Maximum Likelihood

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

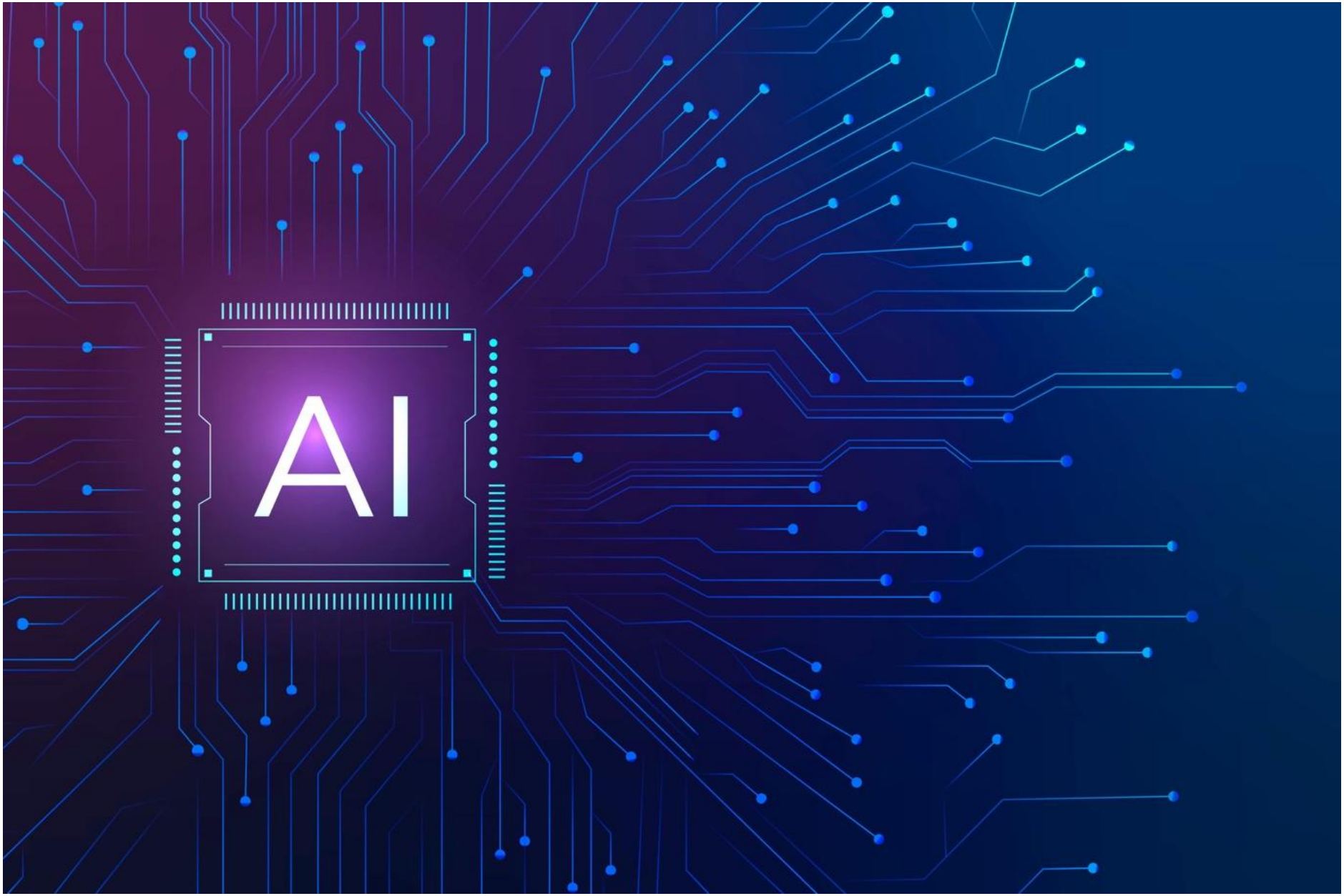
Course website:
https://zhentaoshi.github.io/Econ5121ABC/

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...

In the Era of AI

- Is econometrics a hard subject?

