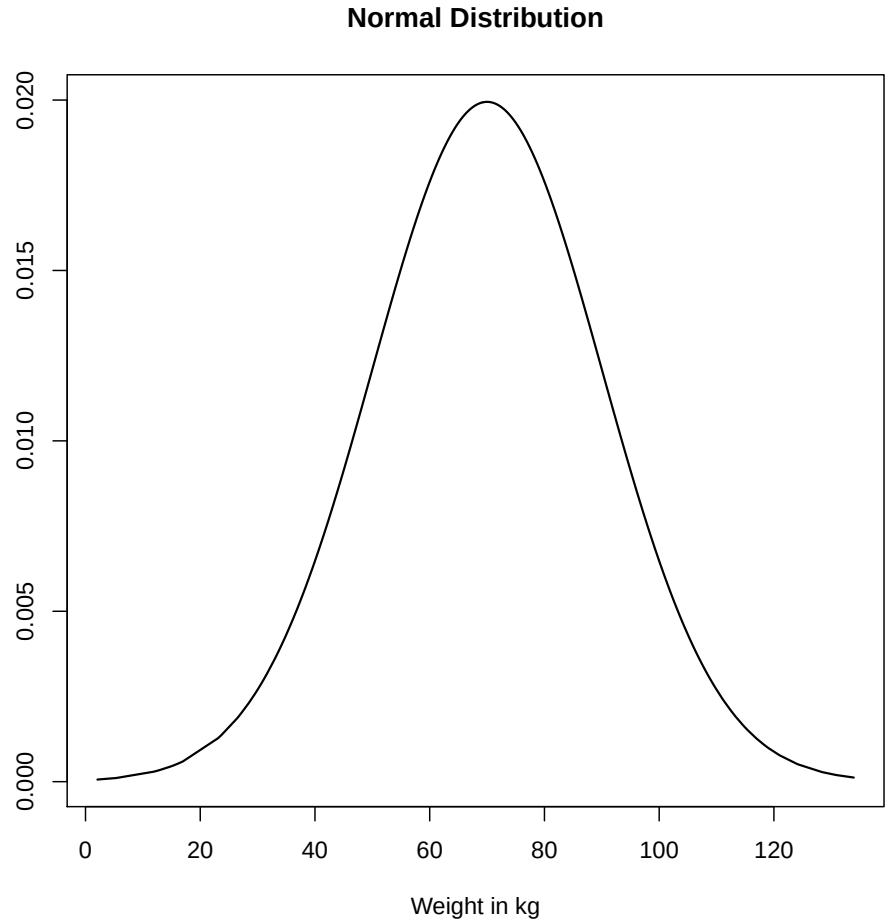



# Motivating Example I

It seems that the normal distribution is the best option.

- We expect most of the weights to be close to the mean.
- We expect the weights to be realitvely symmetrical around the mean.
- Ok ..., but not every normal fits the our data.
- What mean, $\mu$, and variance, $\sigma^2$, are the best "estimates"?

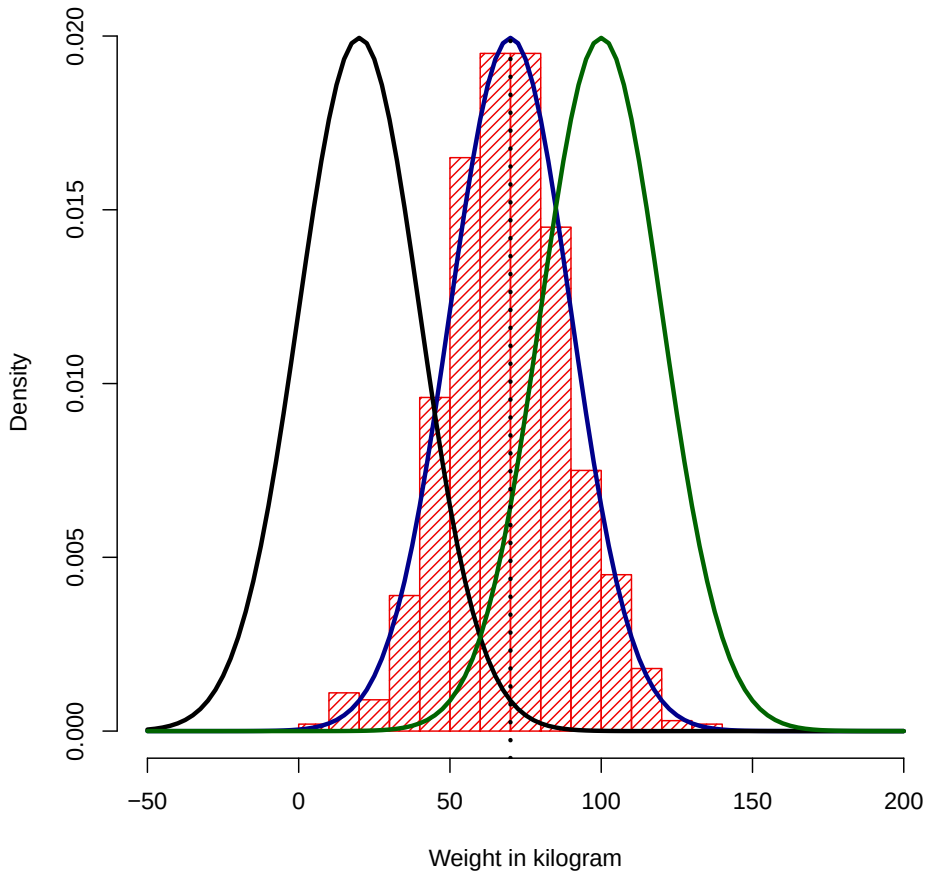


# Motivating Example I

## Maximum Likelihood Principle

1. We observe some data.
2. We pick the distribution we think generated the data.
3. We find the estimator(s) of the distribution, $\widehat{\boldsymbol{\theta}}$, that makes more likely the sample we are observing.

IOW, the problem consists on estimating an unkown parameter of a population when the population distribution is known (up to the unknown parameter)



# Motivating Example II

**Example**

A random sample of 100 trials was performed and 10 resulted in success. What can be inferred about the unknown probability of success $p_0$?

Note that we are observing the sample; somehow we know the distribution; and we are asking what is $\widehat{p}$ that makes more likely the sample we are observing.

# Motivating Example II

For any potential value of $p$ for the probability of success, the probability of $y$ successes from $n$ trail is given by:

$$f(y; n, p) = \Pr(Y = y) = \binom{n}{y} p^y (1-p)^{n-y}$$

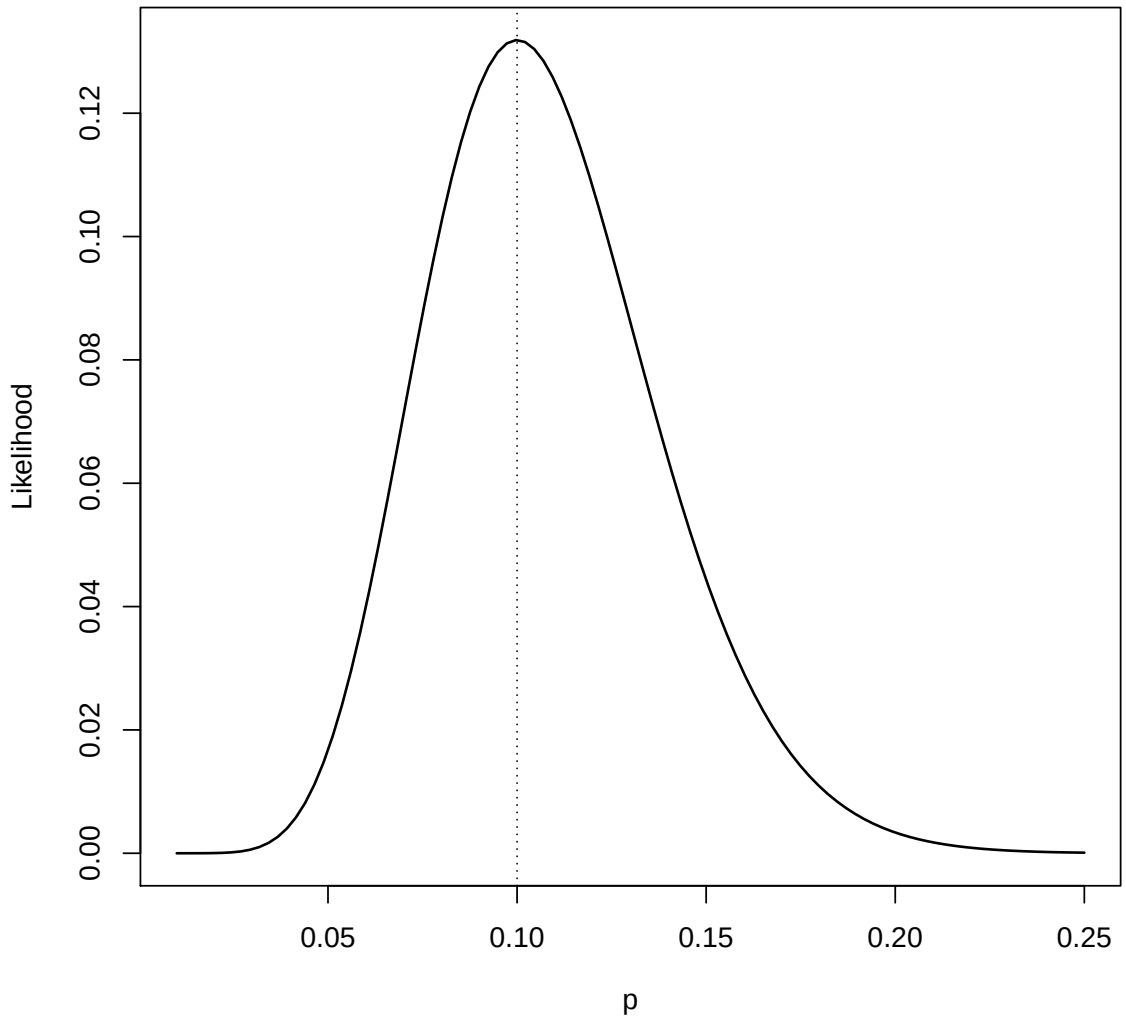
where

$$\binom{n}{y} = \frac{n!}{y!(n-y)!}$$

With $y = 10$ successes from $n = 100$ trials,

$$\begin{aligned} L(p) &= \Pr(Y = 10) \\ &= \frac{100!}{90!10!} p^{10}(1-p)^{90} \\ &= 1.731 \times 10^{13} \times p^{10}(1-p)^{90} \end{aligned}$$

# Motivating Example II



# Likelihood Function

The likelihood function denoted by capital $L$ is:

$$L(\boldsymbol{\theta}, \mathbf{y}|\mathbf{X}) = \prod_{i=1}^{n} L(\boldsymbol{\theta}; y_i|\mathbf{x}_i) = \prod_{i=1}^{n} f(y_i|\mathbf{x}_i; \boldsymbol{\theta})$$

where $\mathbf{y} = (y_1, ..., y_n)$.

- $L(\boldsymbol{\theta}; y_i|\mathbf{x}_i)$ is the likelihood contribution of the $i$-th observation,
- $L(\boldsymbol{\theta}, \mathbf{y}|\mathbf{X})$ is the likelihood function of the whole sample.

- The likelihood function says that, for any given sample $\mathbf{y}|\mathbf{X}$, the likelihood of having obtained that particular sample depends on the parameter $\boldsymbol{\theta}$.
- Whenever we can write down the joint probability function of the sample we can in principle use ML estimation.

# Log Likelihood Function

The log-likelihood function is:

$$\ln L(\boldsymbol{\theta}, \mathbf{y}|\mathbf{X}) = \ln L(\boldsymbol{\theta}) = \ln \underbrace{\left( \prod_{i=1}^{n} f(y_i|\mathbf{x}_i; \boldsymbol{\theta}) \right)}_{f(\mathbf{y}|\mathbf{X};\boldsymbol{\theta})} = \sum_{i=1}^{n} \ln f(y_i|\mathbf{x}_i; \boldsymbol{\theta})$$

- The log-likelihood function is a monotonically increasing function of $L(\boldsymbol{\theta}, \mathbf{y}|\mathbf{X})$:
  - Any maximizing value $\widehat{\boldsymbol{\theta}}$ of $\ln L(\boldsymbol{\theta}, \mathbf{y}|\mathbf{X})$ must also maximize $L(\boldsymbol{\theta}, \mathbf{y}|\mathbf{X})$.
- Taking logarithms which converts products into sums.
  - It allows some simplification in the numerical determination of the MLE.
  - Likelihood values are often extremely small (but can also be extremely large). Numerical optimization of the likelihood highly problematic.
  - Simplification of the study of the properties of the estimator.

# Example

## Example (Binomial Example)

Let $\{Y_n\}$ be a random sample of a binomial r.v with parameters $(n, p)$, where $n$ is assumed to be known and $p$ unknown. The likelihood function for individual $i$ is given by:

$$L_i(p; y_i) = f(y_i; p) = \binom{n}{y_i} p^{y_i} (1-p)^{n-y_i}$$

Since the sample is iid, the likelihood function:

$$L(p; \mathbf{y}) = f(y_1, y_2, ..., y_n; p) = \prod_{i=1}^{n} f(y_i; p) = \prod_{i=1}^{n} \binom{n}{y_i} p^{y_i} (1-p)^{n-y_i}$$

# Example

## Example (Binomial Example)

Taking the log we get the log-likelihood function:

$$
\begin{aligned}
\ln L(p;\mathbf{y}) &= \ln\left(\prod_{i=1}^{n} f(y_i;p)\right) \\
&= \ln\left(\prod_{i=1}^{n} \binom{n}{y_i} p^{y_i}(1-p)^{(n-y_i)}\right) \\
&= \ln\left[p^{\sum_{i=1}^{n} y_i}(1-p)^{\left(n^2-\sum_{n=1}^{n} y_i\right)} \prod_{i=1}^{n} \binom{n}{y_i}\right] \\
&= \ln \prod_{i=1}^{n} \binom{n}{y_i} + \left(\sum_{i=1}^{n} y_i\right) \ln p + \left(n^2 - \sum_{n=1}^{n} y_i\right) \ln(1-p)
\end{aligned}
$$

# Example: Linear Regression

## Example (Linear Regression)

Consider that $\{y_i, \mathbf{x}_i\}$ is i.i.d, and $y_i = \mathbf{x}_i^\intercal \boldsymbol{\beta}_0 + \epsilon_i$, where $\epsilon_i | \mathbf{x}_i \sim \mathrm{N}(0, \sigma_0^2)$. So, with $\boldsymbol{\theta} = (\boldsymbol{\beta}^\intercal, \sigma^2)^\intercal$ and $\mathbf{w}_i = (y_i, \mathbf{x}_i^\intercal)^\intercal$, the conditional pdf is

$$\begin{aligned} f(y_i | \mathbf{x}_i; \boldsymbol{\theta}_0) &= \frac{1}{\sqrt{2\pi\sigma_0^2}} \exp\left[ -\frac{(y_i - \mathbf{x}_i^\intercal \boldsymbol{\beta}_0)^2}{2\sigma_0^2} \right] \\ &= \phi(y_i - \mathbf{x}_i^\intercal \boldsymbol{\beta}_0, \sigma_0^2) \end{aligned}$$

The joint pdf of the sample is:

$$\begin{aligned} \prod_{i=1}^{n} f(y_i | \mathbf{x}_i; \boldsymbol{\theta}_0) &= \left[ 2\pi\sigma_0^2 \right]^{n/2} \exp\left[ -\frac{(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}_0)^\intercal (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}_0)}{2\sigma_0^2} \right] \\ &= \phi(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}_0, \sigma_0^2 \cdot \mathbf{I}_n) \end{aligned}$$

The parameter space is $\boldsymbol{\Theta}$ is $\mathbb{R}^K \times \mathbb{R}_{++}$, where $K$ is the dimension of $\boldsymbol{\beta}$ and $\mathbb{R}_{++}$ is the set of positive real numbers reflecting the a priori restriction that $\sigma_0^2 > 0$

1. Introduction
   - Motivation
   - Maximum Likelihood Estimator
   - Identification
   - The Score Function
   - The Information Matrix
2. Asymptotic Properties
   - Consistency
   - Asymptotic Normality
3. Estimation of Variance
4. Testing
   - Intuition
   - The Trinity
   - Proof, Proof and More Proof

