# Econ 722 – Advanced Econometrics IV, Part II

Francis J. DiTraglia

University of Pennsylvania

# Lecture #1 – AIC-type Information Criteria

Kullback-Leibler Divergence

Bias of Maximized Sample Log-Likelihood

Review of Asymptotics for Mis-specified MLE

Deriving AIC and TIC

Corrected AIC ($\text{AIC}_c$)

# Kullback-Leibler (KL) Divergence

## Motivation

How well does a given density $f(y)$ approximate an unknown true density $g(y)$? Use this to select between parametric models.

# Kullback-Leibler (KL) Divergence

## Motivation

How well does a given density $f(y)$ approximate an unknown true density $g(y)$? Use this to select between parametric models.

## Definition

$$\text{KL}(g; f) = \underbrace{\mathbb{E}_G\left[\log\left\{\frac{g(Y)}{f(Y)}\right\}\right]}_{\text{True density on top}} = \underbrace{\mathbb{E}_G\left[\log g(Y)\right]}_{\substack{\text{Depends only on truth} \\ \text{Fixed across models}}} - \underbrace{\mathbb{E}_G\left[\log f(Y)\right]}_{\substack{\text{Expected} \\ \text{log-likelihood}}}$$

# Kullback-Leibler (KL) Divergence

## Motivation

How well does a given density $f(y)$ approximate an unknown true density $g(y)$? Use this to select between parametric models.

## Definition

$$\text{KL}(g; f) = \underbrace{\mathbb{E}_G\left[\log\left\{\frac{g(Y)}{f(Y)}\right\}\right]}_{\text{True density on top}} = \underbrace{\mathbb{E}_G\left[\log g(Y)\right]}_{\substack{\text{Depends only on truth} \\ \text{Fixed across models}}} - \underbrace{\mathbb{E}_G\left[\log f(Y)\right]}_{\substack{\text{Expected} \\ \text{log-likelihood}}}$$

## Properties

- *Not* symmetric: $\text{KL}(g; f) \neq \text{KL}(f; g)$
- By Jensen's Inequality: $\text{KL}(g; f) \geq 0$ (strict iff $g = f$ a.e.)
- Minimize KL $\iff$ Maximize Expected log-likelihood

# KL Divergence and Mis-specified MLE

## Pseudo-true Parameter Value $\theta_0$

$$\widehat{\theta}_{MLE} \xrightarrow{p} \theta_0 \equiv \underset{\theta \in \Theta}{\arg\min}\ \mathrm{KL}(g; f_\theta) = \underset{\theta \in \Theta}{\arg\max}\ \mathbb{E}_G[\log f(Y|\theta)]$$

# KL Divergence and Mis-specified MLE

## Pseudo-true Parameter Value $\theta_0$

$$\widehat{\theta}_{MLE} \xrightarrow{p} \theta_0 \equiv \underset{\theta \in \Theta}{\arg\min}\ \text{KL}(g; f_\theta) = \underset{\theta \in \Theta}{\arg\max}\ \mathbb{E}_G[\log f(Y|\theta)]$$

## What if $f_\theta$ is correctly specified?

If $g = f_\theta$ for some $\theta$ then $\text{KL}(g; f_\theta)$ is minimized at zero.

# KL Divergence and Mis-specified MLE

## Pseudo-true Parameter Value $\theta_0$

$$\widehat{\theta}_{MLE} \xrightarrow{p} \theta_0 \equiv \underset{\theta \in \Theta}{\arg\min}\; \mathrm{KL}(g; f_\theta) = \underset{\theta \in \Theta}{\arg\max}\; \mathbb{E}_G[\log f(Y|\theta)]$$

## What if $f_\theta$ is correctly specified?

If $g = f_\theta$ for some $\theta$ then $\mathrm{KL}(g; f_\theta)$ is minimized at zero.

## Goal: Compare Mis-specified Models

$$\mathbb{E}_G\left[\log f(Y|\theta_0)\right] \qquad \text{versus} \qquad \mathbb{E}_G\left[\log h(Y|\gamma_0)\right]$$

where $\theta_0$ is the pseudo-true parameter value for $f_\theta$ and $\gamma_0$ is the pseudo-true parameter value for $h_\gamma$.

