Lecture 3: Maximum Likelihood Estimator

Mauricio Sarrias

Universidad Católica del Norte

October 2, 2017

1. What are the consequences of applying OLS to SLM?
   - Finite and Asymptotic Properties
   - Illustration of bias

2. Maximum Likelihood Estimator (MLE)
   - Introduction to MLE
   - Maximum Likelihood Estimator
   - Identification
   - The Score Function
   - The Information Matrix

3. Asymptotic Properties
   - Consistency
   - Asymptotic Normality
   - Estimation of Variance

1. What are the consequences of applying OLS to SLM?
   - Finite and Asymptotic Properties
   - Illustration of bias
2. Maximum Likelihood Estimator (MLE)
   - Introduction to MLE
   - Maximum Likelihood Estimator
   - Identification
   - The Score Function
   - The Information Matrix
3. Asymptotic Properties
   - Consistency
   - Asymptotic Normality
   - Estimation of Variance

Finite and asymptotic properties

Main idea:

We will show that and OLS estimate of ρ will be biased and inconsistent.

Finite and asymptotic properties

Consider the following pure first order spatial autoregressive model:

$$\underset{(n\times 1)}{\mathbf{y}} = \rho_0 \underset{(n\times 1)}{\mathbf{Wy}} + \underset{(n\times 1)}{\boldsymbol{\varepsilon}}, \tag{1}$$

where ρ_0 is the true population parameter of the data generating process (DGP). The reduced form for the **pure SLM** in (1) is:

$$\mathbf{y} = (\mathbf{I}_n - \rho_0 \mathbf{W})^{-1} \boldsymbol{\varepsilon}. \tag{2}$$

As a result, the spatial lag term equals:

$$\mathbf{Wy} = \mathbf{W} (\mathbf{I}_n - \rho_0 \mathbf{W})^{-1} \boldsymbol{\varepsilon}. \tag{3}$$

This result will be useful later. Now, recall that if the model is $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, then the OLS estimator is $\widehat{\boldsymbol{\beta}} = \left(\mathbf{X}^\top \mathbf{X}\right)^{-1} \mathbf{X}^\top \mathbf{y}$. Then, considering (1) the OLS estimate for ρ_0 is:

$$\widehat{\rho}_{OLS} = \left[\underbrace{(\mathbf{Wy})^\top}_{(1\times n)} \underbrace{(\mathbf{Wy})}_{(n\times 1)} \right]^{-1} \underbrace{(\mathbf{Wy})^\top}_{(1\times n)} \underbrace{\mathbf{y}}_{(n\times 1)}. \tag{4}$$

Finite and asymptotic properties

Substituting the expression for $\mathbf{y}$ in the population equation (1) into (4) gives us the following sampling error equation:

$$\begin{aligned}\widehat{\rho}_{OLS} &= \rho_0 + \left[(\mathbf{Wy})^\top (\mathbf{Wy})\right]^{-1} (\mathbf{Wy})^\top \boldsymbol{\varepsilon} \\ &= \rho_0 + \left(\sum_{i=1}^{n} \mathbf{y}_{Li}^2\right)^{-1} \left(\sum_{i=1}^{n} \mathbf{y}_{Li}\epsilon_i\right),\end{aligned}$$

where $\mathbf{y}_{Li}$ is the ith element of the spatial lag operator $\mathbf{Wy} = \mathbf{y}_L$. Assuming that $\mathbf{W}$ is nonstochastic, the mathematical expectation of $\widehat{\rho}_{OLS}$ is

$$\begin{aligned}\mathbb{E}\left(\widehat{\rho}_{OLS} \middle| \mathbf{W}\right) &= \rho_0 + \mathbb{E}\left(\left[(\mathbf{Wy})^\top (\mathbf{Wy})\right]^{-1} (\mathbf{Wy})^\top \boldsymbol{\varepsilon} \middle| \mathbf{W}\right) \\ &= \rho_0 + \left(\sum_{i=1}^{n} \mathbf{y}_{Li}^2\right)^{-1} \mathbb{E}\left(\sum_{i=1}^{n} \mathbf{y}_{Li}\epsilon_i \middle| \mathbf{W}\right).\end{aligned} \tag{5}$$

From (5) it is clear that if the expectation of the last term is zero, then $\widehat{\rho}_{OLS}$ will be unbiased.

