Maximum Likelihood

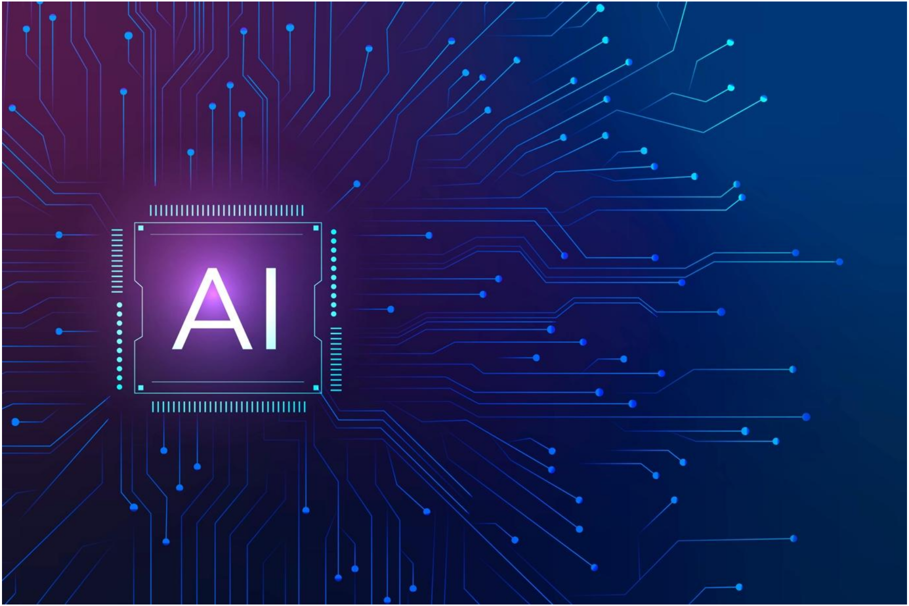
Zhentao Shi
https://zhentaoshi.github.io/

The Chinese University of Hong Kong

Scientific Reasoning

Deductive Reasoning

Inductive Reasoning



Where is Econometrics Going?

Covered topics

- Linear models
- OLS
- Endogeneity
- 2SLS, GMM

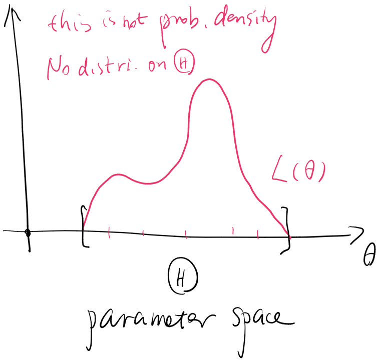
We will continue...

Task

- Purpose: predict y with X
- Beyond continuous random variables
  - Binary
  - Multi-responses
  - Integer
  - Mixed type: censoring, truncation
  - Self-selection

Likelihood

- Philosophy: The most likely outcome (Abductive reasoning).
- Distributional assumption

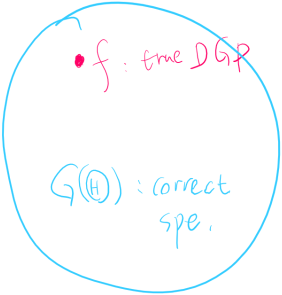


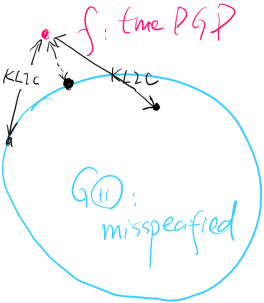
Principles

- Hero: Ronald Fisher (1890–1962)
- General framework
- Special cases of MLE
  - OLS
  - LIML
- Numerical optimization

Model Specification

- Nature: Data z is drawn from a parameter model f
- Human: specify a family of models $g(z;\theta)$ and a parameter space Θ, which span a **model space** $G(\Theta) = \{g(z;\theta) : \theta \in \Theta\}$.





Model and Specification

Parametric model. The distribution of the data $\mathbf{Y} = (Y_1, ..., Y_N)$ is known up to a finite dimensional parameter.

- **Semiparametric model**: If we know $Y \sim i.i.d. \left(\mu, \sigma^2\right)$, we can estimate μ, σ^2 by method of moments.
- **Parametric model**: If we assume $Y \sim N\left(\mu, \sigma^2\right)$, the model has only two parameters μ and σ^2.