# Maximum Likelihood Estimator

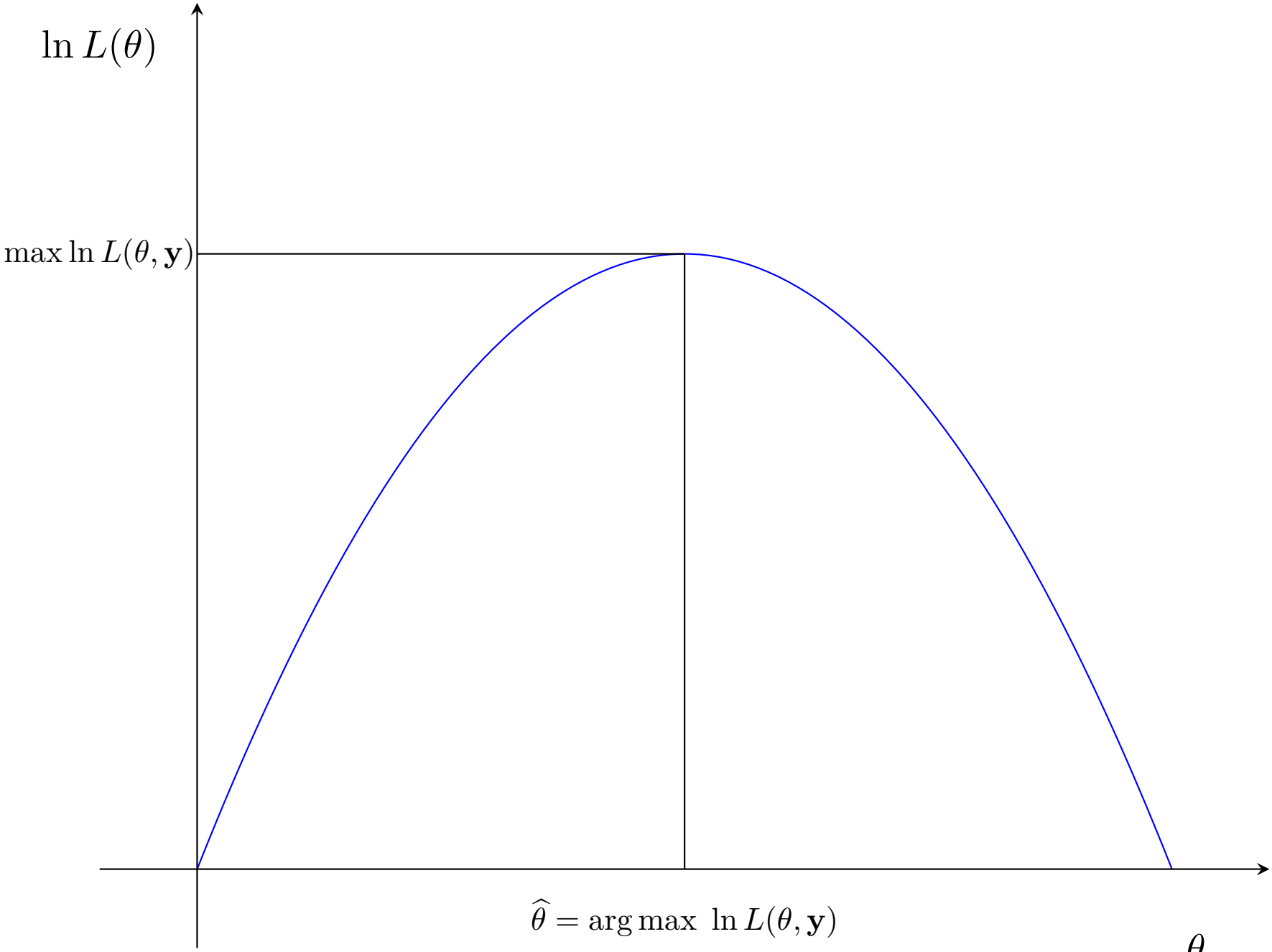
## Definition (ML Estimator)

The MLE is a value of the parameter vector that maximizes the sample average log-likelihood function:

$$\widehat{\boldsymbol{\theta}}_n \equiv \arg\max_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \frac{1}{n} \sum_{i=1}^{n} \ln f(y_i | \mathbf{x}_i; \boldsymbol{\theta})$$

where $\boldsymbol{\Theta}$ denotes the parameter space in which the parameter vector $\boldsymbol{\theta}$ lies. Usually $\boldsymbol{\Theta} = \mathbb{R}^K$.

# Maximum Likelihood Estimator: Maximization



# Maximum Likelihood Estimator

## Remark:

By the nature of the objective function, the MLE is the **estimator which makes the observed data most likely to occur**. In other words, the MLE is the best "rationalization" of what we observed.

## Population analogous

$$\mathbb{E}\left[\ln L(\boldsymbol{\theta}; \mathbf{y}|\mathbf{X})\right] \equiv \int \ln L(\boldsymbol{\theta}; \mathbf{y}|\mathbf{X}) dF(\mathbf{y}|\mathbf{X}; \boldsymbol{\theta}_0)$$

where $F(\mathbf{y}|\mathbf{X}; \boldsymbol{\theta}_0)$ is the joint CDF of $(\mathbf{y}, \mathbf{X})$

# Maximum Likelihood Estimator

## Assumption I: Distribution

The sample $\{y_i, \mathbf{x}_i\}$ is i.i.d with true conditional density $f(y_i|\mathbf{x}_i; \boldsymbol{\theta}_0)$.

Since the sample is i.i.d, we can write:

$$f(\mathbf{y}|\mathbf{X}; \boldsymbol{\theta}) = f(y_1|\mathbf{x}_1; \boldsymbol{\theta}) \times f(y_2|\mathbf{x}_2; \boldsymbol{\theta}) \times ... \times f(y_n|\mathbf{x}_n; \boldsymbol{\theta})$$

# Expected Log-Likelihood Inequality

Is $\mathbb{E}\left[\ln L(\boldsymbol{\theta}; \mathbf{y}|\mathbf{X})\right]$ maximized at $\boldsymbol{\theta}_0$?

## Assumption II: Dominance I

$\mathbb{E}\left[\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} |\ln L(\boldsymbol{\theta}; \mathbf{y}|\mathbf{X})|\right]$ exists.

## Lemma (Expected Log-likelihood Inequality)

*If Dominance I assumption holds, then*

$$\mathbb{E}\left[\ln f(y|\mathbf{x}; \boldsymbol{\theta})\right] \leq \mathbb{E}\left[\ln f(y|\mathbf{x}; \boldsymbol{\theta}_0)\right]$$

# Example

The conditional log-likelihood function of $y_i|\mathbf{x}_i \sim \mathrm{N}(\mathbf{x}_i^\intercal \boldsymbol{\beta}_0, \sigma_0^2)$ is

$$\log f(y_i|\mathbf{x}_i;\boldsymbol{\theta}) = -0.5\log(2\pi\sigma^2) - \frac{(y_i - \mathbf{x}_i^\intercal\boldsymbol{\beta})^2}{2\sigma^2} \tag{1}$$

The conditional expectation is:

$$\mathbb{E}\left[\log f(y_i|\mathbf{x}_i;\boldsymbol{\theta})\middle|\,\mathbf{x}_i\right] = -0.5\log(2\pi\sigma^2) - \frac{\mathbb{E}\left[(y_i - \mathbf{x}_i^\intercal\boldsymbol{\beta})^2\right]}{2\sigma^2}$$

Note that:

$$\begin{aligned}
\mathbb{E}\left[(y_i - \mathbf{x}_i^\intercal\boldsymbol{\beta})^2\right] &= \mathbb{E}\left[(\mathbf{x}_i^\intercal\boldsymbol{\beta}_0 + \epsilon_i - \mathbf{x}_i^\intercal\boldsymbol{\beta})^2\right] \quad \because y_i = \mathbf{x}_i^\intercal\boldsymbol{\beta}_0 + \epsilon_i \\
&= \mathbb{E}\left[(\epsilon_i + \mathbf{x}_i^\intercal(\boldsymbol{\beta}_0 - \boldsymbol{\beta}))^2\right] \\
&= \mathbb{E}\left[\epsilon_i^2 + 2\epsilon_i\mathbf{x}_i^\intercal(\boldsymbol{\beta}_0 - \boldsymbol{\beta}) + (\boldsymbol{\beta}_0 - \boldsymbol{\beta})^\intercal\mathbf{x}_i\mathbf{x}_i^\intercal(\boldsymbol{\beta}_0 - \boldsymbol{\beta})\right] \\
&= \mathbb{E}(\epsilon_i^2) + (\boldsymbol{\beta}_0 - \boldsymbol{\beta})^\intercal\mathbb{E}(\mathbf{x}_i\mathbf{x}_i^\intercal)(\boldsymbol{\beta}_0 - \boldsymbol{\beta}) \\
&= \sigma_0^2 + (\boldsymbol{\beta}_0 - \boldsymbol{\beta})^\intercal\mathbb{E}(\mathbf{x}_i\mathbf{x}_i^\intercal)(\boldsymbol{\beta}_0 - \boldsymbol{\beta})
\end{aligned}$$

When is this expectation finite?

# Example

The last term is finite if $\mathbb{E}(\mathbf{x}_i\mathbf{x}_i^\intercal)$ is. This implies that $\mathbf{X}$ is full-column rank.

$$\mathbb{E}\left[\log f(y_i|\mathbf{x}_i;\boldsymbol{\theta})\middle|\,\mathbf{x}_i\right] = -0.5\log(2\pi\sigma^2) - \frac{\sigma_0^2 + (\boldsymbol{\beta}_0 - \boldsymbol{\beta})^\intercal\mathbb{E}(\mathbf{x}_i\mathbf{x}_i^\intercal)(\boldsymbol{\beta}_0 - \boldsymbol{\beta})}{2\sigma^2}$$

Now, $\mathbb{E}\left[\log f(y_i|\mathbf{x}_i;\boldsymbol{\theta})\middle|\,\mathbf{x}_i\right]$ is uniquely maximized at $\mathbf{x}_i^\intercal\boldsymbol{\beta} = \mathbf{x}_i^\intercal\boldsymbol{\beta}_0$ and $\sigma^2 = \sigma_0^2$

1. Introduction
   - Motivation
   - Maximum Likelihood Estimator
   - Identification
   - The Score Function
   - The Information Matrix
2. Asymptotic Properties
   - Consistency
   - Asymptotic Normality
3. Estimation of Variance
4. Testing
   - Intuition
   - The Trinity
   - Proof, Proof and More Proof

# Identification

- Before employing MLE, it is necessary to check whether the data-generating process is sufficiently informative about the parameters of the model.
- Recall OLS: $\widehat{\boldsymbol{\beta}}$ to be unique $\mathbf{X}$ must be full-column rank. Otherwise, ...
- The question is: is the population $\mathbb{E}\left[\ln f(y_i|\mathbf{x}_i;\boldsymbol{\theta})\right]$ uniquely maximized at $\boldsymbol{\theta}_0$?
  - If there exists another $\boldsymbol{\theta} \neq \boldsymbol{\theta}_0$ that maximized $\mathbb{E}\left[\ln f(y_i|\mathbf{x}_i;\boldsymbol{\theta})\right]$, then MLE is not identified.
- This is satisfied if (conditional density identification):

$$f(y_i|\mathbf{x}_i;\boldsymbol{\theta}) \neq f(y_i|\mathbf{x}_i;\boldsymbol{\theta}_0) \quad \forall \boldsymbol{\theta} \neq \boldsymbol{\theta}_0$$

# Identification

## Definition (Global Identification)

The parameter vector $\boldsymbol{\theta}_0$ is globally identified in $\boldsymbol{\Theta}$ if, for every $\boldsymbol{\theta}_1 \in \boldsymbol{\Theta}$, $\boldsymbol{\theta} \neq \boldsymbol{\theta}_1$ implies that:

$$\Pr\left[f(y_i|\mathbf{x}_i;\boldsymbol{\theta}_0) \neq f(y_i|\mathbf{x}_i;\boldsymbol{\theta}_1)\right] > 0$$

## Assumption III: Global Identification

Every parameter vector $\boldsymbol{\theta}_0 \in \boldsymbol{\Theta}$ is globally identified.

# Identification

## Lemma (Strict Expected Log-Likelihood Inequality)

*Under the Assumptions of* ***Distribution, Dominance I and Global Identification****, then*

$$\boldsymbol{\theta} \neq \boldsymbol{\theta}_0 \implies \mathbb{E}\left[\ln f(y|\mathbf{x};\boldsymbol{\theta})\right] < \mathbb{E}\left[\ln f(y|\mathbf{x};\boldsymbol{\theta}_0)\right]$$

In words, the expected value of the log-likelihood is maximized at the true value of the parameters.

## Proof.

Let $\mathbf{w} = (y, \mathbf{x}')'$ and define

$$a(\mathbf{w}) \equiv f(y|\mathbf{x};\boldsymbol{\theta})/f(y|\mathbf{x};\boldsymbol{\theta}_0)$$

First, WTS that $a(\mathbf{w}) \neq 1$ with positive probability, so that $a(\mathbf{w})$ is nonconstant random variable (so, we can apply Jensen's Inequality).

$$\begin{aligned} a(\mathbf{w}) \neq 1 &\iff f(y|\mathbf{x};\boldsymbol{\theta}) \neq f(y|\mathbf{x};\boldsymbol{\theta}_0) \\ \Pr\left[a(\mathbf{w}) \neq 1\right] &\iff \Pr\left[f(y|\mathbf{x};\boldsymbol{\theta}) \neq f(y|\mathbf{x};\boldsymbol{\theta}_0)\right] \end{aligned}$$

But, by Global Identification:

$$\Pr\left[f(y|\mathbf{x};\boldsymbol{\theta}) \neq f(y|\mathbf{x};\boldsymbol{\theta}_0)\right] > 0 \implies \Pr\left[a(\mathbf{w}) \neq 1\right] > 0$$

Now, WTS $\mathbb{E}\left[\log a(\mathbf{w})\right] < \log\left\{\mathbb{E}\left[a(\mathbf{w})\right]\right\}$. We use the strict version of **Jensen's inequality** which states that if $c(x)$ is a strictly concave function and $x$ is nonconstant random variable, then $\mathbb{E}\left[c(x)\right] < c\left[\mathbb{E}(x)\right]$ □

## Proof.

Set $c(x) = \log(x)$, since $\log(x)$ is strictly concave and $a(\mathbf{w})$ is non-constant. Therefore

$$\mathbb{E}\left[\log a(\mathbf{w})\right] < \log\left\{\mathbb{E}\left[a(\mathbf{w})\right]\right\}$$

Now, WTS that $\mathbb{E}(a(\mathbf{w})) = 1$. Note that the conditional mean of $a(\mathbf{w})$ equals 1 because:

$$\begin{aligned}
\mathbb{E}\left[a(\mathbf{w})|\,\mathbf{x}\right] &= \int a(\mathbf{w}) f(y|\mathbf{x};\boldsymbol{\theta}_0) dy \\
&= \int \frac{f(y|\mathbf{x};\boldsymbol{\theta})}{f(y|\mathbf{x};\boldsymbol{\theta}_0)} f(y|\mathbf{x};\boldsymbol{\theta}_0) dy \\
&= \int f(y|\mathbf{x};\boldsymbol{\theta}) dy \\
&= 1
\end{aligned}$$

By the Law of Total Expectations $\mathbb{E}\left[a(\mathbf{w})\right] = 1$. Combining the results:

$$\mathbb{E}\left[\log(a(\mathbf{w}))\right] < \log(1) = 0$$

But $\log(a(\mathbf{w})) = \log f(y|\mathbf{x};\boldsymbol{\theta}) - \log f(y|\mathbf{x};\boldsymbol{\theta}_0)$. □

1. Introduction
   - Motivation
   - Maximum Likelihood Estimator
   - Identification
   - The Score Function
   - The Information Matrix
2. Asymptotic Properties
   - Consistency
   - Asymptotic Normality
3. Estimation of Variance
4. Testing
   - Intuition
   - The Trinity
   - Proof, Proof and More Proof

# Differentiability

## Assumption IV: Integrability

The pdf $f(y_i|\mathbf{x}_i;\boldsymbol{\theta})$ is twice continuously differentiable in $\boldsymbol{\theta}$ for all $\boldsymbol{\theta} \in \boldsymbol{\Theta}$. Furthermore, the support $\mathcal{S}(\boldsymbol{\theta})$ of $f(y_i|\mathbf{x}_i;\boldsymbol{\theta})$ does not depend on $\boldsymbol{\theta}$, and differentiation and integration are interchangeable in the sense that

$$\begin{aligned}
\frac{\partial}{\partial \boldsymbol{\theta}} \int_{\mathcal{S}} dF(y_i|\mathbf{x}_i;\boldsymbol{\theta}) &= \int_{\mathcal{S}} \frac{\partial}{\partial \boldsymbol{\theta}} dF(y_i|\mathbf{x}_i;\boldsymbol{\theta}) \\
\frac{\partial^2}{\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}'} \int_{\mathcal{S}} dF(y_i|\mathbf{x}_i;\boldsymbol{\theta}) &= \int_{\mathcal{S}} \frac{\partial}{\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}'} dF(y_i|\mathbf{x}_i;\boldsymbol{\theta})
\end{aligned}$$

and

$$\begin{aligned}
\frac{\partial \mathbb{E}\left[\ln f(y_i|\mathbf{x}_i;\boldsymbol{\theta}) \middle| \mathbf{x}_i = x_i\right]}{\partial \boldsymbol{\theta}} &= \mathbb{E}\left[\frac{\partial \ln f(y_i|\mathbf{x}_i;\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \middle| \mathbf{x}_i = x_i\right] \\
\frac{\partial^2 \mathbb{E}\left[\ln f(y_i|\mathbf{x}_i;\boldsymbol{\theta}) \middle| \mathbf{x}_i = x_i\right]}{\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}'} &= \mathbb{E}\left[\frac{\partial^2 \ln f(y_i|\mathbf{x}_i;\boldsymbol{\theta})}{\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}'} \middle| \mathbf{x}_i = x_i\right]
\end{aligned}$$

where all terms exists. In this case, we denote the support of $F(y)$ simply by $\mathcal{S}$.