Finite and asymptotic properties

Note that

$$
\begin{aligned}
\mathbb{E}\left(\sum_{i=1}^{n}\mathbf{y}_{Li}\epsilon_i\middle|\mathbf{W}\right) &= \mathbb{E}\left[(\mathbf{W}\mathbf{y})^{\top}\boldsymbol{\varepsilon}\middle|\mathbf{W}\right] \\
&= \mathbb{E}\left[\boldsymbol{\varepsilon}^{\top}\underset{(1\times 1)}{\left(\mathbf{I}-\rho\mathbf{W}^{\top}\right)^{-1}}\mathbf{W}^{\top}\boldsymbol{\varepsilon}\middle|\mathbf{W}\right] \quad \text{using (3)} \\
&= \mathbb{E}\left[\boldsymbol{\varepsilon}^{\top}\mathbf{C}^{\top}\boldsymbol{\varepsilon}\middle|\mathbf{W}\right] \qquad (6) \\
&= \mathbb{E}\left[\operatorname{tr}\left(\boldsymbol{\varepsilon}^{\top}\mathbf{C}^{\top}\boldsymbol{\varepsilon}\right)\middle|\mathbf{W}\right] \\
&= \mathbb{E}\left[\operatorname{tr}\left(\mathbf{C}^{\top}\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^{\top}\right)\middle|\mathbf{W}\right] \quad \text{since } \operatorname{tr}(ABC)=\operatorname{tr}(BCA) \\
&= \operatorname{tr}(\mathbf{C})\,\mathbb{E}\left(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^{\top}\middle|\mathbf{W}\right) \quad \text{since } \operatorname{tr}(\mathbf{A})=\operatorname{tr}(\mathbf{A}^{\top}) \\
&\neq 0,
\end{aligned}
$$

where $\mathbf{C}=\mathbf{W}\left(\mathbf{I}-\rho\mathbf{W}\right)^{-1}$. Therefore, given the result in (6) we have that $\mathbb{E}\left(\widehat{\rho}_{OLS}\middle|\mathbf{W}\right)=\rho_0$ if and only if $\operatorname{tr}(\mathbf{C})=0$, which occurs if $\rho_0=0$. If $\rho=0$, $\mathbf{C}=\mathbf{W}$, and $\operatorname{tr}(\mathbf{C})=\operatorname{tr}(\mathbf{W})=0$ because the diagonal elements of $\mathbf{W}$ are zeros. In other words, if the true model follows a spatial autoregressive structure, the OLS estimate of ρ will be biased.

What about consistency?

Note that we can write:

$$\widehat{\rho}_{OLS} = \rho_0 + \left(\frac{1}{n} \sum_{i=1}^{N} \mathbf{y}_{Li}^2 \right)^{-1} \left(\frac{1}{n} \sum_{i=1}^{n} \mathbf{y}_{Li} \epsilon_i \right). \tag{7}$$

Under 'some conditions' we can show that:

$$\frac{1}{n} \sum_{i=1}^{n} \mathbf{y}_{Li}^2 \to q, \tag{8}$$

where q is some finite scalar (We need some assumptions here about ρ and the structure of the spatial weight matrix). However, for the second term we obtain

$$\frac{1}{n} \sum_{i=1}^{n} \mathbf{y}_{Li} \epsilon_i \xrightarrow{p} \mathbb{E}(\mathbf{y}_{Li} \boldsymbol{\varepsilon}_i) = \operatorname{tr}(\mathbf{C}) \, \mathbb{E}(\boldsymbol{\varepsilon} \boldsymbol{\varepsilon}^\top) \neq 0. \tag{9}$$

- Quadratic form of error terms, which introduces endogeneity.
- Then, $\widehat{\rho}_{OLS}$ is inconsistent, and we need to account for the simultaneity (ML or IV)