- Many steps of data analysis
- Automating routine tasks
  - Coding and data cleaning
  - Factual math derivation

- Domain knowledge: Interpretation
- Understanding the big picture
- What if AI suggests something completely new to you?

- Welcome to the beautiful new world of AI.

In the Era of AI

- Is econometrics a hard subject?

- Many steps of data analysis
- Automating routine tasks
  - Coding and data cleaning
  - Factual math derivation

- Domain knowledge: Interpretation
- Understanding the big picture
- What if AI suggests something completely new to you?

- Welcome to the beautiful new world of AI.

In the Era of AI

- Is econometrics a hard subject?

- Many steps of data analysis
- Automating routine tasks
  - Coding and data cleaning
  - Factual math derivation

- Domain knowledge: Interpretation
- Understanding the big picture
- What if AI suggests something completely new to you?

- Welcome to the beautiful new world of AI.

In the Era of AI

- Is econometrics a hard subject?

- Many steps of data analysis
- Automating routine tasks
  - Coding and data cleaning
  - Factual math derivation

- Domain knowledge: Interpretation
- Understanding the big picture
- What if AI suggests something completely new to you?

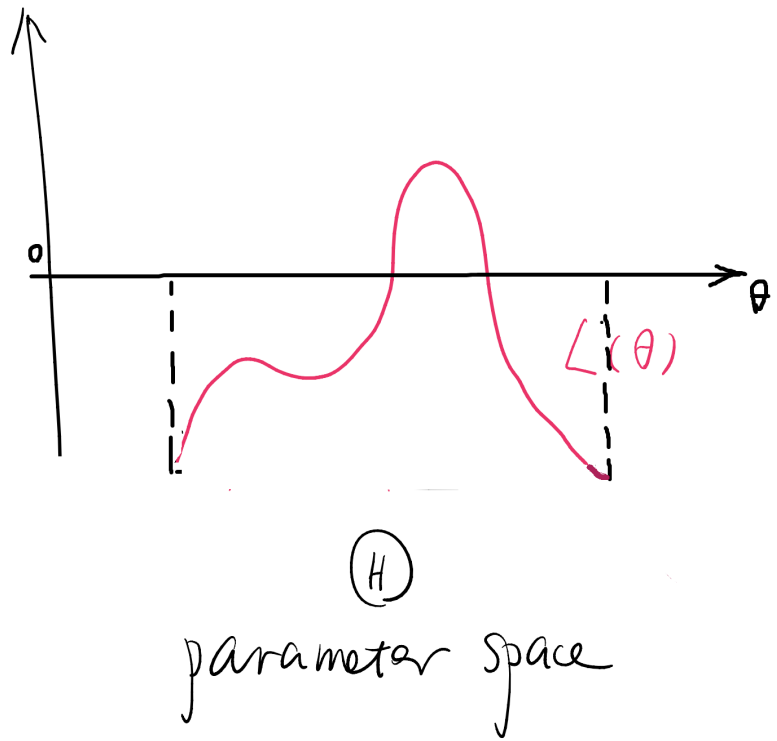
- Welcome to the beautiful new world of AI.

Task

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

Likelihood Approach

- Philosophy: The most likely outcome (Abductive reasoning).
- Distributional assumption (instead of moment conditions)
- Information theory