# How to Estimate Expected Log Likelihood?

For simplicity: $Y_1, \ldots, Y_n \sim$ iid $g(y)$

## Unbiased but Infeasible

$$\mathbb{E}_G\left[\frac{1}{T}\ell(\theta_0)\right] = \mathbb{E}_G\left[\frac{1}{T}\sum_{t=1}^{T} f(Y_t|\theta_0)\right] = \mathbb{E}_G\left[f(Y|\theta_0)\right]$$

# How to Estimate Expected Log Likelihood?

For simplicity: $Y_1, \ldots, Y_n \sim$ iid $g(y)$

## Unbiased but Infeasible

$$\mathbb{E}_G\left[\frac{1}{T}\ell(\theta_0)\right] = \mathbb{E}_G\left[\frac{1}{T}\sum_{t=1}^{T} f(Y_t|\theta_0)\right] = \mathbb{E}_G\left[f(Y|\theta_0)\right]$$

## Biased but Feasible

$T^{-1}\ell(\widehat{\theta}_{MLE})$ is a biased estimator of $\mathbb{E}_G[\log f(Y|\theta_0)]$.

# How to Estimate Expected Log Likelihood?

For simplicity: $Y_1, \ldots, Y_n \sim$ iid $g(y)$

## Unbiased but Infeasible

$$\mathbb{E}_G\left[\frac{1}{T}\ell(\theta_0)\right] = \mathbb{E}_G\left[\frac{1}{T}\sum_{t=1}^{T} f(Y_t|\theta_0)\right] = \mathbb{E}_G\left[f(Y|\theta_0)\right]$$

## Biased but Feasible

$T^{-1}\ell(\widehat{\theta}_{MLE})$ is a biased estimator of $\mathbb{E}_G[\log f(Y|\theta_0)]$.

## Intuition for the Bias

$T^{-1}\ell(\widehat{\theta}_{MLE}) > T^{-1}\ell(\theta_0)$ unless $\widehat{\theta}_{MLE} = \theta_0$. Maximized sample log-like. is an overly optimistic estimator of expected log-like.

# What to do about this bias?

1. General-purpose asymptotic approximation of "degree of over-optimism" of maximized sample log-likelihood.
   - Takeuchi's Information Criterion (TIC)
   - Akaike's Information Criterion (AIC)

# What to do about this bias?

1. General-purpose asymptotic approximation of "degree of over-optimism" of maximized sample log-likelihood.
   - Takeuchi's Information Criterion (TIC)
   - Akaike's Information Criterion (AIC)
2. Problem-specific finite sample approach, assuming $g \in f_\theta$.
   - Corrected AIC ($\text{AIC}_c$) of Hurvich and Tsai (1989)

# What to do about this bias?

1. General-purpose asymptotic approximation of "degree of over-optimism" of maximized sample log-likelihood.
   - Takeuchi's Information Criterion (TIC)
   - Akaike's Information Criterion (AIC)
2. Problem-specific finite sample approach, assuming $g \in f_\theta$.
   - Corrected AIC ($\text{AIC}_c$) of Hurvich and Tsai (1989)

## Tradeoffs

TIC is most general and makes weakest assumptions, but requires very large $T$ to work well. AIC is a good approximation to TIC that requires less data. Both AIC and TIC perform poorly when $T$ is small relative to the number of parameters, hence $\text{AIC}_c$.

# Recall: Asymptotics for Mis-specified ML Estimation

Model $f(y|\theta)$, pseudo-true parameter $\theta_0$. For simplicity $Y_1, \ldots, Y_T \sim$ iid $g(y)$.

# Recall: Asymptotics for Mis-specified ML Estimation

Model $f(y|\theta)$, pseudo-true parameter $\theta_0$. For simplicity $Y_1, \ldots, Y_T \sim$ iid $g(y)$.