1. What are the consequences of applying OLS to SLM?
   - Finite and Asymptotic Properties
   - Illustration of bias
2. Maximum Likelihood Estimator (MLE)
   - Introduction to MLE
   - Maximum Likelihood Estimator
   - Identification
   - The Score Function
   - The Information Matrix
3. Asymptotic Properties
   - Consistency
   - Asymptotic Normality
   - Estimation of Variance

Illustration in R

- We will perform a simple simulation experiment to assess the properties of the OLS estimator.
- We will assume that the true DGP is:

$$\mathbf{y} = \rho_0 \mathbf{W}\mathbf{y} + \boldsymbol{\varepsilon} \tag{10}$$

where

  - the true value $\rho_0 = 0.7$,
  - the sample size for each sample is $n = 225$,
  - $\boldsymbol{\varepsilon} \sim \mathrm{N}(0, 1)$ and $\mathbf{W}$ is an artificial $n \times n$ weight matrix,
  - The $\mathbf{W}$ is constructed from a neighbor list for rook contiguity on a 500×500 regular lattice.

Illustration in R

The syntax for creating the global parameters for the simulation in R is the following:

```
# Global parameters
library("spdep")                          # Load package
set.seed(123)                             # Set seed
S       <- 100                            # Number of simulations
n       <- 225                            # spatial units
rho     <- 0.7                            # True rho
w       <- cell2nb(sqrt(n), sqrt(n)) # Create artifitial W matrix
iw      <- invIrM(w, rho)                 # Compute inverse of (I - rho*W)
rho_hat <- vector(mode = "numeric", length = S) # Vector to save results.
```

Illustration in R

The loop for the simulation is the following

```
# Loop for simulation
for (s in 1:S) {
  e <- rnorm(n, mean = 0 , sd = 1) # Create error term
  y <- iw %*% e                    # True DGP
  Wy <- lag.listw(nb2listw(w), y)  # Create spatial lag
  out <- lm(y ~ Wy)                # Estimate OLS
  rho_hat[s] <- coef(out)["Wy"]    # Save results
}
```

Note that since $\mathbf{W}$ is cosidered as fixed it is created out of simulation loop.

Illustration in R

The results are the following

```
summary(rho_hat)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.8309  0.9981  1.0330  1.0330  1.0750  1.1680
```

- The estimated ρ ranges from 0.8 to 1.2, that is, the range does not include the true parameter $\rho_0 = 0.7$.
- The mean of the estimated parameters is 1, which is very far away from 0.7. We can conclude that the OLS estimator of the pure SLM model is highly biased.