Likelihood Function

- For simplicity, let $\mathbf{Y} = (Y_1, \ldots, Y_N)$ be i.i.d.
- The **likelihood** of the sample under a hypothesized value of $\theta \in \Theta$ is

$$L(\theta; \mathbf{Y}) = f(\mathbf{Y}; \theta) = \prod_{i=1}^{N} f(Y_i; \theta)$$

- Two perspectives:
  - (Probabilist) $f(\mathbf{Y}; \theta)$ is a function of $\mathbf{Y}$ given the parameter θ
  - (Statistician) $L(\theta; \mathbf{Y})$ is a function of θ given the data $\mathbf{Y}$

Section 1

Correct Specification

Log-likelihood

- log-likelihood

$$\ell_N(\theta) = \log L(\theta; \mathbf{Y}) = \sum_{i=1}^{N} \log f(Y_i; \theta)$$

  is easier to compute.
- $\log(\cdot)$ is a monotonically increasing function
- The MLE estimator

$$\hat{\theta} = \arg\max_{\theta \in \Theta} \ell_N(\theta)$$

Why Maximization: Deep Justification

Theorem

If the model is correctly specified, then θ_0 is the maximizer.

- Kullback-Leibler information criterion (KLIC):

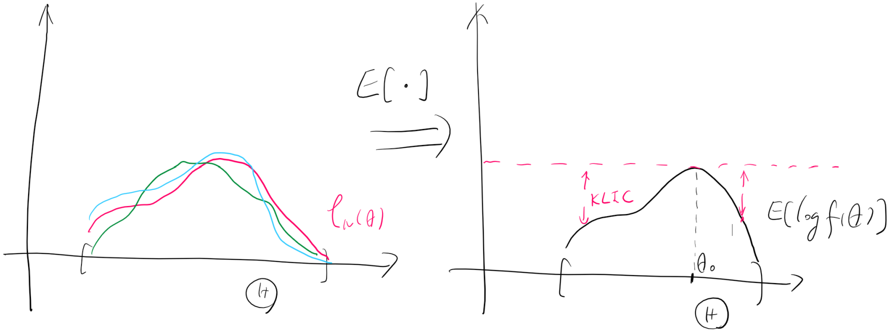
$$KLIC(f,g) = \int f(z) \log \frac{f(z)}{g(z)} dz$$

- $KLIC \geq 0$ because

$$\begin{aligned} & E[\log f(Y;\theta_0)] - E[\log f(Y;\theta)] \\ =& E[\log(f(Y;\theta_0)/f(Y;\theta))] \\ =& -E[\log(f(Y;\theta)/f(Y;\theta_0))] \\ \geq & -\log E[f(Y;\theta)/f(Y;\theta_0)] = 0 \end{aligned}$$

by the Jensen's inequality.

KLIC



Score and Hessian

- Score $s_N(\theta) = \sum_{i=1}^{N} \frac{\partial}{\partial\theta} \log f(Y_i;\theta)$ is a function of θ
- Efficient score $s_{i0} = \frac{\partial}{\partial\theta} \log f\left(Y_i;\theta_0\right)$ is evaluated at the true value θ_0

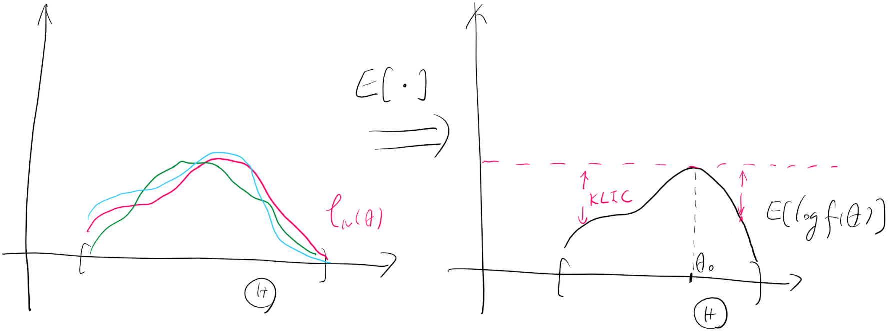
Theorem

If the model is correctly specified, the support of Y does not depend on θ, and θ_0 is in the interior of Θ, then $E\left[s_{i0}\right] = 0$.

MLE is equivalent to looking for roots of $s_N(\theta) = 0$.