# The Score Function

## Definition (Score Function)

The score function is defined as the vector of first partial derivatives of the log-likelihood function with respect to the parameter vector $\boldsymbol{\theta}$:

$$\mathbf{s}(\mathbf{w}, \boldsymbol{\theta}) = \frac{\partial \ln f(\mathbf{y}|\mathbf{X}; \boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \begin{pmatrix} \frac{\partial \ln f(\mathbf{y}|\mathbf{X}; \boldsymbol{\theta})}{\partial \theta_1} \\ \frac{\partial \ln f(\mathbf{y}|\mathbf{X}; \boldsymbol{\theta})}{\partial \theta_2} \\ \vdots \\ \frac{\partial \ln f(\mathbf{y}|\mathbf{X}; \boldsymbol{\theta})}{\partial \theta_K} \end{pmatrix}$$

The score vector for observation $i$ is:

$$\mathbf{s}(\mathbf{w}_i; \boldsymbol{\theta}) = \frac{\partial \ln f(y_i|\mathbf{x}_i; \boldsymbol{\theta})}{\partial \boldsymbol{\theta}}$$

Because of the additivity of terms in the log-likelihood function, we can write:

$$\mathbf{s}(\mathbf{w}, \boldsymbol{\theta}) = \sum_{i=1}^{n} \mathbf{s}(\mathbf{w}_i; \boldsymbol{\theta})$$

# Score Identity

## Lemma (Score Identity)

*Under* ***Integrability and Distribution Assumption:***

$$\mathbb{E}\left[\mathbf{s}(\mathbf{w};\boldsymbol{\theta})\right] = 0$$

- We have to be clear whether we are speaking about the score of a single observation $\mathbf{s}(\mathbf{w}_i;\boldsymbol{\theta})$ or the score of the sample $\mathbf{s}(\mathbf{w};\boldsymbol{\theta})$.
- Since under random sampling, $\mathbf{s}(\mathbf{w},\boldsymbol{\theta}) = \sum_{i=1}^{n} \mathbf{s}(\mathbf{w}_i;\boldsymbol{\theta})$, it is sufficient to establish that $\mathbb{E}\left[\mathbf{s}(\mathbf{w}_i;\boldsymbol{\theta})\right] = \mathbf{0}$

## Proof.

First, we derive an integral property of pdf. Because we are **assuming** $F(y|x;\boldsymbol{\theta})$ is a proper cdf.,

$$\int_{\mathcal{S}} dF(y_i|\mathbf{x}_i;\boldsymbol{\theta}) = \int_{\mathcal{S}} f(y_i|\mathbf{x}_i;\boldsymbol{\theta})dy_i = 1 \tag{2}$$

$\forall \boldsymbol{\theta} \in \boldsymbol{\Theta}$. Given **differentiability**, we can differentiate both sides of this equality with respect to $\boldsymbol{\theta}$:

$$\mathbf{0} = \int_{\mathcal{S}} \frac{\partial}{\partial \boldsymbol{\theta}} f(y_i|\mathbf{x}_i;\boldsymbol{\theta})dy_i \tag{3}$$

This equation states how changes in $f(y_i|\mathbf{x}_i;\boldsymbol{\theta})$ resulting from changes in $\boldsymbol{\theta}$ are restricted by (2). We can rewrite (3) as

$$\begin{aligned}
\mathbf{0} &= \int_{\mathcal{S}} \frac{f(y_i|\mathbf{x}_i;\boldsymbol{\theta})}{f(y_i|\mathbf{x}_i;\boldsymbol{\theta})} \frac{\partial}{\partial \boldsymbol{\theta}} f(y_i|\mathbf{x}_i;\boldsymbol{\theta})dy_i \\
\mathbf{0} &= \int_{\mathcal{S}} \frac{1}{f(y_i|\mathbf{x}_i;\boldsymbol{\theta})} \frac{\partial f(y_i|\mathbf{x}_i;\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \underbrace{dF(y_i|\mathbf{x}_i;\boldsymbol{\theta})}_{f(y_i|\mathbf{x}_i;\boldsymbol{\theta})dy_i}
\end{aligned} \tag{4}$$

□

## Proof.

Now we interpret this integral equation as an expectation. Consider:

$$\begin{aligned}
\frac{\partial}{\partial \boldsymbol{\theta}} \ln f(y_i|\mathbf{x}_i;\boldsymbol{\theta}) &\equiv \frac{1}{f(y_i|\mathbf{x}_i;\boldsymbol{\theta})} \frac{\partial}{\partial \boldsymbol{\theta}} f(y_i|\mathbf{x}_i;\boldsymbol{\theta}) \\
\mathbf{s}(\mathbf{w}_i;\boldsymbol{\theta}) &\equiv \frac{1}{f(y_i|\mathbf{x}_i;\boldsymbol{\theta})} \frac{\partial}{\partial \boldsymbol{\theta}} f(y_i|\mathbf{x}_i;\boldsymbol{\theta}) \qquad (5) \\
\mathbf{s}(\mathbf{w}_i;\boldsymbol{\theta}) f(y_i|\mathbf{x}_i;\boldsymbol{\theta}) &\equiv \frac{\partial}{\partial \boldsymbol{\theta}} f(y_i|\mathbf{x}_i;\boldsymbol{\theta})
\end{aligned}$$

Then, substituting into (4)

$$\int_{\mathcal{S}} \mathbf{s}(\mathbf{w}_i;\boldsymbol{\theta}) dF(y_i|\mathbf{x}_i;\boldsymbol{\theta}) = \mathbf{0}$$

This hold for any $\boldsymbol{\theta} \in \boldsymbol{\Theta}$, in particular, for $\boldsymbol{\theta} = \boldsymbol{\theta}_0$. Setting $\boldsymbol{\theta} = \boldsymbol{\theta}_0$, we obtain:

$$\int_{\mathcal{S}} \mathbf{s}(\mathbf{w}_i;\boldsymbol{\theta}) dF(y_i|\mathbf{x}_i;\boldsymbol{\theta}) = \mathbf{0}$$

$$\int_{\mathcal{S}} \mathbf{s}(\mathbf{w}_i;\boldsymbol{\theta}_0) dF(y_i|\mathbf{x}_i;\boldsymbol{\theta}_0) = \mathbb{E}\left[\mathbf{s}(\mathbf{w}_i;\boldsymbol{\theta}_0)\middle|\, \mathbf{x}\right] = \mathbf{0}$$

Then, by Law of Total Expectations, we obtain the desired result.

□

# What if the support depend on $\boldsymbol{\theta}$?

In this case the support is $\mathcal{S}(\boldsymbol{\theta}) = A(\boldsymbol{\theta}) \leq y \leq B(\boldsymbol{\theta})$. By definition:

$$\int_{A(\boldsymbol{\theta})}^{B(\boldsymbol{\theta})} f(y|x;\boldsymbol{\theta})dy = 1$$

Now, using the Leibnitz's theorem gives:

$$\begin{aligned} \frac{\partial \int_{A(\boldsymbol{\theta})}^{B(\boldsymbol{\theta})} f(y|x;\boldsymbol{\theta})dy}{\partial \boldsymbol{\theta}} &= \mathbf{0} \\ \int_{A(\boldsymbol{\theta})}^{B(\boldsymbol{\theta})} \frac{\partial f(y|x;\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} dy + f(B(\boldsymbol{\theta})|\boldsymbol{\theta}) \frac{\partial B(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} - f(A(\boldsymbol{\theta})|\boldsymbol{\theta}) \frac{\partial A(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} &= \mathbf{0} \end{aligned}$$

To interchange the operations of differentiation and integration we need the second and third terms go to zero. The **necessary condition** is that

$$\begin{aligned} \lim_{y \to A(\boldsymbol{\theta})} f(y|x;\boldsymbol{\theta}) &= \mathbf{0} \\ \lim_{y \to B(\boldsymbol{\theta})} f(y|x;\boldsymbol{\theta}) &= \mathbf{0} \end{aligned}$$

**Sufficient conditions** are that the support does not depend on the parameter, which means that $\partial A(\boldsymbol{\theta})/\partial \boldsymbol{\theta} = \partial B(\boldsymbol{\theta})/\partial \boldsymbol{\theta} = \mathbf{0}$ or that the density is zero at the terminal points.

1. Introduction
   - Motivation
   - Maximum Likelihood Estimator
   - Identification
   - The Score Function
   - The Information Matrix
2. Asymptotic Properties
   - Consistency
   - Asymptotic Normality
3. Estimation of Variance
4. Testing
   - Intuition
   - The Trinity
   - Proof, Proof and More Proof

# Hessian

- Since we are doing an optimization analysis, we need the Hessian Matrix.

$$\mathbf{H}(\mathbf{w};\boldsymbol{\theta}) = \frac{\partial^2 \ln f(\mathbf{y}|\mathbf{X};\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^\intercal} = \begin{pmatrix} \frac{\partial^2 \ln f(\mathbf{y}|\mathbf{X};\boldsymbol{\theta})}{\partial\theta_1^2} & \frac{\partial^2 \ln f(\mathbf{y}|\mathbf{X};\boldsymbol{\theta})}{\partial\theta_1\partial\theta_2} & \cdots & \frac{\partial^2 \ln f(\mathbf{y}|\mathbf{X};\boldsymbol{\theta})}{\partial\theta_1\partial\theta_K} \\ \frac{\partial^2 \ln f(\mathbf{y}|\mathbf{X};\boldsymbol{\theta})}{\partial\theta_2\partial\theta_1} & \frac{\partial^2 \ln f(\mathbf{y}|\mathbf{X};\boldsymbol{\theta})}{\partial\theta_2^2} & \cdots & \frac{\partial^2 \ln f(\mathbf{y}|\mathbf{X};\boldsymbol{\theta})}{\partial\theta_2\partial\theta_K} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 \ln f(\mathbf{y}|\mathbf{X};\boldsymbol{\theta})}{\partial\theta_K\partial\theta_1} & \frac{\partial^2 \ln f(\mathbf{y}|\mathbf{X};\boldsymbol{\theta})}{\partial\theta_2\partial\theta_K} & \cdots & \frac{\partial^2 \ln f(\mathbf{y}|\mathbf{X};\boldsymbol{\theta})}{\partial\theta_K^2} \end{pmatrix}$$

If the log-likelihood function is concave in $\boldsymbol{\theta}$, $\mathbf{H}(\mathbf{w};\boldsymbol{\theta})$ is said to be negative definite. In the scalar case, for $K = 1$, this simply means that the second derivative of the log-likelihood function is negative.

# Hessian

Because of the additivity of terms in the log-likelihood function:

$$\mathbf{H}(\mathbf{w}_i;\boldsymbol{\theta}) = \sum_{i=1}^{n} \mathbf{H}(\mathbf{w}_i;\boldsymbol{\theta}) \quad \text{where} \quad \mathbf{H}(\mathbf{w}_i;\boldsymbol{\theta}) = \frac{\partial^2 \ln f(y_i|\mathbf{x}_i;\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^\intercal}$$

**Remark**

It is important to keep in mind that both the score and Hessian depend on the sample and are therefore random variables (they differ in repeated samples).

# Information Identity

- To analyze the variance and the limiting distribution of the ML estimator, we require some results on the **Fisher information matrix**.
- It is very related to the Hessian matrix.
- The information matrix of a sample is simply defined as the negative expectation of the Hessian Matrix:

$$\mathbf{I}(\boldsymbol{\theta}) = -\mathbb{E}\left[\mathbf{H}(\mathbf{w}, \boldsymbol{\theta})\right]$$

- Why is it useful?
  - It can be used to assess whether the likelihood function is "well behaved" (Identification)
  - Important result: the information matrix is the inverse of the variance of the maximum likelihood estimator.
  - Cramér Rao lower bound.

# Information matrix equality

## Information matrix equality

The information matrix can be derived in two ways, either as minus the expected Hessian, or alternative as the variance of the score function, both evaluated at the true parameter $\boldsymbol{\theta}_0$

# Information Identity

## Assumption V: Finite Information

$\operatorname{Var}\left[\frac{\partial}{\partial \boldsymbol{\theta}} \ln f(\mathbf{y}|\mathbf{X};\boldsymbol{\theta})\right] \equiv \operatorname{Var}\left[\mathbf{s}(\mathbf{w};\boldsymbol{\theta})\right]$ exists.

## Lemma (Information Identity)

*Under* ***Distribution, Differentiability and Finite Information*** *Assumption:*

$$\mathbb{E}\left[\frac{\partial^2}{\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}^\intercal} \ln f(\mathbf{y}|\mathbf{X};\boldsymbol{\theta})\right] = -\operatorname{Var}\left[\mathbf{s}(\mathbf{w};\boldsymbol{\theta})\right]$$

Proof: (Homework)

# Information Identity

Note the following:

$$\begin{aligned}\operatorname{Var}\left[\mathbf{s}(\mathbf{w}_i;\boldsymbol{\theta}_0)\right] &= \mathbb{E}\left[\underbrace{\mathbf{s}(\mathbf{w}_i;\boldsymbol{\theta}_0)}_{(K\times 1)}\underbrace{\mathbf{s}(\mathbf{w}_i;\boldsymbol{\theta}_0)^{\intercal}}_{(1\times K)}\right] + \underbrace{\mathbb{E}\left[\mathbf{s}(\mathbf{w}_i;\boldsymbol{\theta}_0)\right]}_{=\mathbf{0}}\mathbb{E}\left[\mathbf{s}(\mathbf{w}_i;\boldsymbol{\theta}_0)\right]^{\intercal} \\ &= \mathbb{E}\left[\mathbf{s}(\mathbf{w}_i;\boldsymbol{\theta}_0)\mathbf{s}(\mathbf{w}_i;\boldsymbol{\theta}_0)^{\intercal}\right]\end{aligned}$$

Therefore we can write:

$$-\mathbf{I}(\boldsymbol{\theta}_0) = \mathbb{E}\left[\mathbf{H}(\mathbf{w}_i;\boldsymbol{\theta}_0)\right] = -\operatorname{Var}\left[\mathbf{s}(\mathbf{w}_i;\boldsymbol{\theta}_0)\right] = -\mathbb{E}\left[\mathbf{s}(\mathbf{w}_i;\boldsymbol{\theta}_0)\mathbf{s}(\mathbf{w}_i;\boldsymbol{\theta}_0)^{\intercal}\right]$$

# Example

Recall that:

$$\log f(y_i|\mathbf{x}_i;\boldsymbol{\theta}) = -0.5\log(2\pi\sigma^2) - \frac{(y_i - \mathbf{x}_i^\intercal\boldsymbol{\beta})^2}{2\sigma^2}$$

We have:

$$\mathbf{s}(\mathbf{w}_i;\boldsymbol{\theta}) = \begin{pmatrix} \frac{1}{\sigma^2}\mathbf{x}_i \cdot \widehat{\epsilon}_i \\ -\frac{1}{2\sigma^2} + \frac{1}{2\sigma^4}\widehat{\epsilon}_i^2 \end{pmatrix}$$

$$\mathbf{H}(\mathbf{w}_i;\boldsymbol{\theta}) = \begin{pmatrix} -\frac{1}{\sigma^2}\mathbf{x}_i\mathbf{x}_i^\intercal & -\frac{1}{\sigma^4}\mathbf{x}_i \cdot \widehat{\epsilon}_i \\ -\frac{1}{\sigma^4}\mathbf{x}_i^\intercal \cdot \widehat{\epsilon}_i & \frac{1}{2\sigma^4} - \frac{1}{\sigma^6}\widehat{\epsilon}_i^2 \end{pmatrix}$$

$$\mathbf{s}(\mathbf{w}_i;\boldsymbol{\theta})\mathbf{s}(\mathbf{w}_i;\boldsymbol{\theta})^\intercal = \begin{pmatrix} \frac{1}{\sigma^4}\mathbf{x}_i\mathbf{x}_i^\intercal\widehat{\epsilon}_i^2 & -\frac{1}{2\sigma^4}\mathbf{x}_i \cdot \widehat{\epsilon}_i + \frac{1}{2\sigma^6}\mathbf{x}_i \cdot \widehat{\epsilon}_i^3 \\ -\frac{1}{2\sigma^4}\mathbf{x}_i^\intercal \cdot \widehat{\epsilon}_i + \frac{1}{2\sigma^6}\mathbf{x}_i^\intercal \cdot \widehat{\epsilon}_i^3 & \frac{1}{4\sigma^4} - \frac{1}{2\sigma^6}\widehat{\epsilon}_i^2 + \frac{1}{4\sigma^8}\widehat{\epsilon}_i^4 \end{pmatrix}$$

where $\mathbf{w}_i = (y_i, \mathbf{x}_i^\intercal)^\intercal$, $\boldsymbol{\theta} = (\boldsymbol{\theta}^\intercal, \sigma^2)^\intercal$ and $\widehat{\epsilon}_i \equiv y_i - \mathbf{x}_i^\intercal\boldsymbol{\beta}$

# Example

So for $\boldsymbol{\theta} = \boldsymbol{\theta}_0$ the $\widehat{\epsilon}_i$ in these expressions can be replaced by $\epsilon_i$. In the linear regression model, $\mathbb{E}(\epsilon_i|\mathbf{x}_i) = 0$. Also, since $\epsilon_i \sim \mathrm{N}(0, \sigma_0^2)$, we have $\mathbb{E}(\epsilon_i^3) = 0$ and $\mathbb{E}(\epsilon_i^4) = 3\sigma_0^4$. Using these relations, we have:

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

If $\mathbb{E}(\mathbf{x}_i\mathbf{x}_i^\intercal)$ is nonsingular, then $\mathbb{E}\left[\mathbf{H}(\mathbf{w}_i; \boldsymbol{\theta}_0)\right]$ is nonsingular.