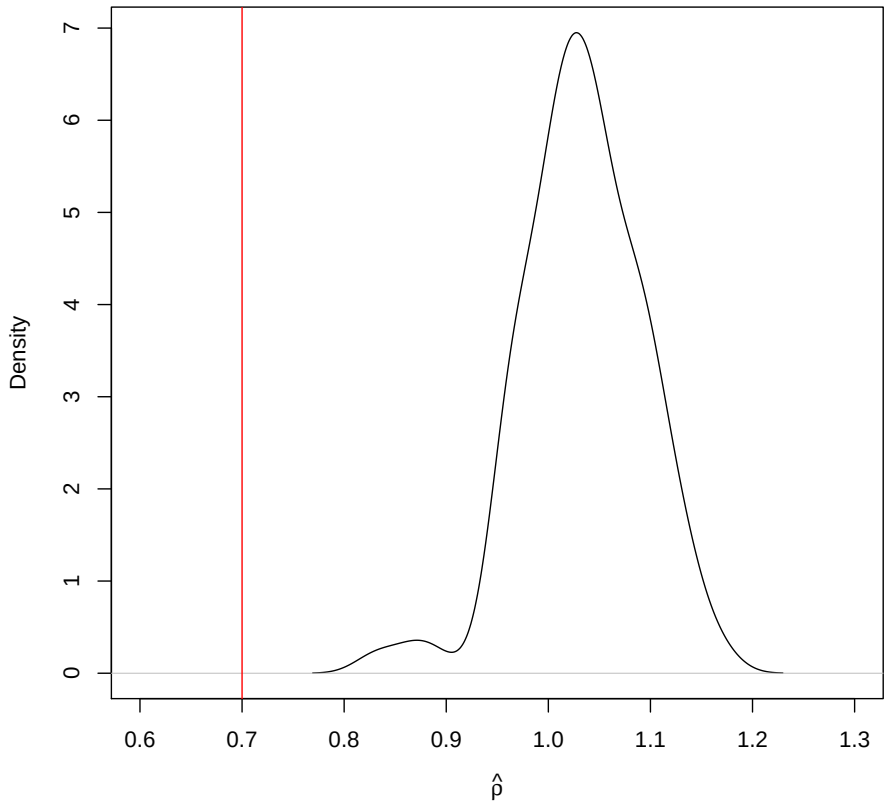
Illustration in R

Finally, we can plot the sampling distribution of the estimated parameters in the following way:

```
plot(density(rho_hat),
     xlab = expression(hat(rho)),
     main = "")
abline(v = rho, col = "red")
```

Illustration in R

Figure: Distribution of $\widehat{\rho}$



Notes: This graph shows the sampling distribution of ρ estimated by OLS for each sample in the Monte Carlo simulation study. The true DGP follows a pure Spatial Lag Model where the true parameter is $\rho_0 = 0.7$

Maximum Likelihood Estimator

- The maximum likelihood estimate (MLE) is a way to estimate the value of a parameter of interest
- The MLE is the value of $\boldsymbol{\theta}$ that **maximizes** the likelihood.

Likelihood Function

- Let $f(y_i|\mathbf{x}_i;\boldsymbol{\theta})$ the conditional density...
- ... that is, the probability of observing $y_i|\mathbf{x}_i$.

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 estimation is to find a set of parameters estimates, say $\widehat{\boldsymbol{\theta}}$, such that this likelihood is maximized.

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 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:

Linear Regression

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

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

The joint p.d.f of the sample is:

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

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 $\sigma_0^2 > 0$

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 $\boldsymbol{\Theta} = \mathbb{R}^K$.

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

- We will assume that the the sample is i.i.d.
- We also assume that we know the true conditional density (this is a strong assumption!).

Assumption: Distribution

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

Expected Log-Likelihood Inequality

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

Assumption: 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

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

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:

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

Assumption: 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*

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

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

The score function

- Note that the MLE is the solution to a maximization problem.
- Therefore as any optimization problem, we need the first and second order conditions.
- The problem is that sometimes the FOC do not have a closed form solution.

Differentiability

Assumption: 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 bout 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.,

$$\int_{\mathcal{S}} dF(y_i|\mathbf{x}_i;\boldsymbol{\theta}) = \int_{\mathcal{S}} f(y_i|\mathbf{x}_i;\boldsymbol{\theta})dy_i = 1 \tag{11}$$

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

$$\mathbf{0} = \int_{\mathcal{S}} \frac{\partial}{\partial \boldsymbol{\theta}} f(y_i|\mathbf{x}_i;\boldsymbol{\theta})dy_i \tag{12}$$

This equation states how changes in $f(y_i|\mathbf{x}_i;\boldsymbol{\theta})$ resulting from changes in $\boldsymbol{\theta}$ are restricted by (11). We can rewrite (12) as

$$\begin{aligned}
\mathbf{0} &= \int_{\mathcal{S}} \frac{f(y_i|\mathbf{x}_i;\boldsymbol{\theta})}{f(y_i|\mathbf{x}_i;\boldsymbol{\theta})} \frac{\partial}{\partial \boldsymbol{\theta}} f(y_i|\mathbf{x}_i;\boldsymbol{\theta})dy_i \\
\mathbf{0} &= \int_{\mathcal{S}} \frac{1}{f(y_i|\mathbf{x}_i;\boldsymbol{\theta})} \frac{\partial f(y_i|\mathbf{x}_i;\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \underbrace{dF(y_i|\mathbf{x}_i;\boldsymbol{\theta})}_{f(y_i|\mathbf{x}_i;\boldsymbol{\theta})dy_i}
\end{aligned} \tag{13}$$

□

Proof.