## Fundamental Expansion

$$\sqrt{T}(\widehat{\theta} - \theta_0) = J^{-1}\left(\sqrt{T}\bar{U}_T\right) + o_p(1)$$

$$J = -\mathbb{E}_G\left[\frac{\partial \log f(Y|\theta_0)}{\partial\theta\partial\theta'}\right], \quad \bar{U}_T = \frac{1}{T}\sum_{t=1}^{T}\frac{\partial \log f(Y_t)}{\partial\theta}$$

# Recall: Asymptotics for Mis-specified ML Estimation

Model $f(y|\theta)$, pseudo-true parameter $\theta_0$. For simplicity $Y_1, \ldots, Y_T \sim$ iid $g(y)$.

## Fundamental Expansion

$$\sqrt{T}(\widehat{\theta} - \theta_0) = J^{-1}\left(\sqrt{T}\bar{U}_T\right) + o_p(1)$$

$$J = -\mathbb{E}_G\left[\frac{\partial \log f(Y|\theta_0)}{\partial\theta\partial\theta'}\right], \quad \bar{U}_T = \frac{1}{T}\sum_{t=1}^{T}\frac{\partial \log f(Y_t)}{\partial\theta}$$

## Central Limit Theorem

$$\sqrt{T}\bar{U}_T \to_d U \sim N_p(0, K), \quad K = \text{Var}_G\left[\frac{\partial \log f(Y|\theta_0)}{\partial\theta}\right]$$

$$\sqrt{T}(\widehat{\theta} - \theta_0) \to_d J^{-1}U \sim N_p(0, J^{-1}KJ^{-1})$$

# Recall: Asymptotics for Mis-specified ML Estimation

Model $f(y|\theta)$, pseudo-true parameter $\theta_0$. For simplicity $Y_1, \dots, Y_T \sim$ iid $g(y)$.

## Fundamental Expansion

$$\sqrt{T}(\widehat{\theta} - \theta_0) = J^{-1}\left(\sqrt{T}\bar{U}_T\right) + o_p(1)$$

$$J = -\mathbb{E}_G\left[\frac{\partial \log f(Y|\theta_0)}{\partial\theta\partial\theta'}\right], \quad \bar{U}_T = \frac{1}{T}\sum_{t=1}^{T}\frac{\partial \log f(Y_t)}{\partial\theta}$$

## Central Limit Theorem

$$\sqrt{T}\bar{U}_T \to_d U \sim N_p(0, K), \quad K = \text{Var}_G\left[\frac{\partial \log f(Y|\theta_0)}{\partial\theta}\right]$$

$$\sqrt{T}(\widehat{\theta} - \theta_0) \to_d J^{-1}U \sim N_p(0, J^{-1}KJ^{-1})$$

## Information Matrix Equality

If $g = f_\theta$ for some $\theta \in \Theta$ then $K = J \implies \text{AVAR}(\widehat{\theta}) = J^{-1}$

# Bias Relative to Infeasible Plug-in Estimator

## Definition of Bias Term $B$

$$B = \underbrace{\frac{1}{T}\ell(\widehat{\theta})}_{\substack{\text{feasible} \\ \text{overly-optimistic}}} - \underbrace{\int g(y) \log f(y|\widehat{\theta})\, dy}_{\substack{\text{uses data only once} \\ \text{infeas. not overly-optimistic}}}$$

# Bias Relative to Infeasible Plug-in Estimator

## Definition of Bias Term $B$

$$B = \underbrace{\frac{1}{T}\ell(\widehat{\theta})}_{\substack{\text{feasible}\\ \text{overly-optimistic}}} - \underbrace{\int g(y) \log f(y|\widehat{\theta})\, dy}_{\substack{\text{uses data only once}\\ \text{infeas. not overly-optimistic}}}$$

## Question to Answer

On average, over the sampling distribution of $\widehat{\theta}$, how large is $B$?

AIC and TIC construct an asymptotic approximation of $\mathbb{E}[B]$.