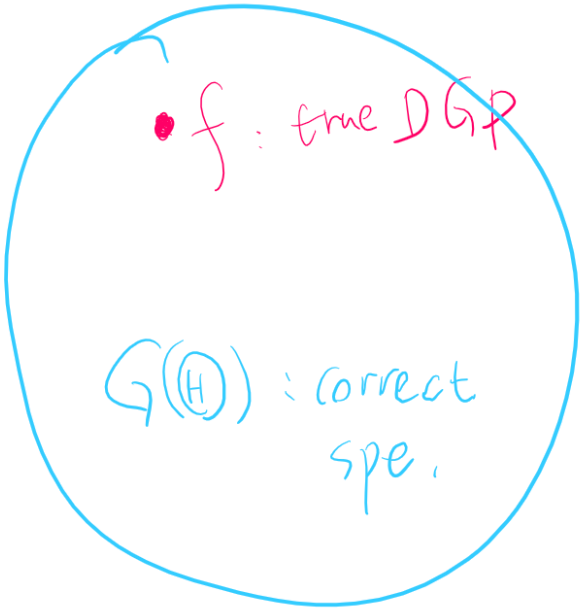
Principles

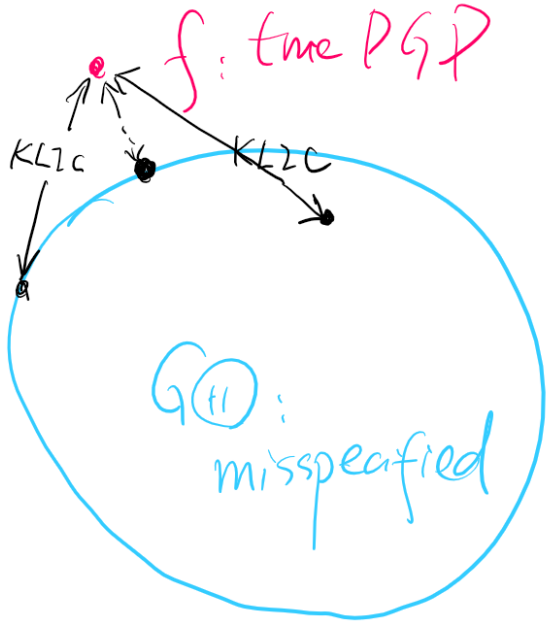
- Hero: Ronald Fisher (1890–1962)

- General framework
- Special cases of MLE
  - OLS
  - LIML
- Numerical optimization

Model Specification

- Nature: draws data Y from a parametric model f
- Human: specifies a family of models $g(y;\theta)$ and a parameter space Θ, which span a **model space** $G(\Theta) = \{g(y;\theta) : \theta \in \Theta\}$.





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 $Y \sim (\mu, \sigma^2)$, we can estimate (μ, σ^2) by method of moments.
- **Parametric model**: If we assume $Y \sim N(\mu, \sigma^2)$, 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; generative) $f(\mathbf{Y}; \theta)$ is a function of $\mathbf{Y}$ given the parameter θ
  - (Statistician; reverse engineering) $L(\theta; \mathbf{Y})$ 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

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

Why Maximization: Deep Justification

Theorem

If the model is correctly specified, then θ_0 is the maximizer of the likelihood function in expectation.

- **Kullback-Leibler information criterion** (KLIC):