1 Introduction
- Motivation
- Maximum Likelihood Estimator
- Identification
- The Score Function
- The Information Matrix

2 Asymptotic Properties
- Consistency
- Asymptotic Normality

3 Estimation of Variance

4 Testing
- Intuition
- The Trinity
- Proof, Proof and More Proof

# Some Ideas

- For OLS estimator consistency can be shown by finding the sampling error function and applying LLN.
- This cannot be done for nonlinear estimator such as MLE since closed form solution for finite sample estimators do not exists.
- That is, the MLE is an implicit function of the random sample.

**Question**

How can we proceed?

# Some Ideas

Using some LLN we know that:

$$\frac{1}{n}\sum_{i=1}^{n}\log f(y_i|\mathbf{x}_i;\boldsymbol{\theta}) \xrightarrow{p} \mathbb{E}\left[\log f(y_i|\mathbf{x}_i;\boldsymbol{\theta})\right] \tag{6}$$

for any fixed parameter value $\boldsymbol{\theta}$. That is, the sample average log-likelihood function converges to the expected log-likelihood for any value of $\boldsymbol{\theta}$. Recall that:

$$\widehat{\boldsymbol{\theta}}_n \equiv \arg\max_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \frac{1}{n}\sum_{i=1}^{n}\log f(y_i|\mathbf{x}_i;\boldsymbol{\theta})$$

$$\boldsymbol{\theta}_0 \equiv \arg\max_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \mathbb{E}\left[\log f(y_i|\mathbf{x}_i;\boldsymbol{\theta})\right]$$

We would like to say that, given that

$\frac{1}{n}\sum_{i=1}^{n}\log f(y_i|\mathbf{x}_i;\boldsymbol{\theta}) \xrightarrow{p} \mathbb{E}\left[\log f(y_i|\mathbf{x}_i;\boldsymbol{\theta})\right]$, then $\widehat{\boldsymbol{\theta}}_n \xrightarrow{p} \boldsymbol{\theta}_0$

# Some Ideas

We might be able to do this using the **continuous mapping theorem**.

- Let $X_n = \frac{1}{n}\sum_{i=1}^{n} \log f(y_i|\mathbf{x}_i;\boldsymbol{\theta})$,
- and $g(\cdot) = \underset{\boldsymbol{\theta}\in\boldsymbol{\Theta}}{\arg\max}(\cdot)$

Then we would like to say that if $X_n \xrightarrow{p} X$ then $g(X_n) \xrightarrow{p} g_0(X)$.

**In words:**

If the sample average of the log likelihood function is close to the true expected value of the log likelihood function, then we would expect that $\widehat{\boldsymbol{\theta}}_n$ will be close to the maximum of the expected likelihood (as $n$ increases without bound)

However, we cannot do that!

# What is the problem?

- The problem is that the argument of the $\arg\max_{\boldsymbol{\theta}\in\boldsymbol{\Theta}}(\cdot)$ is a function of $\boldsymbol{\theta}$, not a real vector:
  - The concept of convergence in probability was defined for **sequence of random variables**
- Therefore, we need to define what we mean by the probability limit of **sequence of random functions**, as opposed to a sequence of random variables:

$$\text{Convergence for sequence of random variables} \implies X_n = X_n(\omega), \omega \in \Omega$$
$$\text{Convergence for sequence of random function} \implies Q_n = Q_n(\omega, \boldsymbol{\theta}), \omega \in \Omega$$

# Example

## Example

In ML estimation, the log-likelihood is a function of the sample data (a random vector that depends on $\omega$) and of a parameter $\boldsymbol{\theta}$. By increasing the sample size, we obtain a sequence of log-likelihoods that depend on $\omega$ and $\boldsymbol{\theta}$.

# Consistency

How is the distance between two functions over a set containing an infinite number of possible comparisons at different values of $\boldsymbol{\theta}$ measured?

- IOW, since we are dealing with convergence on a **function space** we need to define when two functions are close to one another.
- To reduce the infinite dimensional character of a function to a one-dimensional concept of convergence, we take the supremum of the absolute difference of the function values over all $\boldsymbol{\theta}$ in $\boldsymbol{\Theta}$

# Uniform Convergence in Probability

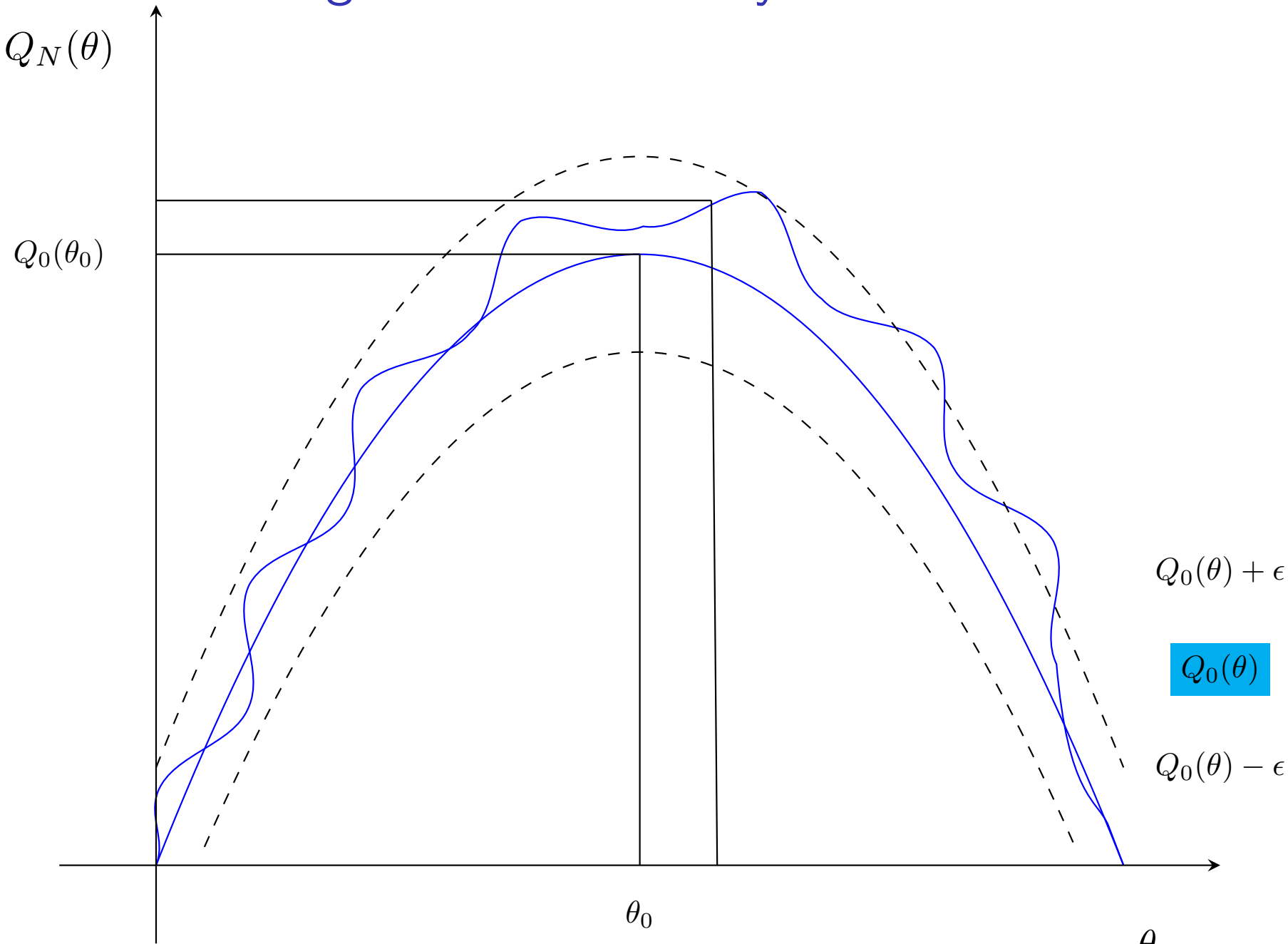
## Definition (Uniform Convergence in Probability)

The sequence of real-valued functions $\{Q_n(\boldsymbol{\theta})\}$ converges uniformly in probability to the limit function $Q_0(\boldsymbol{\theta})$ if $\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} |Q_n(\boldsymbol{\theta}) - Q_0(\boldsymbol{\theta})| \xrightarrow{p} 0$. We will say that $Q_n(\boldsymbol{\theta}) \xrightarrow{p} Q_0(\boldsymbol{\theta})$ **uniformly**.
Another way to express uniform convergence in probability is:

$$\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} |Q_n(\boldsymbol{\theta}) - Q_0(\boldsymbol{\theta})| = o_p(1)$$

IOW, instead of requiring that the distance $|Q_n(\boldsymbol{\theta}) - Q_0(\boldsymbol{\theta})|$ converge in probability to 0 for each $\boldsymbol{\theta}$, we require convergence of $\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} |Q_n(\boldsymbol{\theta}) - Q_0(\boldsymbol{\theta})|$, which is the maximum distance that can be found by ranging over the space parameters.

# Uniform Convergence in Probability



# Uniform Convergence in Probability

Extending the concept to random vectors is straightforward. Now suppose that $\{Q_n(\boldsymbol{\theta})\}$ is a sequence of $K \times 1$ random vectors that depend both on the data and on the parameter $\boldsymbol{\theta} \in \boldsymbol{\Theta}$. This sequence of **random vectors** is uniformly convergent in probability to $Q_0(\boldsymbol{\theta})$ if and only if

$$\sup_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \|Q_n(\boldsymbol{\theta}) - Q_0(\boldsymbol{\theta})\| = o_p(1)$$

where $\|Q_n(\boldsymbol{\theta}) - Q_0(\boldsymbol{\theta})\|$ denotes the Euclidean norm of the vector $Q_n(\boldsymbol{\theta}) - Q_0(\boldsymbol{\theta})$. By taking the supremum over $\boldsymbol{\theta}$ we obtain another random quantity that does not depend on $\boldsymbol{\theta}$.

# Pointwise Convergence in probability

## Definition (Pointwise Convergence in probability)

The sequence of real-valued functions $\{Q_n(\boldsymbol{\theta})\}$ converges pointwise in probability if and only if $|Q_n(\boldsymbol{\theta}) - Q_0(\boldsymbol{\theta})| \xrightarrow{p} 0$ **for each** $\boldsymbol{\theta} \in \boldsymbol{\Theta}$

Uniform convergence is stronger than pointwise convergence.

# Uniform LLN

Now we present the **uniform LLN** to study sequences of random functions which is analogous to the Chebychev's LLN for averages of random variables.

## Theorem (Uniform LLN)

*Suppose that* $Q(\boldsymbol{\theta}, U)$ *is continuous function over* $\boldsymbol{\theta} \in \boldsymbol{\Theta}$*, a closed and bounded subset of* $\mathbb{R}^p$*, and that* $\{U_n\}$ *is a sequence of i.i.d. random variables with cdf* $F_U(u)$*. If* $\mathbb{E}\left[\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \|Q(\boldsymbol{\theta}; U)\|\right]$ *exits, then*

1. $\mathbb{E}\left[Q(\boldsymbol{\theta}; U)\right]$ *is continuous over* $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ *and,*
2. $\frac{1}{n}\sum_{i=1}^{n} Q(\boldsymbol{\theta}; u_i) \xrightarrow{p} \mathbb{E}\left[Q(\boldsymbol{\theta}; U)\right]$ *uniformly.*

*Or as Newey and McFadden (1994) (Lemma 2.4 state):*

$$\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \left\| \frac{1}{n}\sum_{i=1}^{n} Q(\boldsymbol{\theta}; u_i) - \mathbb{E}\left[Q(\boldsymbol{\theta}; U)\right] \right\| \xrightarrow{p} \mathbf{0}$$

# Uniform LLN

The following Theorem makes the connection between the uniform convergence of $\frac{1}{n}\sum_{i=1}^{n} Q(\boldsymbol{\theta}; u_i)$ to $\mathbb{E}\left[Q(\boldsymbol{\theta}; U)\right]$ and the convergence of $\widehat{\boldsymbol{\theta}}_n$ to $\boldsymbol{\theta}_0$ using the assumption of compact parameter space.

# Consistency

## Theorem (Consistency of Maxima with Compact Parameter Space)

*Suppose that:*

1. *(compact parameter space)* $\boldsymbol{\Theta} \subset \mathbb{R}^p$ *is a closed and bounded parameter space,*
2. *(uniform convergence)* $Q_n(\boldsymbol{\theta})$ *is a sequence of function that convergence in probability uniformly to a function* $Q_0(\boldsymbol{\theta})$ *on* $\boldsymbol{\Theta}$*,*
3. *(continuity)* $Q_n(\boldsymbol{\theta})$ *is continuous in* $\boldsymbol{\theta}$ *for any data* $(\mathbf{w}_1, ..., \mathbf{w}_n)$*,*
4. *(identification)* $Q_0(\boldsymbol{\theta})$ *is uniquely maximized at* $\boldsymbol{\theta}_0 \in \boldsymbol{\Theta}$

*then* $\widehat{\boldsymbol{\theta}}_n \equiv \underset{\boldsymbol{\theta} \in \boldsymbol{\Theta}}{\arg\max}\, Q_n(\boldsymbol{\theta})$ *converges in probability to* $\boldsymbol{\theta}_0$*.*

# Consistency

**Intuition:**

If $Q_n(\boldsymbol{\theta})$ converges uniformly to $Q_0(\boldsymbol{\theta})$, then the characteristics of $Q_n(\boldsymbol{\theta})$ will be close to the characteristics of $Q_0(\boldsymbol{\theta})$ as $n \to \infty$. One particular characteristic is the point $\boldsymbol{\theta}_0$ where $Q_0(\boldsymbol{\theta})$ is uniquely maximized. Then, it is expected that the maximizer of $Q_n(\boldsymbol{\theta})$, $\widehat{\boldsymbol{\theta}}$, will be close to the maximizer of $Q_0(\boldsymbol{\theta})$.

# Consistency

## Theorem (Consistency of Maxima without Compactness)

*Suppose that:*

1. *(interior)* $\boldsymbol{\theta}_0$ *is an element of the interior of a convex parameter space* $\boldsymbol{\Theta}$,
2. *(pointwise convergence)* $Q_n(\boldsymbol{\theta})$ *converges in probability to* $Q_0(\boldsymbol{\theta})$ *for all* $\boldsymbol{\theta} \in \boldsymbol{\Theta}$,
3. *(concavity)* $Q_n(\boldsymbol{\theta})$ *is concave over the parameter space for any data* $(\mathbf{w}_1, ..., \mathbf{w}_n)$,
4. *(identification)* $Q_0(\boldsymbol{\theta})$ *is uniquely maximized at* $\boldsymbol{\theta}_0 \in \boldsymbol{\Theta}$

*then, as* $n \to \infty$, $\widehat{\boldsymbol{\theta}}_n$ *exists with probability approaching 1 and* $\widehat{\boldsymbol{\theta}}_n \xrightarrow{p} \boldsymbol{\theta}_0$.