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

$$\begin{aligned}
\frac{\partial}{\partial \boldsymbol{\theta}} \ln f(y_i|\mathbf{x}_i;\boldsymbol{\theta}) &\equiv \frac{1}{f(y_i|\mathbf{x}_i;\boldsymbol{\theta})} \frac{\partial}{\partial \boldsymbol{\theta}} f(y_i|\mathbf{x}_i;\boldsymbol{\theta}) \\
\mathbf{s}(\mathbf{w}_i;\boldsymbol{\theta}) &\equiv \frac{1}{f(y_i|\mathbf{x}_i;\boldsymbol{\theta})} \frac{\partial}{\partial \boldsymbol{\theta}} f(y_i|\mathbf{x}_i;\boldsymbol{\theta}) \\
\mathbf{s}(\mathbf{w}_i;\boldsymbol{\theta}) f(y_i|\mathbf{x}_i;\boldsymbol{\theta}) &\equiv \frac{\partial}{\partial \boldsymbol{\theta}} f(y_i|\mathbf{x}_i;\boldsymbol{\theta})
\end{aligned} \tag{14}$$

Then, substituting into (13)

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

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.

Hessian

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

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

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:

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

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: Finite Information

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

Lemma (Information Identity)

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

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

Proof: (Homework)

Information Identity

Note the following:

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

Therefore we can write:

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

Example

Recall that:

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

We have:

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

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

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

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

Example

So for $\boldsymbol{\theta} = \boldsymbol{\theta}_0$ the $\widehat{\epsilon}_i$ in these expressions can be replaced by ϵ_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:

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

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

1. What are the consequences of applying OLS to SLM?
   - Finite and Asymptotic Properties
   - Illustration of bias
2. Maximum Likelihood Estimator (MLE)
   - Introduction to MLE
   - Maximum Likelihood Estimator
   - Identification
   - The Score Function
   - The Information Matrix
3. Asymptotic Properties
   - Consistency
   - Asymptotic Normality
   - Estimation of Variance

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.

Question

How can we proceed?

Some Ideas

Using some LLN we know that:

$$\frac{1}{n}\sum_{i=1}^{n}\log f(y_i|\mathbf{x}_i;\boldsymbol{\theta}) \xrightarrow{p} \mathbb{E}\left[\log f(y_i|\mathbf{x}_i;\boldsymbol{\theta})\right] \tag{15}$$

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 $g(\cdot) = \arg\max_{\boldsymbol{\theta}\in\boldsymbol{\Theta}}(\cdot)$

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 ω) and of a parameter $\boldsymbol{\theta}$. By increasing the sample size, we obtain a sequence of log-likelihoods that depend on ω 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

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.*

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

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* $(y_i, \mathbf{x}_i)$,
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

$$\frac{1}{n}\sum_{i=1}^{n} \log f(y_i|\mathbf{x}_i; \boldsymbol{\theta}) \xrightarrow{p} \mathbb{E}\left[\log f(y_i|\mathbf{x}_i; \boldsymbol{\theta})\right] \tag{16}$$

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

Asymptotic

Theorem (Asymptotic Normality of Conditional ML)

Let $\mathbf{w} \equiv (y_i, \mathbf{x}_i')'$ *be i.i.d. Suppose the conditions of either Theorem 14 or 15 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* $-\mathbb{E}\left[\mathbf{H}(\mathbf{w}_i;\boldsymbol{\theta}_0)\right] = \mathbb{E}\left[\mathbf{s}(\mathbf{w}_i;\boldsymbol{\theta}_0)\mathbf{s}(\mathbf{w}_i;\boldsymbol{\theta}_0)'\right]$,
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:

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

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,

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

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

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

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

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

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

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:

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

□

1. What are the consequences of applying OLS to SLM?
   - Finite and Asymptotic Properties
   - Illustration of bias
2. Maximum Likelihood Estimator (MLE)
   - Introduction to MLE
   - Maximum Likelihood Estimator
   - Identification
   - The Score Function
   - The Information Matrix
3. Asymptotic Properties
   - Consistency
   - Asymptotic Normality
   - Estimation of Variance

Variance Estimation

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

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

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:

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

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

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

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 $\sqrt{n}(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0)$