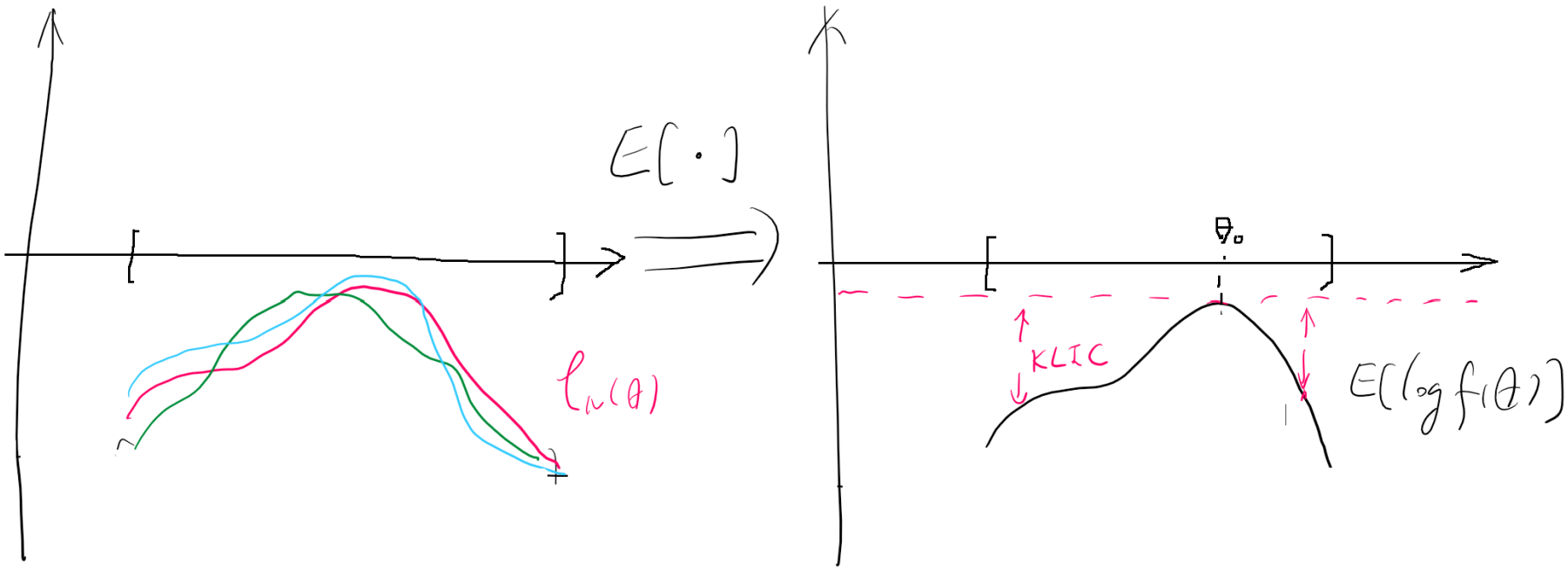
$$KLIC\left(f,g\right)=\int f\left(z\right)\log\frac{f\left(z\right)}{g\left(z\right)}dz$$

- $KLIC\geq 0$ because

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

by the Jensen's inequality.

KLIC



- Numerical demon: Normal MLE
  `https://www.kaggle.com/code/frankshi0/normal-mle`

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 $s_{i0} = \frac{\partial}{\partial \theta} \log f(Y_i; \theta_0)$ is evaluated at the true value θ_0