# Consistency

## Theorem (Consistency of conditional ML with compact parameter)

*Let* $\{y_i, \mathbf{x}_i\}$ *be i.i.d with conditional density* $f(y_i|\mathbf{x}_i;\boldsymbol{\theta}_0)$ *and let* $\widehat{\boldsymbol{\theta}}$ *be the conditional ML estimator, which maximizes the average log conditional likelihood:*

$$\widehat{\boldsymbol{\theta}}_n = \arg\max_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \frac{1}{n}\sum_{i=1}^{n} \log f(y_i|\mathbf{x}_i;\boldsymbol{\theta})$$

*Suppose the model is correctly specified so that* $\boldsymbol{\theta}_0$ *is in* $\boldsymbol{\Theta}$*. Suppose that*

1. *(Compactness) the parameter space* $\boldsymbol{\Theta}$ *is compact subset of* $\mathbb{R}^K$*,*
2. $f(y_i|\mathbf{x}_i;\boldsymbol{\theta})$ *is continuous in* $\boldsymbol{\theta}$ *for all* $(y_i, \mathbf{x}_i)$ *(Here, note the Weierstrass's theorem),*
3. $f(y_i|\mathbf{x}_i;\boldsymbol{\theta})$ *is measurable in* $(y_i, \mathbf{x}_i)$ *for all* $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ *(so* $\widehat{\boldsymbol{\theta}}$ *is well-defined random variable),*
4. *(identification)* $\Pr\left[f(y_i|\mathbf{x}_i;\boldsymbol{\theta}) \neq f(y_i|\mathbf{x}_i;\boldsymbol{\theta}_0)\right] > 0$ *for all* $\boldsymbol{\theta} \neq \boldsymbol{\theta}_0$ *in* $\boldsymbol{\Theta}$*,*
5. *(dominance)* $\mathbb{E}\left[\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} |\log f(y_i|\mathbf{x}_i;\boldsymbol{\theta})|\right] < \infty$ *(note: the expectation is over* $y_i$ *and* $\mathbf{x}_i$*)*

*Then* $\widehat{\boldsymbol{\theta}} \xrightarrow{p} \boldsymbol{\theta}_0$

## Sketch of Proof.

We would like to apply **Consistency of Maxima with Compact Parameter Space** Theorem. In this case, let $Q(\boldsymbol{\theta}; U) = \log f(\mathbf{y}|\mathbf{x}; \boldsymbol{\theta})$. Now we verify that the condition of the theorem are met:

- $f(y_i|\mathbf{x}_i; \boldsymbol{\theta})$ is continuous,
- Compactness states that $\boldsymbol{\Theta}$ is a closed and bounded subset of $E^K$,
- $(y_i, \mathbf{x}_i)$ are i.i.d with conditional density $f(y_i|\mathbf{x}_i; \boldsymbol{\theta}_0)$,
- Dominance I states that $\mathbb{E}\left[\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} |\log f(y_i|\mathbf{x}_i; \boldsymbol{\theta})|\right]$ exists.

Therefore, $\mathbb{E}\left[\log f(y_i|\mathbf{x}_i; \boldsymbol{\theta})\right]$ is continuous and

$$\frac{1}{n}\sum_{i=1}^{n} \log f(y_i|\mathbf{x}_i; \boldsymbol{\theta}) \xrightarrow{p} \mathbb{E}\left[\log f(y_i|\mathbf{x}_i; \boldsymbol{\theta})\right] \tag{7}$$

uniformly. Let $Q_n(\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^{n} \log f(y_i|\mathbf{x}_i; \boldsymbol{\theta})$ and $Q_0(\boldsymbol{\theta}) = \mathbb{E}\left[\log f(y_i|\mathbf{x}_i; \boldsymbol{\theta})\right]$. Under the additional assumption of Likelihood Identification, we can invoke the strict expected log-likelihood inequality: $\boldsymbol{\theta} \neq \boldsymbol{\theta}_0$ that $\mathbb{E}\left[\log f(\mathbf{y}|\mathbf{x}; \boldsymbol{\theta})\right] < \mathbb{E}\left[\log f(\mathbf{y}|\mathbf{x}; \boldsymbol{\theta}_0)\right]$. This implies that $Q_0(\boldsymbol{\theta})$ is uniquely maximized at $\boldsymbol{\theta}_0$. Therefore

$$\widehat{\boldsymbol{\theta}}_n = \arg\max_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \frac{1}{n}\sum_{i=1}^{n} \log f(y_i|\mathbf{x}_i; \boldsymbol{\theta}) \xrightarrow{p} \boldsymbol{\theta}_0$$

□

# Consistency

## Theorem (Consistency of conditional ML without Compactness)

*Let* $\{y_i, \mathbf{x}_i\}$ *be i.i.d with conditional density* $f(y_i|\mathbf{x}_i;\boldsymbol{\theta}_0)$ *and let* $\widehat{\boldsymbol{\theta}}$ *be the conditional ML estimator, which maximizes the average log conditional likelihood:*

$$\widehat{\boldsymbol{\theta}}_n = \arg\max_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \frac{1}{n}\sum_{i=1}^{n} \log f(y_i|\mathbf{x}_i;\boldsymbol{\theta})$$

*Suppose the model is correctly specified so that* $\boldsymbol{\theta}_0$ *is in* $\boldsymbol{\Theta}$*. Suppose that*

1. *the true parameter vector* $\boldsymbol{\theta}_0$ *is an element of the interior of convex parameter space* $\boldsymbol{\Theta} \subset \mathbb{R}^K$,
2. $\log f(y_i|\mathbf{x}_i;\boldsymbol{\theta})$ *is concave in* $\boldsymbol{\theta}$ *for all* $(y_i, \mathbf{x}_i)$ ,
3. $\log f(y_i|\mathbf{x}_i;\boldsymbol{\theta})$ *is measurable in* $(y_i, \mathbf{x}_i)$,
4. *(identification)* $\Pr\left[f(y_i|\mathbf{x}_i;\boldsymbol{\theta}) \neq f(y_i|\mathbf{x}_i;\boldsymbol{\theta}_0)\right] > 0$ *for all* $\boldsymbol{\theta} \neq \boldsymbol{\theta}_0$ *in* $\boldsymbol{\Theta}$,
5. $\mathbb{E}\left[|\log f(y_i|\mathbf{x}_i;\boldsymbol{\theta})|\right] < \infty$ *(i.e.,* $\mathbb{E}\left[\log f(y_i|\mathbf{x}_i;\boldsymbol{\theta})\right]$ *exists and is finite) for all* $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ *(note: the expectation is over* $y_i$ *and* $\mathbf{x}_i$*)*

*Then,* $n \to \infty$, $\widehat{\boldsymbol{\theta}}$ *exists with probability approaching 1 and* $\widehat{\boldsymbol{\theta}}_n \xrightarrow{p} \boldsymbol{\theta}_0$

1 Introduction
- Motivation
- Maximum Likelihood Estimator
- Identification
- The Score Function
- The Information Matrix

2 Asymptotic Properties
- Consistency
- Asymptotic Normality

3 Estimation of Variance

4 Testing
- Intuition
- The Trinity
- Proof, Proof and More Proof

# Asymptotic

## Theorem (Asymptotic Normality of Conditional ML)

*Let* $\mathbf{w} \equiv (y_i, \mathbf{x}_i^\intercal)^\intercal$ *be iid. Suppose the conditions of either Theorem 18 or 19 are satisfied, so that* $\widehat{\boldsymbol{\theta}}_n \xrightarrow{p} \boldsymbol{\theta}_0$. *Suppose, in addition, that:*

1. $\boldsymbol{\theta}_0$ *is in interior of* $\boldsymbol{\Theta}$,
2. $f(y_i|\mathbf{x}_i;\boldsymbol{\theta}_0)$ *is twice continuously differentiable in* $\boldsymbol{\theta}$ *for all* $(y_i, \mathbf{x}_i)$,
3. $\mathbb{E}\left[\mathbf{s}(\mathbf{w}_i;\boldsymbol{\theta}_0)\right] = \mathbf{0}$ *and* $-\mathbb{E}\left[\mathbf{H}(\mathbf{w}_i;\boldsymbol{\theta}_0)\right] = \mathbb{E}\left[\mathbf{s}(\mathbf{w}_i;\boldsymbol{\theta}_0)\mathbf{s}(\mathbf{w}_i;\boldsymbol{\theta}_0)^\intercal\right]$,
4. *(local dominance condition on the Hessian) for some neighborhood* $\mathcal{N}$ *of* $\boldsymbol{\theta}_0$,

$$\mathbb{E}\left[\sup_{\boldsymbol{\theta}\in\mathcal{N}} \left\|\mathbf{H}(\mathbf{w}_i;\boldsymbol{\theta})\right\|\right] < \infty$$

   *so that for any consistent estimator* $\widetilde{\boldsymbol{\theta}}$, $\frac{1}{n}\sum_{i=1}^n \mathbf{H}(\mathbf{w}_i;\widetilde{\boldsymbol{\theta}}) \xrightarrow{p} \mathbb{E}\left[\mathbf{H}(\mathbf{w}_i;\boldsymbol{\theta}_0)\right]$
5. $\mathbb{E}\left[\mathbf{H}(\mathbf{w}_i;\boldsymbol{\theta}_0)\right]$ *is nonsingular.*

*Then:*

$$\sqrt{n}\left(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\right) \xrightarrow{d} \mathrm{N}(\mathbf{0}, \mathbf{V}), \mathbf{V} = -\left\{\mathbb{E}\left[\mathbf{H}(\mathbf{w}_i;\boldsymbol{\theta}_0)\right]\right\}^{-1} = \left\{\mathbb{E}\left[\mathbf{s}(\mathbf{w}_i;\boldsymbol{\theta}_0)\mathbf{s}(\mathbf{w}_i;\boldsymbol{\theta}_0)^\intercal\right]\right\}^{-1}$$

# Asymptotic

The intuition is the following:

- Since usually we don't have an explicit solution for the estimator, we need to focus on the asymptotic behaviour of the score function.
- Assuming that $\widehat{\boldsymbol{\theta}}_n \xrightarrow{p} \boldsymbol{\theta}$, the behaviour of the score function matters only within an arbitrary small neighbourhood of $\boldsymbol{\theta}_0$.
- ... after all, $\widehat{\boldsymbol{\theta}}_n$ will fall withing such neighborhooods with arbitrary high probability for a large enough sample size...
- ... and within such neighborhood the score function is essentially linear.. (Taylor series expansion)

# Mean Value Theorem

## Theorem (Mean Value Theorem)

*Let* $s : \mathbb{R}^K \to \mathbb{R}$ *be defined on an open convex set* $\boldsymbol{\Theta} \subset \mathbb{R}^K$ *such that* $s$ *is continuously differentiable on* $\boldsymbol{\Theta}$ *with gradient* $\nabla s$. *Then for any points* $\boldsymbol{\theta}$ *and* $\boldsymbol{\theta}_0$ *such that* $s(\boldsymbol{\theta}) = s(\boldsymbol{\theta}_0) + \nabla s(\bar{\boldsymbol{\theta}})(\boldsymbol{\theta} - \boldsymbol{\theta}_0)$

# Asymptotic

## Proof.

The objective function is:

$$Q_n(\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^{n} \ln f(y_i|\mathbf{x}_i;\boldsymbol{\theta})$$

Given that $f(y_i|\mathbf{x}_i;\boldsymbol{\theta}_0)$ is twice continuously differentiable in $\boldsymbol{\theta}$, and given that $\boldsymbol{\theta}_0$ is in interior of $\boldsymbol{\Theta}$, then the maximum likelihood estimator satisfies

$$\frac{\partial \log L(\widehat{\boldsymbol{\theta}})}{\partial \boldsymbol{\theta}} = \mathbf{s}(\mathbf{w};\widehat{\boldsymbol{\theta}}) = \mathbf{0}$$

We need to know about the behavior of the gradient around the true parameter. Expand this set of equations in a Taylor series around the true parameters $\boldsymbol{\theta}_0$. We will use the mean value theorem to truncate the Taylor series at the second term,

$$\underbrace{\frac{\partial \log L(\widehat{\boldsymbol{\theta}})}{\partial \boldsymbol{\theta}}}_{(K\times 1)} = \underbrace{\frac{\partial \log L(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}}}_{(K\times 1)} + \underbrace{\frac{\partial \log L(\overline{\boldsymbol{\theta}})}{\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}^{\intercal}}}_{(K\times K)} \underbrace{\left(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\right)}_{(K\times 1)}$$

$$= \frac{1}{n}\sum_{i=1}^{n} \mathbf{s}(\mathbf{w}_i;\boldsymbol{\theta}_0) + \left[\frac{1}{n}\sum_{i=1}^{n} \mathbf{H}(\mathbf{w}_i;\overline{\boldsymbol{\theta}})\right]\left(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\right)$$

where $\overline{\boldsymbol{\theta}} = \alpha\widehat{\boldsymbol{\theta}} + (1-\alpha)\boldsymbol{\theta}_0$ for some $\alpha \in (0,1)$ □

## Proof.

So,

$$\sqrt{n}\left(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\right) = \left[-\frac{1}{n}\sum_{i=1}^{n}\mathbf{H}(\mathbf{w}_i;\overline{\boldsymbol{\theta}})\right]^{-1}\sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n}\mathbf{s}(\mathbf{w}_i;\boldsymbol{\theta}_0)\right)$$

We know that

$$\widehat{\boldsymbol{\theta}} \xrightarrow{p} \boldsymbol{\theta}_0 \implies \overline{\boldsymbol{\theta}} \xrightarrow{p} \boldsymbol{\theta}_0$$

By uniform LLN, we know that

$$\frac{1}{n}\sum_{i=1}^{n}\mathbf{H}(\mathbf{w}_i;\boldsymbol{\theta}) \xrightarrow{p} \mathbb{E}\left[\mathbf{H}(\mathbf{w}_i;\boldsymbol{\theta})\right] \quad \text{uniformly in } \boldsymbol{\theta} \in \boldsymbol{\Theta}$$

Then, applying our Lemma:

$$\frac{1}{n}\sum_{i=1}^{n}\mathbf{H}(\mathbf{w}_i;\overline{\boldsymbol{\theta}}_n) \xrightarrow{p} \mathbb{E}\left[\mathbf{H}(\mathbf{w}_i;\boldsymbol{\theta}_0)\right]$$

since $\mathbb{E}\left[\mathbf{H}(\mathbf{w}_i;\boldsymbol{\theta}_0)\right]$ exists. Finally, using probability limit continuity and nonsingular information, then:

$$\left[\frac{1}{n}\sum_{i=1}^{n}\mathbf{H}(\mathbf{w}_i;\overline{\boldsymbol{\theta}}_n)\right]^{-1} \xrightarrow{p} \mathbb{E}\left[\mathbf{H}(\mathbf{w}_i;\boldsymbol{\theta}_0)\right]^{-1}$$

□

## Proof.

Since $(y_i, \mathbf{x}_i)$ are i.i.d. $\left( \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \mathbf{s}(\mathbf{w}_i; \boldsymbol{\theta}_0) \right)$ is the sum of variables $\mathbf{s}(\mathbf{w}_i; \boldsymbol{\theta}_0)$. The score identity lemma implies that $\mathbb{E}(\mathbf{s}(\mathbf{w}_i; \boldsymbol{\theta}_0)) = \mathbf{0}$, and the Information Identity implies that

$$\operatorname{Var}\left[\mathbf{s}(\mathbf{w}_i; \boldsymbol{\theta}_0)\right] = \mathbb{E}\left[\mathbf{s}(\mathbf{w}_i; \boldsymbol{\theta}_0)\mathbf{s}(\mathbf{w}_i; \boldsymbol{\theta}_0)^\intercal\right] = -\mathbb{E}\left[\mathbf{H}(\mathbf{w}_i; \boldsymbol{\theta}_0)\right]$$

The Lindberg-Levy CLT therefore implies:

$$\sqrt{n}\left( \frac{1}{n} \sum_{i=1}^{n} \mathbf{s}(\mathbf{w}_i; \boldsymbol{\theta}_0) \right) \xrightarrow{d} \mathrm{N}\left(\mathbf{0}, -\mathbb{E}\left[\mathbf{H}(\mathbf{w}_i; \boldsymbol{\theta}_0)\right]\right)$$

□

## Proof.

Then:

$$\begin{aligned}
\sqrt{n}\left(\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0\right) &= \left[-\frac{1}{n}\sum_{i=1}^{n}\mathbf{H}(\mathbf{w}_i;\overline{\boldsymbol{\theta}})\right]^{-1}\sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n}\mathbf{s}(\mathbf{w}_i;\boldsymbol{\theta}_0)\right) \\
&\xrightarrow{d} -\mathbb{E}\left[\mathbf{H}(\mathbf{w}_i;\boldsymbol{\theta}_0)\right]^{-1}\mathrm{N}\left(\mathbf{0},-\mathbb{E}\left[\mathbf{H}(\mathbf{w}_i;\boldsymbol{\theta}_0)\right]\right) \\
&= \mathrm{N}\left[\mathbf{0},-\mathbb{E}\left[\mathbf{H}(\mathbf{w}_i;\boldsymbol{\theta}_0)\right]^{-1}\mathbb{E}\left[\mathbf{H}(\mathbf{w}_i;\boldsymbol{\theta}_0)\right]\mathbb{E}\left[\mathbf{H}(\mathbf{w}_i;\boldsymbol{\theta}_0)\right]^{-1}\right] \\
&= \mathrm{N}\left[\mathbf{0},-\mathbb{E}\left[\mathbf{H}(\mathbf{w}_i;\boldsymbol{\theta}_0)\right]^{-1}\right] \\
&= \mathrm{N}\left[\mathbf{0},\left[\mathbf{I}(\mathbf{w}_i;\boldsymbol{\theta}_0)\right]^{-1}\right]
\end{aligned}$$

□

# Variance Estimation

For large but finite samples, we can therefore write the approximate distribution of $\widehat{\boldsymbol{\theta}}_n$ as

$$\widehat{\boldsymbol{\theta}} \overset{a}{\sim} \mathrm{N}\left[\boldsymbol{\theta}_0, n \cdot [\mathbf{I}(\boldsymbol{\theta}_0)]^{-1}\right]$$

we have three potential estimators of $\mathbf{I}(\boldsymbol{\theta}_0)$:

- The empirical mean of minus the Hessian,

$$\widehat{\mathbf{V}}^1 = \left(\frac{1}{n}\sum_{i=1}^{n} -\mathbf{H}(\mathbf{w}_i, \widehat{\boldsymbol{\theta}})\right)^{-1}$$

- The empirical variance of the score:

$$\widehat{\mathbf{V}}^2 = \left(\frac{1}{n}\sum_{i=1}^{n} \mathbf{s}(\mathbf{w}_i, \widehat{\boldsymbol{\theta}})\mathbf{s}(\mathbf{w}_i; \widehat{\boldsymbol{\theta}})^\intercal\right)^{-1}$$

- Minus the expected Hessian evaluated at $\widehat{\boldsymbol{\theta}}$:

$$\widehat{\mathbf{V}}^3 = \left(\frac{1}{n}\sum_{i=1}^{n} -\mathbb{E}\left[\mathbf{H}(\mathbf{w}_i, \widehat{\boldsymbol{\theta}})\right]\right)^{-1}$$

# Proof of Consistency

- Evaluated at a $\boldsymbol{\theta} \in \boldsymbol{\Theta}$, each estimator converges in probability uniformly to its expectation.
- Because $\widehat{\boldsymbol{\theta}}_n \xrightarrow{p} \boldsymbol{\theta}_0$, evaluated at $\widehat{\boldsymbol{\theta}}_n$ each estimator converges in probability to $\mathbf{I}(\boldsymbol{\theta}_0)$.
- Because matrix inversion is a continuous transformation, the inverse of each matrix is also a consistent estimator for the variance matrix of the asymptotic distribution of $\sqrt{n}(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0)$

1. Introduction
   - Motivation
   - Maximum Likelihood Estimator
   - Identification
   - The Score Function
   - The Information Matrix
2. Asymptotic Properties
   - Consistency
   - Asymptotic Normality
3. Estimation of Variance
4. Testing
   - Intuition
   - The Trinity
   - Proof, Proof and More Proof

# Hypothesis Testing

- ML estimator are distributed asymptotically normally:
  - As the sample size increases, the sampling distribution of an ML estimator becomes approximately normal

$$\widehat{\boldsymbol{\beta}} \overset{a}{\sim} \mathrm{N}\left(\boldsymbol{\beta}, \mathbf{V}_{\widehat{\boldsymbol{\beta}}}\right)$$

where, for three coefficients:

$$\mathbf{V}_{\widehat{\boldsymbol{\beta}}} = \mathrm{Var}\begin{pmatrix}\widehat{\beta}_0 \\ \widehat{\beta}_1 \\ \widehat{\beta}_2\end{pmatrix} = \begin{pmatrix}\sigma^2_{\widehat{\beta}_0} & \sigma_{\widehat{\beta}_0,\widehat{\beta}_1} & \sigma_{\widehat{\beta}_0,\widehat{\beta}_2} \\ \sigma_{\widehat{\beta}_1,\widehat{\beta}_0} & \sigma^2_{\widehat{\beta}_1} & \sigma_{\widehat{\beta}_1,\widehat{\beta}_2} \\ \sigma_{\widehat{\beta}_2,\widehat{\beta}_0} & \sigma_{\widehat{\beta}_2,\widehat{\beta}_1} & \sigma^2_{\widehat{\beta}_2}\end{pmatrix}$$

# Hypothesis Testing

Consider the simple hypothesis

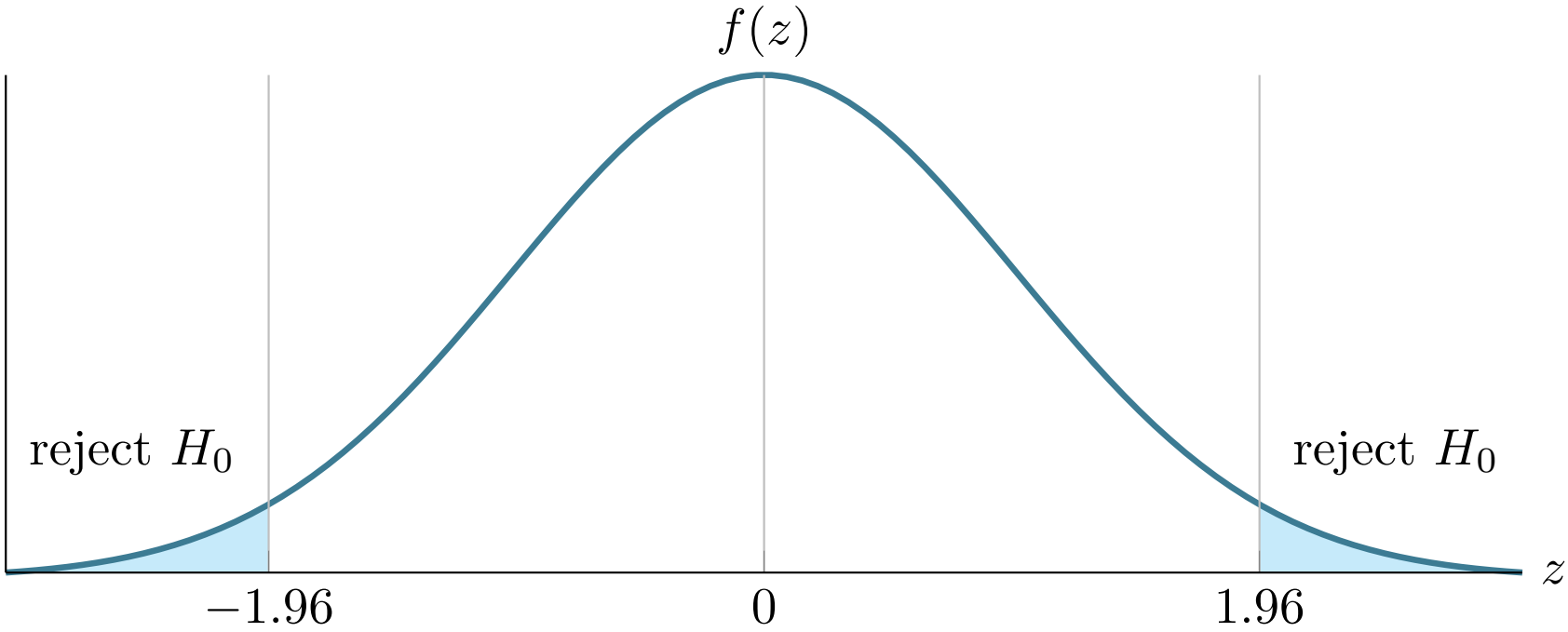
$$H_0 : \beta_k = \beta^*$$

where $\beta^*$ is the hypothesized value, often equal to 0. Since $\sigma^2_{\widehat{\beta}_k}$ is unknown, it must be estimated, which results in the test:

$$z = \frac{\widehat{\beta}_k - \beta^*}{\widehat{\sigma}^2_{\widehat{\beta}_k}}$$

Under the assumptions justifying ML, if $H_0$ is true, then $z$ is distributed approximately normally with mean of 0 and variance of 1 for large samples.

# Testing



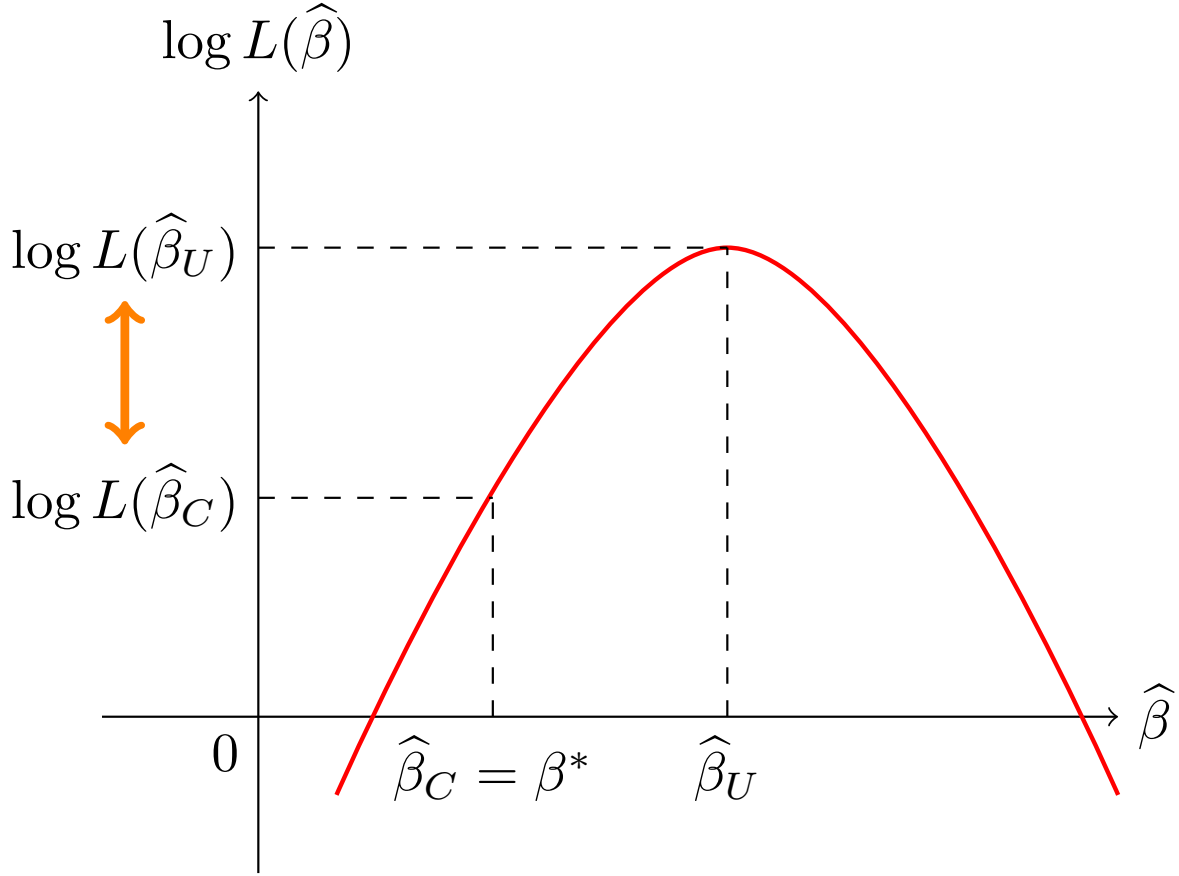
1. Introduction
   - Motivation
   - Maximum Likelihood Estimator
   - Identification
   - The Score Function
   - The Information Matrix
2. Asymptotic Properties
   - Consistency
   - Asymptotic Normality
3. Estimation of Variance
4. Testing
   - Intuition
   - The Trinity
   - Proof, Proof and More Proof

# The Trinity

For more complex hypotheses we can use the Wald, likelihood ratio (LR), or Lagrange multiplier (LM) test. These test can be though of as a comparison between the estimates obtained after the constrains implied by the hypothesis have been imposed to the estimates obtained without the constraints.

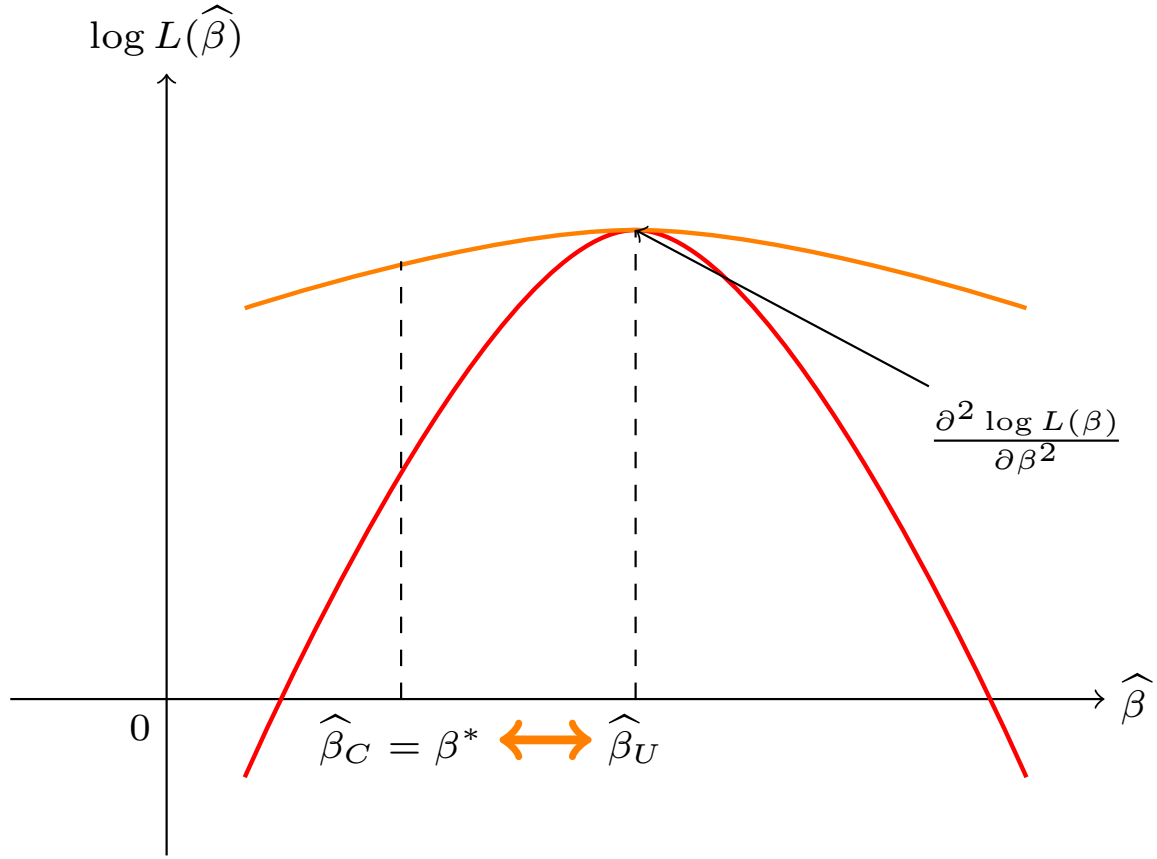
# The Trinity: LR Test

- Log-likelihood function is the red solid line;
- $\widehat{\beta}_U$: unconstrained estimator.
- The $H_0 : \beta = \beta^*$ imposes the constraint $\beta = \beta^*$, so that the constrained estimate is $\widehat{\beta}_C = \beta^*$.
- Unless $\widehat{\beta}_U = \beta^*$, $\ln L(\widehat{\beta}_C) \leq \ln L(\widehat{\beta}_U)$.
- If the constraint **significantly** reduces the likelihood, then the null hypothesis is rejected



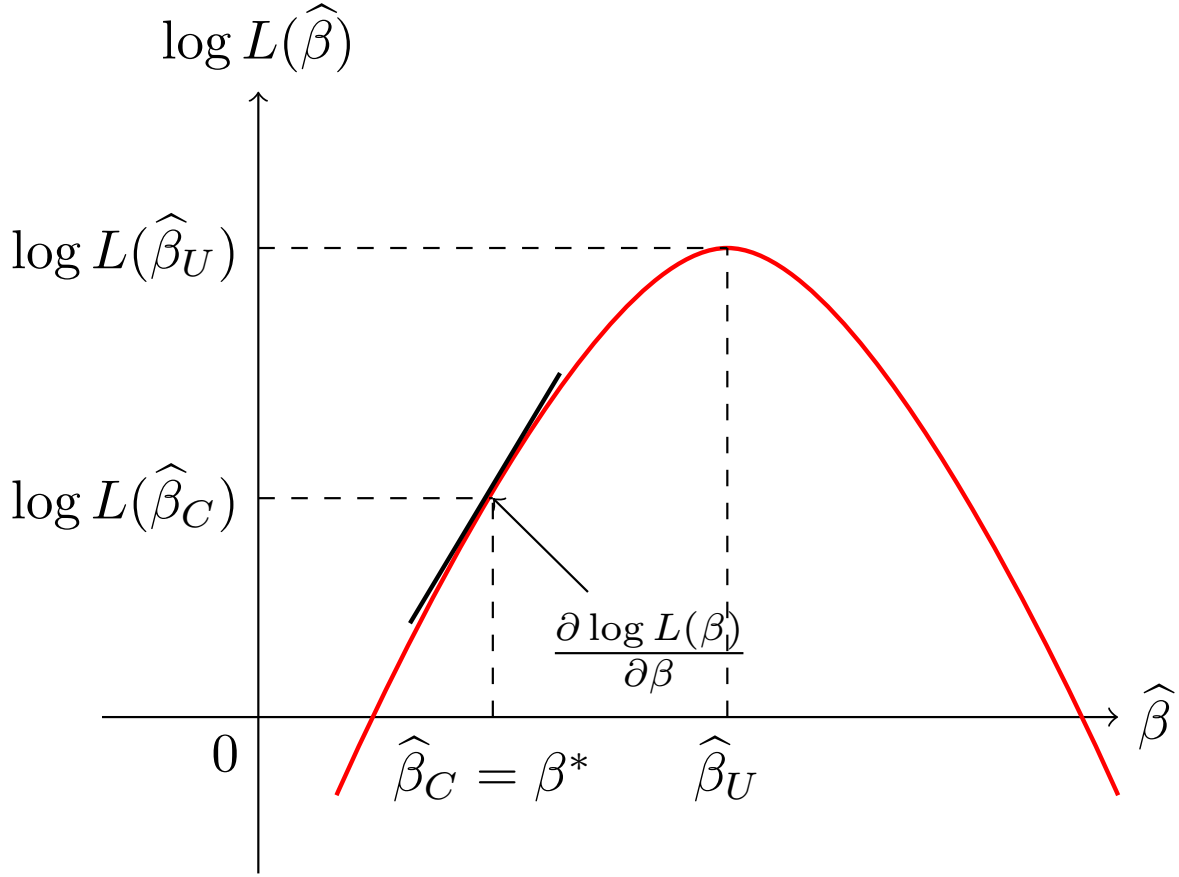
# The Trinity: Wald Test

- The Wald test estimate the model without constraints, and assesses the constraint by considering 2 things:
- 1. It measures the distance $\widehat{\beta}_U - \widehat{\beta}_C = \widehat{\beta}_U - \beta^*$.
- The larger the distance, the less likely it is that the constraint is true. .
- 2. The distance is weighted by the curvature of the log likelihood function $\partial^2 \log L(\beta)/\partial\beta^2$
- The larger the second derivative, the faster the curve is changing.
- The LL function (dashed line) in nearly flat, so the second derivative evaluated at $\widehat{\beta}_U$ is relatively small.
- When the second derivative is small, the distance $\widehat{\beta}_U$ and $\widehat{\beta}_C$ is minor relative to the sampling variation.
- The second LL function has a larger second derivative.
- With a larger second derivative, the same distance between $\widehat{\beta}_U$ and $\widehat{\beta}_C$ might be significant.