- Hessian: $H_N(\theta) = -\sum_{i=1}^{N} \frac{\partial^2}{\partial\theta\partial\theta'} \log f(Y_i;\theta)$
- Expected Hessian: $H_0 = -E\left[\frac{\partial^2}{\partial\theta\partial\theta'} \log f\left(Y;\theta_0\right)\right]$

Score and Hessian: Illustration



Information Equality

- **Fisher Information Matrix**: $I_0 = E[s_{i0}s'_{i0}]$

Theorem

If the model is correctly specified, the support of Υ does not depend on θ, and θ_0 is in the interior of Θ, then

$$I_0 = H_0.$$

- Information equality fails when the model is misspecified

Cramér-Rao Lower Bound

Theorem

Suppose the model is correctly specified, the support of Υ does not depend on θ, and θ_0 is in the interior of Θ. If $\widetilde{\theta}$ is unbiased estimator, then

$$var(\widetilde{\theta}) \geq (NI_0)^{-1}.$$

- More general than "BLUE"
- A lower bound for variance of unbiased estimator
- When reached, an estimator is called **Cramér-Rao efficient**.

Example: Normal MLE

- Normal distribution $Y_i \sim N(\mu, \sigma^2)$ gives density

$$f(Y_i|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\gamma}} \exp\left(-\frac{(Y_i - \mu)^2}{2\gamma}\right)$$

  where $\gamma = \sigma^2$ to simplify notations and derivatives.
- The log-likelihood is

$$\ell_N(Y; \theta) = -\frac{N}{2} \log \gamma - \frac{N}{2} \log 2\pi - \frac{1}{2\gamma} \sum_{i=1}^{N} (Y_i - \mu)^2$$

  where $\theta = (\mu, \gamma)$.

MLE Estimator

Set the score equal to zero

$$s_N(\theta) = \begin{bmatrix} \frac{1}{\gamma}\sum_{i=1}^N (Y_i - \mu) \\ -\frac{N}{2\gamma} + \frac{1}{2\gamma^2}\sum_{i=1}^N (Y_i - \mu)^2 \end{bmatrix} = 0$$

and we solve the MLE estimator

$$\hat{\mu} = \frac{1}{N}\sum_{i=1}^N Y_i = \bar{Y}$$

$$\hat{\gamma} = \frac{1}{N}\sum_{i=1}^N (Y_i - \hat{\mu})^2 = \frac{1}{N}\sum_{i=1}^N (Y_i - \bar{Y})^2$$

Lower Bound

- The Hessian

$$H_N(\theta) = \left[\begin{array}{cc} \frac{N}{\gamma} & \frac{1}{\gamma^2}\sum_{i=1}^{N}(Y_i-\mu) \\ \star & -\frac{N}{2\gamma^2}+\frac{1}{\gamma^3}\sum_{i=1}^{N}(Y_i-\mu)^2 \end{array} \right]$$

- Expectation: $E\left[H_N(\theta_0)\right] = \left[\begin{array}{cc} \frac{N}{\gamma} & 0 \\ 0 & \frac{N}{2\gamma^2} \end{array} \right] = N \times H_0$
- Take inverse: $\left[\begin{array}{cc} \frac{\gamma}{N} & 0 \\ 0 & \frac{2\gamma^2}{N} \end{array} \right]$
- Information equality can be verified.

MLE for the Mean

- The sample mean

$$var\left(\frac{1}{N}\sum_{i=1}^{N} Y_i\right) = \frac{\sigma^2}{N}$$

reaches the Cramér-Rao lower bound

MLE for the Variance

- $E\left(S_N^2\right) = \sigma^2$ is unbiased, where

$$s_N^2 = \frac{1}{N-1}\sum_{i=1}^{N}\left(Y_i - \bar{Y}\right)^2 = \frac{1}{N-1}Y'\left(I - \frac{1}{N}1_N 1_N'\right)Y$$

- It follows $s_N^2 = \frac{\sigma^2}{N-1}\cdot\chi^2\left(N-1\right)$ because

$$\left(N-1\right)\frac{s_N^2}{\sigma^2} = \left(\frac{Y}{\sigma}\right)'\left(I - \frac{1}{N}1_N 1_N'\right)\left(\frac{Y}{\sigma}\right) \sim \chi^2\left(N-1\right).$$

- As a result,

$$var\left(s_N^2\right) = \frac{\sigma^4}{\left(N-1\right)^2}\cdot 2\left(N-1\right) = \frac{2\sigma^4}{N-1} > \frac{2\sigma^4}{N}$$

Is not Cramér-Rao efficient

Return to Normal Regression

- The normal regression models is

$$Y_i = X_i'\beta + \varepsilon_i$$

- Under the assumption $\varepsilon_i \mid X_i \sim N(0, \gamma)$, the conditional distribution is

$$Y_i \mid X_i \sim N\left(X_i'\beta, \gamma\right).$$

- Parameter $\theta = (\beta, \gamma)$
- The joint likelihood

$$f(Y_i, X_i) = f(Y_i|X_i)f(X_i),$$

where the specification of $f(X_i)$ is irrelevant to θ.