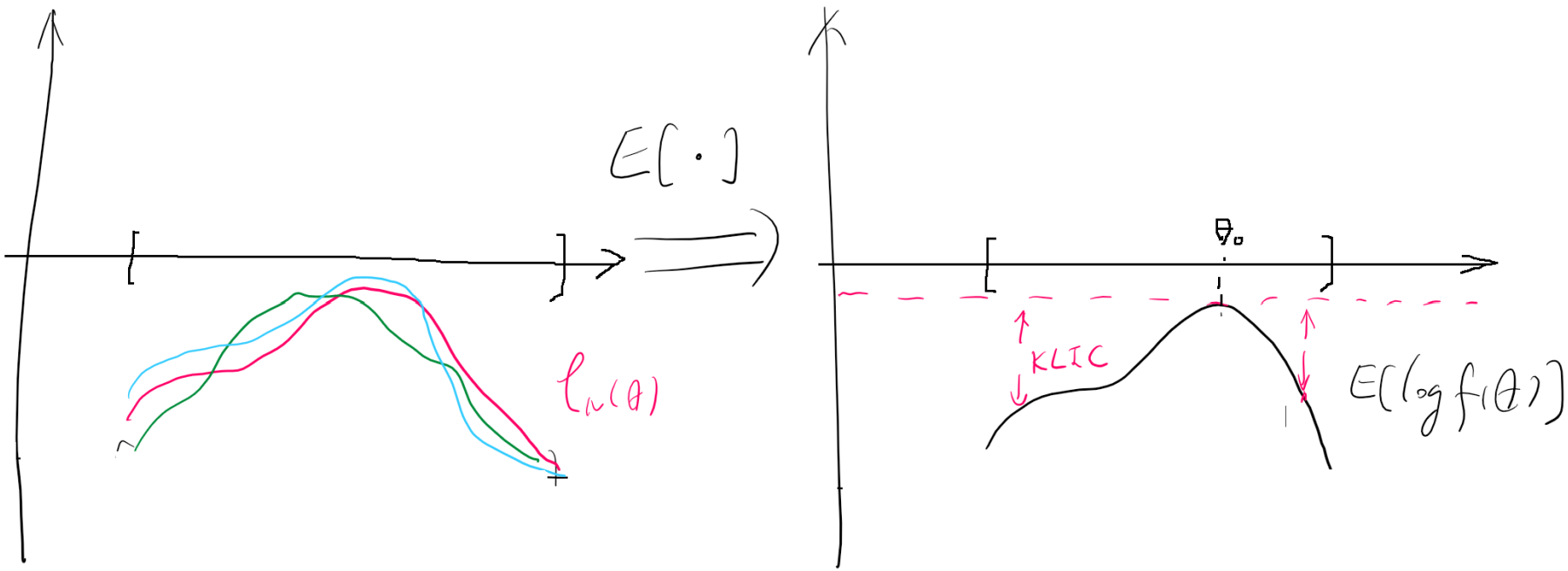
Theorem

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 $E[s_{i0}] = 0$.

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: $H_0 = -E\left[\frac{\partial^2}{\partial \theta \partial \theta'} \log f(Y; \theta_0)\right]$

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 Y 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 Y 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 OLS's "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

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

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

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

Lower Bound

- The Hessian

$$H_N(\theta) = \begin{bmatrix} \frac{N}{\gamma} & \frac{1}{\gamma^2}\sum_{i=1}^{N}(Y_i - \mu) \\ \star & -\frac{N}{2\gamma^2} + \frac{1}{\gamma^3}\sum_{i=1}^{N}(Y_i - \mu)^2 \end{bmatrix}$$

- Expectation: $E[H_N(\theta_0)] = \begin{bmatrix} \frac{N}{\gamma} & 0 \\ 0 & \frac{N}{2\gamma^2} \end{bmatrix} = N \times H_0$
- Take inverse: $\begin{bmatrix} \frac{\gamma}{N} & 0 \\ 0 & \frac{2\gamma^2}{N} \end{bmatrix}$
- 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

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!

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

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

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

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

Return to Normal Regression

- Supervised learning vs unsupervised learning
- 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

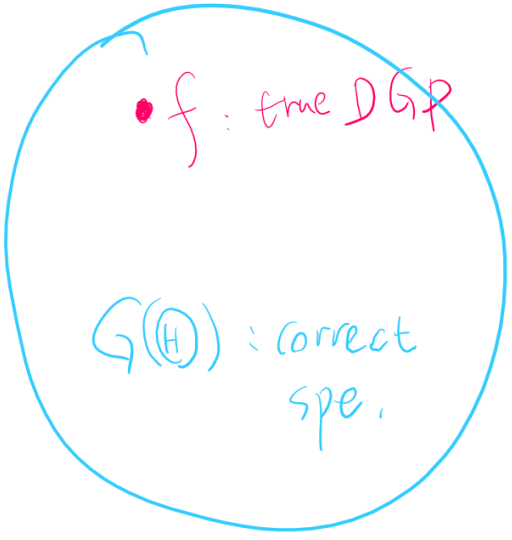
$$f(Y_i, X_i) = f(Y_i|X_i)f(X_i),$$

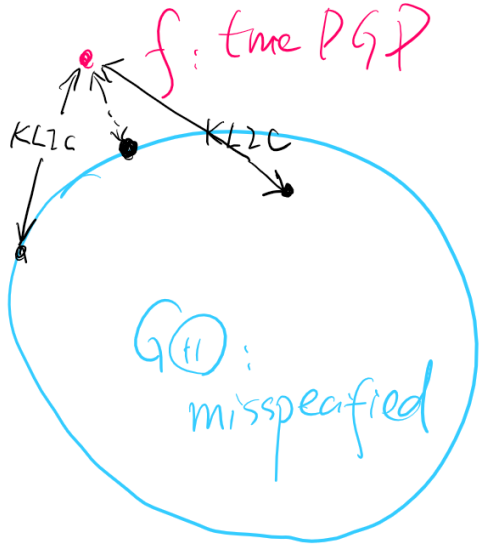
where the specification of $f(X_i)$ is irrelevant to θ.