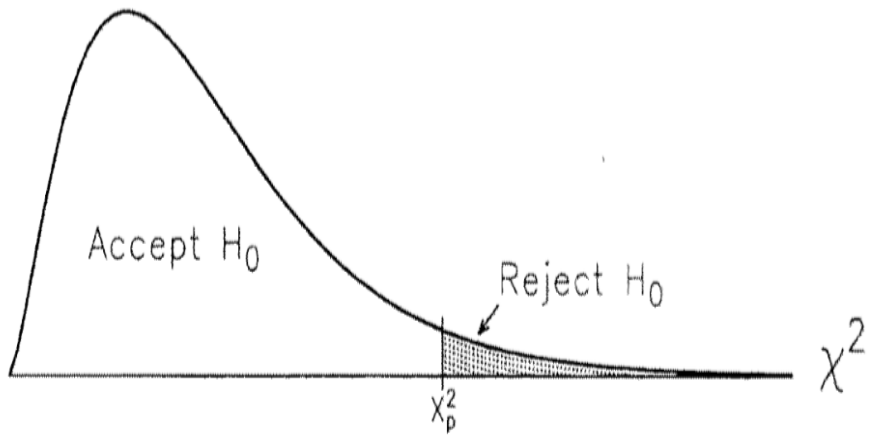
# The Trinity: LM (Score) Test

- It only estimate the constrained model.
- It assesses the slope of the log likelihood function at the constraint.
- If $H_0$ is true, the slope (score) at the constraint should be close to 0.
- As with the Wald test, the curvature of the log likelihood function at the constraint is used to assess the significance of a nonzero slope.



# The Trinity

- When $H_0$ is true, the Wald, LR, and LM tests are asymptotically equivalent.
- As $n \to \infty$, the sampling distributions of the three test $\xrightarrow{d} \chi^2_r$, where $r$ is the number of constraints being tested.




They are similar only when $n \to \infty$. In small samples this is not necessary true.

# Test Statistics

Assume that we have $r$ $(r < K)$ nonlinear restrictions (which includes linear restriction as special case)

$$\underbrace{\mathbf{r}(\boldsymbol{\theta}_0)}_{r\times 1} = \underbrace{\mathbf{0}}_{r\times 1}$$

The hypotheses are:

$$H_0 : \mathbf{r}(\boldsymbol{\theta}_0) = \mathbf{0}$$
$$H_1 : \mathbf{r}(\boldsymbol{\theta}_0) \neq \mathbf{0}$$

Also, denote $\widehat{\boldsymbol{\theta}}$ as the unconstrained estimator and $\widetilde{\boldsymbol{\theta}}$ as constrained estimator. So:

$$\widetilde{\boldsymbol{\theta}}_{ML} = \arg\max_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \left\{ \frac{1}{n} \sum_{i=1}^{m} \log f(y_i|\mathbf{x}_i;\boldsymbol{\theta}) \right\} \quad s.t \quad \mathbf{r}(\boldsymbol{\theta}_0) = \mathbf{0}$$

# Test Statistics

Then, Wald, LM and LR statistics are defined as:

$$W = \underbrace{\mathbf{r}(\widehat{\boldsymbol{\theta}})^{\intercal}}_{1\times r}\left[\underbrace{\mathbf{R}(\widehat{\boldsymbol{\theta}})}_{r\times K}\underbrace{\left(\frac{1}{n}\cdot\widehat{\mathbf{V}}\right)}_{K\times K}\underbrace{\mathbf{R}(\widehat{\boldsymbol{\theta}})^{\intercal}}_{K\times r}\right]^{-1}\underbrace{\mathbf{r}(\widehat{\boldsymbol{\theta}})}_{r\times 1}$$

$$LM = \underbrace{\mathbf{s}(\widetilde{\boldsymbol{\theta}})^{\intercal}}_{1\times K}\underbrace{\left(\frac{1}{n}\cdot\widetilde{\mathbf{V}}\right)}_{K\times K}\underbrace{\mathbf{s}(\widetilde{\boldsymbol{\theta}})}_{K\times 1}$$

$$LR = 2\cdot n\left(\underbrace{\frac{1}{n}\sum_{i=1}^{n}\ln f(y_i|\mathbf{x}_i;\widehat{\boldsymbol{\theta}})}_{1\times 1} - \underbrace{\frac{1}{n}\sum_{i=1}^{n}\ln f(y_i|\mathbf{x}_i;\widetilde{\boldsymbol{\theta}})}_{1\times 1}\right)$$

# Test Statistics

where:

$$\underbrace{\mathbf{R}(\boldsymbol{\theta}_0)}_{r \times K} = \frac{\partial \mathbf{r}(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}^\intercal} \quad \text{Jacobian of } \mathbf{r}(\boldsymbol{\theta}_0)$$

$$\underbrace{\mathbf{V}}_{K \times K} = \mathbf{I}^{-1} = -\frac{1}{n} \sum_{i=1} \frac{\partial}{\partial \boldsymbol{\theta} \boldsymbol{\theta}^\intercal} \ln f(y_i | \mathbf{x}_i; \widehat{\boldsymbol{\theta}})$$

1. Introduction
   - Motivation
   - Maximum Likelihood Estimator
   - Identification
   - The Score Function
   - The Information Matrix
2. Asymptotic Properties
   - Consistency
   - Asymptotic Normality
3. Estimation of Variance
4. Testing
   - Intuition
   - The Trinity
   - Proof, Proof and More Proof

## Proof: Wald Statistic.

Write W as:

$$\mathbf{W} = \mathbf{c}_n^\intercal \mathbf{Z}_n^{-1} \mathbf{c}_n, \quad \mathbf{c}_n \equiv \sqrt{n}\mathbf{r}(\widehat{\boldsymbol{\theta}}), \quad \mathbf{Z}_n \equiv \mathbf{R}(\widehat{\boldsymbol{\theta}})\widehat{\mathbf{V}}\mathbf{R}(\widehat{\boldsymbol{\theta}})^\intercal \tag{8}$$

First, we will show that $\mathbf{c}_n \xrightarrow{d} \mathrm{N}(\mathbf{0}, \mathbf{V_f})$. Applying the MVT (21) (to truncate the Taylor series) to $\mathbf{r}(\widehat{\boldsymbol{\theta}})$ around $\boldsymbol{\theta}_0$:

$$\begin{aligned}
\mathbf{r}(\widehat{\boldsymbol{\theta}}) = & \underbrace{\mathbf{r}(\boldsymbol{\theta}_0)}_{=\mathbf{0} \text{ under } H_0} + \mathbf{R}(\bar{\boldsymbol{\theta}})(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \\
\underbrace{\sqrt{n}\mathbf{r}(\widehat{\boldsymbol{\theta}})}_{r\times 1} = & \underbrace{\mathbf{R}(\bar{\boldsymbol{\theta}})}_{r\times K} \underbrace{\sqrt{n}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)}_{K\times 1} \quad \text{multiplying by } \sqrt{n} \\
= & \mathbf{R}(\bar{\boldsymbol{\theta}})\sqrt{n}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) + \mathbf{R}(\boldsymbol{\theta}_0)\sqrt{n}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) - \mathbf{R}(\boldsymbol{\theta}_0)\sqrt{n}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \\
= & \left(\mathbf{R}(\bar{\boldsymbol{\theta}}) - \mathbf{R}(\boldsymbol{\theta}_0)\right)\sqrt{n}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) + \mathbf{R}(\boldsymbol{\theta}_0)\sqrt{n}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)
\end{aligned}$$

where $\bar{\boldsymbol{\theta}} = \alpha\widehat{\boldsymbol{\theta}} + (1-\alpha)\boldsymbol{\theta}_0$

## Proof: Wald Statistic.

Note that

$$\because \widehat{\boldsymbol{\theta}} \xrightarrow{p} \boldsymbol{\theta}_0 \implies \bar{\boldsymbol{\theta}} \xrightarrow{p} \boldsymbol{\theta}_0$$

$$\because F(\cdot) \quad \text{is continuous} \quad \implies F(\bar{\boldsymbol{\theta}}) \xrightarrow{p} F(\boldsymbol{\theta}_0)$$

Note that $\sqrt{n}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \xrightarrow{d}$ some random variable, then we can write:

$$\sqrt{n}\mathbf{r}(\widehat{\boldsymbol{\theta}}) = \underbrace{\left(\mathbf{R}(\bar{\boldsymbol{\theta}}) - \mathbf{R}(\boldsymbol{\theta}_0)\right)\sqrt{n}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)}_{\xrightarrow{p}\mathbf{0}} + \mathbf{R}(\boldsymbol{\theta}_0)\sqrt{n}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)$$

$$\sqrt{n}\mathbf{r}(\widehat{\boldsymbol{\theta}}) = \mathbf{R}(\boldsymbol{\theta}_0)\sqrt{n}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) + o_p(1)$$

Furthermore, we know that:

$$\sqrt{n}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) = \mathbb{E}\left[\mathbf{H}(\mathbf{w}_i, \theta_0)\right]^{-1} \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \mathbf{s}(\mathbf{w}_i, \boldsymbol{\theta}_0) + o_p(1) \tag{9}$$

□

## Proof: Wald Statistic.

Then:

$$\begin{aligned}\sqrt{n}\mathbf{r}(\widehat{\boldsymbol{\theta}}) &= \mathbf{R}(\boldsymbol{\theta}_0)\sqrt{n}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) + o_p(1) \\ &= \mathbf{R}(\boldsymbol{\theta}_0)\left[\mathbb{E}\left[\mathbf{H}(\mathbf{w}_i, \theta_0)\right]^{-1} \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \mathbf{s}(\mathbf{w}_i, \boldsymbol{\theta}_0) + o_p(1)\right] + o_p(1) \\ &= \mathbf{R}(\boldsymbol{\theta}_0)\mathbb{E}\left[\mathbf{H}(\mathbf{w}_i, \theta_0)\right]^{-1} \underbrace{\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \mathbf{s}(\mathbf{w}_i, \boldsymbol{\theta}_0)}_{\xrightarrow{d} \mathrm{N}(\mathbf{0}, -\mathbb{E}[\mathbf{H}(\mathbf{w}_i; \boldsymbol{\theta}_0)])} + \underbrace{o_p(1)}_{o_p(1)+o_p(1)}\end{aligned}$$

Then:

$$\sqrt{n}\mathbf{r}(\widehat{\boldsymbol{\theta}}) \xrightarrow{d} \mathrm{N}\left(\mathbf{0}, -\mathbf{R}(\boldsymbol{\theta}_0)\mathbb{E}\left[\mathbf{H}(\mathbf{w}_i, \theta_0)\right]^{-1} \mathbb{E}\left[\mathbf{H}(\mathbf{w}_i; \boldsymbol{\theta}_0)\right] \mathbb{E}\left[\mathbf{H}(\mathbf{w}_i, \theta_0)\right]^{-1} \mathbf{R}(\boldsymbol{\theta}_0)^{\intercal}\right)$$

$$\xrightarrow{d} \mathrm{N}\left(\mathbf{0}, \mathbf{R}(\boldsymbol{\theta}_0) \underbrace{\left[-\mathbb{E}\left[\mathbf{H}(\mathbf{w}_i, \theta_0)\right]^{-1}\right]}_{\mathbf{V}} \mathbf{R}(\boldsymbol{\theta}_0)'\right)$$

□

## Proof: Wald Statistic.

It follows that:

$$\mathbf{W} = \mathbf{c}_n^\intercal \mathbf{Z}_n^{-1} \mathbf{c}_n = \sqrt{n}\mathbf{r}(\widehat{\boldsymbol{\theta}}) \left[\mathbf{R}(\boldsymbol{\theta}_0)\mathbf{V}_0\mathbf{R}(\boldsymbol{\theta}_0)^\intercal\right]^{-1} \sqrt{n}\mathbf{r}(\widehat{\boldsymbol{\theta}}) \xrightarrow{d} \chi^2(\#r)$$

If we have consistent estimators of $\mathbf{R}(\boldsymbol{\theta}_0)$ and $\mathbf{V}$, then limit results for continuous functions imply that:

$$\sqrt{n}\mathbf{r}(\widehat{\boldsymbol{\theta}}) \left[\widehat{\mathbf{R}}(\boldsymbol{\theta}_0)\widehat{\mathbf{V}}\widehat{\mathbf{R}}(\boldsymbol{\theta}_0)^\intercal\right]^{-1} \sqrt{n}\mathbf{r}(\widehat{\boldsymbol{\theta}}) \xrightarrow{d} \chi^2(\#r)$$

□

# Preliminaries to the next two statistics

The Lagrangian for the constraint is

$$\mathcal{L} = \frac{1}{n}\sum_{i=1}^{n} \log f(y_i|\mathbf{x}_i, \boldsymbol{\theta}) - \underbrace{\boldsymbol{\lambda}^\intercal}_{1\times r}\underbrace{\mathbf{r}(\boldsymbol{\theta})}_{r\times 1}$$

Then FOC:

$$\begin{aligned} \sqrt{n}\mathbf{s}\left(\widetilde{\boldsymbol{\theta}}\right) + \sqrt{n}\mathbf{R}(\widetilde{\boldsymbol{\theta}})^\intercal\boldsymbol{\lambda} &= \mathbf{0} \\ \sqrt{n}\mathbf{r}(\widetilde{\boldsymbol{\theta}}) &= \mathbf{0} \end{aligned} \tag{10}$$

Following a similar reasoning as in previous proof:

$$\sqrt{n}\mathbf{r}(\widetilde{\boldsymbol{\theta}}) = \mathbf{R}(\boldsymbol{\theta}_0)\sqrt{n}(\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) + o_p(1) = O_p(1)$$

# Preliminaries to the next two statistics

A Taylor expansion of $\mathbf{s}(\widetilde{\boldsymbol{\theta}})$ yields:

$$\begin{aligned}
\frac{\partial \log L(\widetilde{\boldsymbol{\theta}})}{\partial \boldsymbol{\theta}} &= \frac{\partial \log L(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}} + \frac{\partial^2 \log L(\bar{\boldsymbol{\theta}})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^\intercal} \left(\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\right) \\
&= \frac{\partial \log L(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}} + \frac{\partial^2 \log L(\bar{\boldsymbol{\theta}})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} \left(\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\right) + \frac{\partial^2 \log L(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^\intercal} \left(\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\right) - \frac{\partial^2 \log L(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} \left(\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\right) \\
&= \frac{\partial \log L(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}} + \frac{\partial^2 \log L(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^\intercal} \left(\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\right) + \left(\frac{\partial^2 \log L(\bar{\boldsymbol{\theta}})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^\intercal} - \frac{\partial^2 \log L(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^\intercal}\right) \left(\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\right)
\end{aligned}$$

$$\begin{aligned}
\sqrt{n}\frac{\partial \log L(\widetilde{\boldsymbol{\theta}})}{\partial \boldsymbol{\theta}} &= \sqrt{n}\frac{\partial \log L(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}} + \frac{\partial^2 \log L(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^\intercal} \sqrt{n} \left(\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\right) + \\
&+ \underbrace{\left(\frac{\partial^2 \log L(\bar{\boldsymbol{\theta}})}{n^{-1} \partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^\intercal} - n^{-1} \frac{\partial^2 \log L(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^\intercal}\right)}_{o_p(1)} \underbrace{\sqrt{n} \left(\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\right)}_{O_p(1)} \\
&= \underbrace{\sqrt{n}\frac{\partial \log L(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}}}_{\xrightarrow{d} \mathrm{N}(\mathbf{0}, -\mathbb{E}[\mathbf{H}(\mathbf{w}_i; \boldsymbol{\theta}_0)])} + \underbrace{\frac{\partial^2 \log L(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^\intercal}}_{\mathbb{E}[\mathbf{H}(\mathbf{w}_i; \boldsymbol{\theta}_0)]} \sqrt{n} \left(\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\right) + o_p(1) \\
&= O_p(1) + O_p(1) O_p(1) + o_p(1) \\
&= O_p(1) + O_p(1) + o_p(1) = O_p(1)
\end{aligned}$$