Conditional Log-Likelihood

- The **conditional log-likelihood** of the sample is

$$\ell_N(\theta) = -\frac{N}{2}\log\gamma - \frac{N}{2}\log 2\pi - \frac{1}{2\gamma}\sum_{i=1}^{N}\left(Y_i - X_i'\beta\right)^2,$$

  where the distribution of X_i is unspecified.
- The MLE estimator

$$\hat{\beta} = (X'X)^{-1}X'Y$$

$$\hat{\sigma}^2 = \frac{1}{N}\sum_{i=1}^{N}\hat{\varepsilon}_i^2$$

  where $\hat{\varepsilon}_i = Y_i - X_i'\hat{\beta}$.

Asymptotic Normality

- Under regularity conditions, $\hat{\theta} \xrightarrow{p} \theta_0$, and

$$\sqrt{N}\left(\hat{\theta} - \theta_0\right) \xrightarrow{d} N\left(0, H_0^{-1} I_0 H_0^{-1}\right)$$

- When the information equality holds, we have

$$\sqrt{N}\left(\hat{\theta} - \theta_0\right) \xrightarrow{d} N\left(0, I_0^{-1}\right),$$

  or equivalently

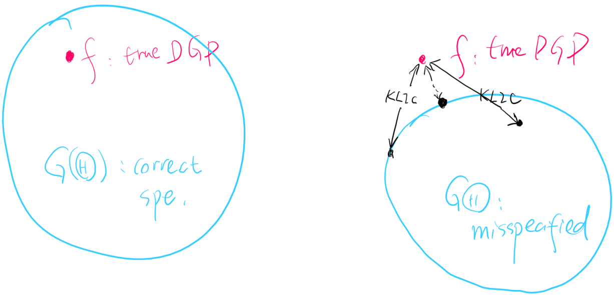
$$\hat{\theta} - \theta_0 \overset{a}{\sim} N\left(0, \frac{I_0^{-1}}{N}\right),$$

- The variance $(NI_0)^{-1}$ is efficient!

Section 2

Mispecification

KLIC for Misspecified Models



- If $f \notin G(\Theta)$, the model is misspecified.

$$
\begin{aligned}
KLIC\left(f, g\left(z;\theta\right)\right) &= \int f\left(z\right)\log f\left(z\right)dz - \int f\left(z\right)\log g\left(z;\theta\right)dz \\
&= E\left[\log f\left(z\right)\right] - E\left[\log g\left(z;\theta\right)\right] > 0
\end{aligned}
$$

Misspecified Model

- Misspecified: $\min_{\theta\in\Theta} KLIC\left(f,g\left(z;\theta\right)\right) > 0$
- MLE is still meaningful
- Pseudo-true parameter:

$$\theta^* = \arg\max_{\theta\in\Theta} E[\ell\left(\theta\right)]$$

the minimizer of $KLIC\left(f,g\left(z;\theta\right)\right)$ in the parameter space Θ

- Under standard assumption, the MLE estimator $\widehat{\theta} \xrightarrow{p} \theta^*$ and

$$\sqrt{N}\left(\hat{\theta}-\theta^*\right) \xrightarrow{d} N\left(0, H_*^{-1} I_* H_*^{-1}\right)$$

Three Tests

A linear hypothesis: $R\theta_0 = q$.

- Wald test: unconstrained estimation, $\ell_N(\hat{\theta})$
- Lagrange multiplier test: constrained estimation under the null, $\ell_N(\tilde{\theta})$
- Likelihood ratio test: $\ell_N(\hat{\theta}) - \ell_N(\tilde{\theta})$

Also works for nonlinear hypothesis.

Are specification tests important?

Summary

- Parametric models
- Specification of distribution family
- MLE
- Score, Hessian, information matrix
- Misspecification

My Related Research

- Ming Li, Zhentao Shi, and Yapeng Zheng (2024) "Estimation and Inference in Dyadic Network Formation Models with Nontransferable Utilities", *working paper*. `https://arxiv.org/abs/2410.23852`
- Jinyuan Chang, Zhentao Shi and Jia Zhang (2023): "Culling the Herd of Moments with Penalized Empirical Likelihood," *Journal of Business & Economic Statistics*, 41(3), 791-805. `https://doi.org/10.1080/07350015.2022.2071903`
- Zhentao Shi (2016): "Econometric Estimation with High-Dimensional Moment Equalities," *Journal of Econometrics*, 195, 104-119. `https://doi.org/10.1016/j.jeconom.2016.07.004`