Section 2

Mispecification

KLIC for Misspecified Models





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

$$KLIC(f, g(y;\theta)) = \int f(y) \log f(y)\, dy - \int f(y) \log g(y;\theta)\, dy$$
$$= E[\log f(y)] - E[\log g(y;\theta)] > 0$$

Misspecified Model

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

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

  the minimizer of $KLIC\left(f, g\left(y;\theta\right)\right)$ 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

Testing a hypothesis: $H_0 : R\theta_0 = q$ vs. $H_1 : R\theta_0 \neq q$.

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

- All three are asymptotically equivalent
- Also work for nonlinear hypothesis
- Trade-off: computational convenience vs. finite sample performance

Wald Test

- Null hypothesis: $H_0 : R\theta_0 = q$, where R is $r \times k$ with rank r
- Uses the **unconstrained MLE** $\hat{\theta}$
- Test statistic:

$$W = N(R\hat{\theta} - q)' \left[R\hat{I}_0^{-1} R' \right]^{-1} (R\hat{\theta} - q)$$

  where $\hat{I}_0 = \frac{1}{N} \sum_{i=1}^{N} \hat{s}_i \hat{s}_i'$ is the estimated information matrix
- Under H_0:

$$W \xrightarrow{d} \chi^2(r)$$

- Reject H_0 if $W > \chi_r^2(\alpha)$, the $(1 - \alpha)$ quantile

Wald Test: Intuition

- Measures how far the unrestricted estimate $\hat{\theta}$ is from satisfying the constraint
- If H_0 is true, $R\hat{\theta}$ should be close to q
- The distance $(R\hat{\theta} - q)$ is standardized by its asymptotic variance

- **Advantages**:
  - Only requires unconstrained estimation
  - Easy to compute
  - Most commonly used in practice
- **Disadvantages**:
  - Not invariant to reparameterization
  - May have poor finite sample properties

Lagrange Multiplier (LM) Test

- Also called the **Score test**
- Uses the **constrained MLE** $\tilde{\theta}$ under $H_0 : R\theta = q$
- Test statistic:

$$LM = \frac{1}{N} s_N(\tilde{\theta})' \tilde{I}_0^{-1} s_N(\tilde{\theta})$$

  where $s_N(\tilde{\theta}) = \sum_{i=1}^{N} \frac{\partial}{\partial \theta} \log f(Y_i; \tilde{\theta})$ is the score at $\tilde{\theta}$
- Under H_0:

$$LM \xrightarrow{d} \chi^2(r)$$

- Reject H_0 if $LM > \chi^2_r(\alpha)$

LM Test: Intuition

- If H_0 is true, the score at $\tilde{\theta}$ should be close to zero
- If the constraint is violated, imposing it will push us away from the true maximum, making $s_N(\tilde{\theta})$ large

- **Advantages**:
  - Only requires constrained estimation
  - Useful when unconstrained estimation is difficult
  - Common in misspecification tests
- **Disadvantages**:
  - Need to solve constrained optimization

Likelihood Ratio (LR) Test

- Uses both $\hat{\theta}$ (unconstrained) and $\tilde{\theta}$ (constrained)
- Test statistic:

$$LR = 2[\ell_N(\hat{\theta}) - \ell_N(\tilde{\theta})]$$

- Under H_0:

$$LR \xrightarrow{d} \chi^2(r)$$

- Reject H_0 if $LR > \chi^2_r(\alpha)$

- Measures the improvement in log-likelihood from relaxing the constraint

LR Test: Intuition

- If H_0 is true, imposing the constraint should not significantly reduce the likelihood
- If H_0 is false, the constrained model fits much worse

- **Advantages**:
  - Invariant to reparameterization
  - Often has better finite sample properties
  - Natural interpretation as likelihood comparison
- **Disadvantages**:
  - Requires both constrained and unconstrained estimation

Summary

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