# Preliminaries to the next two statistics

Then

$$\mathbf{R}(\widetilde{\boldsymbol{\theta}})^{\intercal}\sqrt{n}\boldsymbol{\lambda}_n = -\underbrace{\sqrt{n}\mathbf{s}\left(\widetilde{\boldsymbol{\theta}}\right)}_{O_p(1)} \implies \mathbf{R}(\widetilde{\boldsymbol{\theta}})'\sqrt{n}\boldsymbol{\lambda}_n = O_p(1)$$

Since $\mathbf{R}(\widetilde{\boldsymbol{\theta}}) \xrightarrow{p} \mathbf{R}_0$, we obtain:

$$\mathbf{R}(\widetilde{\boldsymbol{\theta}})\sqrt{n}\boldsymbol{\lambda}_n = \mathbf{R}_0^{\intercal}\sqrt{n}\boldsymbol{\lambda}_n + \left(\mathbf{R}(\widetilde{\boldsymbol{\theta}}) - \mathbf{R}_0\right)'\sqrt{n}\boldsymbol{\lambda}_n = \mathbf{R}_0^{\intercal}\sqrt{n}\boldsymbol{\lambda}_n + o_p(1)$$

Substituting these three equations into the FOCs, we obtain:

$$\begin{pmatrix} \underset{(K\times K)}{\mathbb{E}\left[\mathbf{H}(\mathbf{w}_i;\boldsymbol{\theta}_0)\right]} & \underset{K\times r}{\mathbf{R}_0^{\intercal}} \\ \underset{(r\times K)}{\mathbf{R}_0} & \underset{(r\times r)}{\mathbf{0}} \end{pmatrix} \begin{pmatrix} \underset{(K\times 1)}{\sqrt{n}\left(\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\right)} \\ \underset{(r\times 1)}{\sqrt{n}\boldsymbol{\lambda}_n} \end{pmatrix} = \begin{pmatrix} \underset{(K\times 1)}{-\sqrt{n}\frac{\partial \log L(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}}} \\ \underset{(r\times 1)}{\mathbf{0}} \end{pmatrix} + o_p(1)$$

# Preliminaries to the next two statistics

This can be solved using the partitioned inverse formula:

$$\begin{pmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} \\ \mathbf{A}_{21} & \mathbf{A}_{22} \end{pmatrix}^{-1} = \begin{pmatrix} \mathbf{A}_{11}^{-1} + \mathbf{A}_{11}^{-1}\mathbf{A}_{12}(\mathbf{A}_{22} - \mathbf{A}_{21}\mathbf{A}_{11}^{-1}\mathbf{A}_{12})^{-1}\mathbf{A}_{21}\mathbf{A}_{11}^{-1} & -\mathbf{A}_{11}^{-1}\mathbf{A}_{12}(\mathbf{A}_{22} - \mathbf{A}_{21}\mathbf{A}_{11}^{-1}\mathbf{A}_{12}) \\ -(\mathbf{A}_{22} - \mathbf{A}_{21}\mathbf{A}_{11}^{-1}\mathbf{A}_{12})^{-1}\mathbf{A}_{21}\mathbf{A}_{11}^{-1} & (\mathbf{A}_{22} - \mathbf{A}_{21}\mathbf{A}_{11}^{-1}\mathbf{A}_{12}) \end{pmatrix}$$

Then:

$$\begin{pmatrix} \sqrt{n}\left(\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\right) \\ (K \times 1) \\ \sqrt{n}\boldsymbol{\lambda}_n \\ (r \times 1) \end{pmatrix} = \begin{pmatrix} -\mathbb{E}\left[\mathbf{H}(\mathbf{w}_i; \boldsymbol{\theta}_0)\right]^{-1} + \mathbb{E}\left[\mathbf{H}(\mathbf{w}_i; \boldsymbol{\theta}_0)\right]^{-1}\mathbf{R}_0^{\intercal}(\mathbf{R}_0\mathbb{E}\left[\mathbf{H}(\mathbf{w}_i; \boldsymbol{\theta}_0)\right]^{-1}\mathbf{R}_0^{\intercal})^{-1}\mathbf{R}_0\mathbb{E}\left[\mathbf{H}(\mathbf{w}_i; \boldsymbol{\theta}_0)\right]^{-} \\ -(\mathbf{R}_0\mathbb{E}\left[\mathbf{H}(\mathbf{w}_i; \boldsymbol{\theta}_0)\right]^{-1}\mathbf{R}_0^{\intercal})^{-1}\mathbf{R}_0\mathbb{E}\left[\mathbf{H}(\mathbf{w}_i; \boldsymbol{\theta}_0)\right]^{-1} \end{pmatrix}$$

## Proof LR.

By second order Taylor series:

$$\log L(\widetilde{\boldsymbol{\theta}}) = \log L(\widehat{\boldsymbol{\theta}}) + \frac{\partial \log L(\widehat{\boldsymbol{\theta}})}{\partial \boldsymbol{\theta}} \left(\widetilde{\boldsymbol{\theta}} - \widehat{\boldsymbol{\theta}}\right) + \frac{1}{2} \left(\widetilde{\boldsymbol{\theta}} - \widehat{\boldsymbol{\theta}}\right)' \frac{\partial^2 \log L(\bar{\boldsymbol{\theta}})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} \left(\widetilde{\boldsymbol{\theta}} - \widehat{\boldsymbol{\theta}}\right)$$

where $\bar{\boldsymbol{\theta}} = \alpha \widetilde{\boldsymbol{\theta}} + (1 - \alpha) \widehat{\boldsymbol{\theta}}$ for some $\alpha \in \alpha\,[0, 1]$. Recall that $\frac{\partial \log L(\widehat{\boldsymbol{\theta}})}{\partial \boldsymbol{\theta}} = \mathbf{0}$ and $n^{-1} \frac{\partial^2 \log L(\bar{\boldsymbol{\theta}})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} \xrightarrow{p} \mathbb{E}\left[\mathbf{H}(\mathbf{w}_i, \boldsymbol{\theta}_0)\right]$. It follows that:

$$\log L(\widetilde{\boldsymbol{\theta}}) - \log L(\widehat{\boldsymbol{\theta}}) = \frac{1}{2} \left(\widetilde{\boldsymbol{\theta}} - \widehat{\boldsymbol{\theta}}\right)' \frac{\partial^2 \log L(\bar{\boldsymbol{\theta}})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} \left(\widetilde{\boldsymbol{\theta}} - \widehat{\boldsymbol{\theta}}\right)$$

$$2 \cdot \left(\log L(\widetilde{\boldsymbol{\theta}}) - \log L(\widehat{\boldsymbol{\theta}})\right) = \left(\widetilde{\boldsymbol{\theta}} - \widehat{\boldsymbol{\theta}}\right)^{\intercal} \frac{\partial^2 \log L(\bar{\boldsymbol{\theta}})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^{\intercal}} \left(\widetilde{\boldsymbol{\theta}} - \widehat{\boldsymbol{\theta}}\right)$$

$$2\sqrt{n}\sqrt{n} \cdot \left(\log L(\widetilde{\boldsymbol{\theta}}) - \log L(\widehat{\boldsymbol{\theta}})\right) = \sqrt{n} \left(\widetilde{\boldsymbol{\theta}} - \widehat{\boldsymbol{\theta}}\right)' n^{-1} \frac{\partial^2 \log L(\bar{\boldsymbol{\theta}})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^{\intercal}} \sqrt{n} \left(\widetilde{\boldsymbol{\theta}} - \widehat{\boldsymbol{\theta}}\right)$$

$$2 \cdot n \cdot \left(\log L(\widetilde{\boldsymbol{\theta}}) - \log L(\widehat{\boldsymbol{\theta}})\right) = \sqrt{n} \left(\widetilde{\boldsymbol{\theta}} - \widehat{\boldsymbol{\theta}}\right)^{\intercal} n^{-1} \frac{\partial^2 \log L(\bar{\boldsymbol{\theta}})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} \sqrt{n} \left(\widetilde{\boldsymbol{\theta}} - \widehat{\boldsymbol{\theta}}\right)$$

## Proof LR.

Adding and subtracting $\sqrt{n}\left(\widetilde{\boldsymbol{\theta}}-\widehat{\boldsymbol{\theta}}\right)^{\intercal} n^{-1}\frac{\partial^2 \log L(\boldsymbol{\theta}_0)}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^{\intercal}}\sqrt{n}\left(\widetilde{\boldsymbol{\theta}}-\widehat{\boldsymbol{\theta}}\right)$:

$$
\begin{aligned}
2\cdot n\cdot\left(\log L(\widetilde{\boldsymbol{\theta}})-\log L(\widehat{\boldsymbol{\theta}})\right) &= \sqrt{n}\left(\widetilde{\boldsymbol{\theta}}-\widehat{\boldsymbol{\theta}}\right)^{\intercal} n^{-1}\frac{\partial^2 \log L(\boldsymbol{\theta}_0)}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}\sqrt{n}\left(\widetilde{\boldsymbol{\theta}}-\widehat{\boldsymbol{\theta}}\right)+ \\
&\quad+\underbrace{\sqrt{n}\left(\widetilde{\boldsymbol{\theta}}-\widehat{\boldsymbol{\theta}}\right)^{\intercal}\left[n^{-1}\frac{\partial^2 \log L(\bar{\boldsymbol{\theta}})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}-n^{-1}\frac{\partial^2 \log L(\boldsymbol{\theta}_0)}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}\right]\sqrt{n}\left(\widetilde{\boldsymbol{\theta}}-\widehat{\boldsymbol{\theta}}\right)}_{o_p(1)=O_p(1)o_p(1)O_p(1)} \\
&= \sqrt{n}\left(\widetilde{\boldsymbol{\theta}}-\widehat{\boldsymbol{\theta}}\right)^{\intercal} n^{-1}\frac{\partial^2 \log L(\boldsymbol{\theta}_0)}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^{\intercal}}\sqrt{n}\left(\widetilde{\boldsymbol{\theta}}-\widehat{\boldsymbol{\theta}}\right)+o_p(1) \\
&= \sqrt{n}\left(\widetilde{\boldsymbol{\theta}}-\widehat{\boldsymbol{\theta}}\right)' \mathbb{E}\left[\mathbf{H}(\mathbf{w}_i,\boldsymbol{\theta}_0)\right]\sqrt{n}\left(\widetilde{\boldsymbol{\theta}}-\widehat{\boldsymbol{\theta}}\right)+o_p(1) \\
2\cdot n\cdot\left(\log L(\widehat{\boldsymbol{\theta}})-\log L(\widetilde{\boldsymbol{\theta}})\right) &= -\sqrt{n}\left(\widetilde{\boldsymbol{\theta}}-\widehat{\boldsymbol{\theta}}\right)' \mathbb{E}\left[\mathbf{H}(\mathbf{w}_i,\boldsymbol{\theta}_0)\right]\sqrt{n}\left(\widetilde{\boldsymbol{\theta}}-\widehat{\boldsymbol{\theta}}\right)+o_p(1) \quad \times -1
\end{aligned}
$$

□

# Proof LR.

We know that

$$\begin{aligned}
\sqrt{n}\left(\widetilde{\boldsymbol{\theta}}-\widehat{\boldsymbol{\theta}}\right) &= \sqrt{n}\left(\widetilde{\boldsymbol{\theta}}-\boldsymbol{\theta}_0\right)-\sqrt{n}\left(\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0\right) \\
&= -\left(\mathbb{E}\left[\mathbf{H}(\mathbf{w}_i;\boldsymbol{\theta}_0)\right]^{-1}-\mathbb{E}\left[\mathbf{H}(\mathbf{w}_i;\boldsymbol{\theta}_0)\right]^{-1}\mathbf{R}_0^{\intercal}(\mathbf{R}_0\mathbb{E}\left[\mathbf{H}(\mathbf{w}_i;\boldsymbol{\theta}_0)\right]^{-1}\mathbf{R}_0^{\intercal})^{-1}\mathbf{R}_0\mathbb{E}\left[\mathbf{H}(\mathbf{w}_i;\boldsymbol{\theta}_0)\right]^{-1}\right)\times \\
&\times\sqrt{n}\frac{\partial\log L(\boldsymbol{\theta}_0)}{\partial\boldsymbol{\theta}}-\left(-\mathbb{E}\left[\mathbf{H}(\mathbf{w}_i;\boldsymbol{\theta}_0)\right]^{-1}\sqrt{n}\frac{\partial\log L(\boldsymbol{\theta}_0)}{\partial\boldsymbol{\theta}}\right)+o_p(1) \\
&= \mathbb{E}\left[\mathbf{H}(\mathbf{w}_i;\boldsymbol{\theta}_0)\right]^{-1}\mathbf{R}_0^{\intercal}(\mathbf{R}_0\mathbb{E}\left[\mathbf{H}(\mathbf{w}_i;\boldsymbol{\theta}_0)\right]^{-1}\mathbf{R}_0')^{-1}\mathbf{R}_0\mathbb{E}\left[\mathbf{H}(\mathbf{w}_i;\boldsymbol{\theta}_0)\right]^{-1}\sqrt{n}\frac{\partial\log L(\boldsymbol{\theta}_0)}{\partial\boldsymbol{\theta}}+o_p(1)
\end{aligned}$$

Then:

$$\begin{aligned}
2\cdot n\cdot\left(\log L(\widehat{\boldsymbol{\theta}})-\log L(\widetilde{\boldsymbol{\theta}})\right) &= -\sqrt{n}\left(\widetilde{\boldsymbol{\theta}}-\widehat{\boldsymbol{\theta}}\right)'\mathbb{E}\left[\mathbf{H}(\mathbf{w}_i,\boldsymbol{\theta}_0)\right]\sqrt{n}\left(\widetilde{\boldsymbol{\theta}}-\widehat{\boldsymbol{\theta}}\right)+o_p(1) \\
&= \left(\sqrt{n}\frac{\partial\log L(\boldsymbol{\theta}_0)}{\partial\boldsymbol{\theta}}\right)'\mathbb{E}\left[\mathbf{H}(\mathbf{w}_i;\boldsymbol{\theta}_0)\right]^{-1}\mathbf{R}_0^{\intercal}\left(\mathbf{R}_0\mathbb{E}\left[\mathbf{H}(\mathbf{w}_i;\boldsymbol{\theta}_0)\right]^{-1}\mathbf{R}_0^{\intercal}\right)^{-1}\times \\
&= \times\mathbf{R}_0\mathbb{E}\left[\mathbf{H}(\mathbf{w}_i;\boldsymbol{\theta}_0)\right]^{-1}\underbrace{\left(\sqrt{n}\frac{\partial\log L(\boldsymbol{\theta}_0)}{\partial\boldsymbol{\theta}}\right)}_{\xrightarrow{d}\mathrm{N}\left(\mathbf{0},-\mathbb{E}\left[\mathbf{H}\left(\mathbf{w}_i;\boldsymbol{\theta}_0\right)\right]\right)}
\end{aligned}$$

□

## Proof LR.

Then :

$$\mathbf{R}_0\mathbb{E}\left[\mathbf{H}(\mathbf{w}_i;\boldsymbol{\theta}_0)\right]^{-1}\left(\sqrt{n}\frac{\partial \log L(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}}\right) \xrightarrow{d} \mathrm{N}\left(\mathbf{0}, \mathbf{R}_0\mathbb{E}\left[\mathbf{H}(\mathbf{w}_i;\boldsymbol{\theta}_0)\right]^{-1} - \mathbb{E}\left[\mathbf{H}(\mathbf{w}_i;\boldsymbol{\theta}_0)\right]\mathbb{E}\left[\mathbf{H}\left(\mathbf{w}_i;\boldsymbol{\theta}_0\right)\right]^{-1}\mathbf{R}_0^{\intercal}\right)$$
$$\xrightarrow{d} \mathrm{N}\left(\mathbf{0}, -\mathbf{R}_0\mathbb{E}\left[\mathbf{H}(\mathbf{w}_i;\boldsymbol{\theta}_0)\right]^{-1}\mathbf{R}_0'\right)$$

This asymptotic variance cancels against the central term of the quadratic form , and hence we are looking at the norm of a $\#r$-dimensional standard normal vector:

$$LR \equiv 2 \cdot n \cdot \left(\log L(\widehat{\boldsymbol{\theta}}) - \log L(\widetilde{\boldsymbol{\theta}})\right) \xrightarrow{d} \chi^2(\#r)$$

□