# Derivation of AIC/TIC

## Step 1: Taylor Expansion

$$B = \bar{Z}_T + (\widehat{\theta} - \theta_0) + (\widehat{\theta} - \theta_0)' J(\widehat{\theta} - \theta_0) + o_p(T^{-1})$$

$$\bar{Z}_T = \frac{1}{T} \sum_{t=1}^{T} \{\log f(Y_t|\theta_0) - \mathbb{E}_G[\log f(Y|\theta_0)\}$$

# Derivation of AIC/TIC

## Step 1: Taylor Expansion

$$B = \bar{Z}_T + (\widehat{\theta} - \theta_0) + (\widehat{\theta} - \theta_0)' J(\widehat{\theta} - \theta_0) + o_p(T^{-1})$$

$$\bar{Z}_T = \frac{1}{T}\sum_{t=1}^{T}\{\log f(Y_t|\theta_0) - \mathbb{E}_G[\log f(Y|\theta_0)\}$$

## Step 2: $\mathbb{E}[\bar{Z}_T] = 0$

$$\mathbb{E}[B] \approx \mathbb{E}\left[(\widehat{\theta} - \theta_0)' J(\widehat{\theta} - \theta_0)\right]$$

# Derivation of AIC/TIC

## Step 1: Taylor Expansion

$$B = \bar{Z}_T + (\widehat{\theta} - \theta_0) + (\widehat{\theta} - \theta_0)' J(\widehat{\theta} - \theta_0) + o_p(T^{-1})$$

$$\bar{Z}_T = \frac{1}{T} \sum_{t=1}^{T} \{\log f(Y_t|\theta_0) - \mathbb{E}_G[\log f(Y|\theta_0)\}$$

## Step 2: $\mathbb{E}[\bar{Z}_T] = 0$

$$\mathbb{E}[B] \approx \mathbb{E}\left[(\widehat{\theta} - \theta_0)' J(\widehat{\theta} - \theta_0)\right]$$

## Step 3: $\sqrt{T}(\widehat{\theta} - \theta_0) \to_d J^{-1}U$

$$T(\widehat{\theta} - \theta_0)' J(\widehat{\theta} - \theta_0) \to_d U' J^{-1} U$$

# Derivation of AIC/TIC Continued...

Step 3: $\sqrt{T}(\widehat{\theta} - \theta_0) \to_d J^{-1}U$

$$T(\widehat{\theta} - \theta_0)' J(\widehat{\theta} - \theta_0) \to_d U' J^{-1} U$$

# Derivation of AIC/TIC Continued...

Step 3: $\sqrt{T}(\widehat{\theta} - \theta_0) \to_d J^{-1}U$

$$T(\widehat{\theta} - \theta_0)' J(\widehat{\theta} - \theta_0) \to_d U' J^{-1} U$$

Step 4: $U \sim N_p(0, K)$

$$\mathbb{E}[B] \approx \frac{1}{T}\mathbb{E}[U' J^{-1} U] = \text{tr}\left\{J^{-1}K\right\}$$

# Derivation of AIC/TIC Continued...

Step 3: $\sqrt{T}(\widehat{\theta} - \theta_0) \to_d J^{-1}U$

$$T(\widehat{\theta} - \theta_0)'J(\widehat{\theta} - \theta_0) \to_d U'J^{-1}U$$

Step 4: $U \sim N_p(0, K)$

$$\mathbb{E}[B] \approx \frac{1}{T}\mathbb{E}[U'J^{-1}U] = \text{tr}\left\{J^{-1}K\right\}$$

Final Result:

$T^{-1}\text{tr}\left\{J^{-1}K\right\}$ is an asymp. unbiased estimator of the over-optimism of $T^{-1}\ell(\widehat{\theta})$ relative to $\int g(y) \log f(y|\widehat{\theta})\, dy$.

# TIC and AIC

### Takeuchi's Information Criterion

Multiply by $2T$, estimate $J, K \Rightarrow \text{TIC} = 2\left[\ell(\widehat{\theta}) - \text{tr}\left\{\widehat{J}^{-1}\widehat{K}\right\}\right]$

# TIC and AIC

### Takeuchi's Information Criterion

Multiply by $2T$, estimate $J, K \Rightarrow \text{TIC} = 2\left[\ell(\widehat{\theta}) - \text{tr}\left\{\widehat{J}^{-1}\widehat{K}\right\}\right]$

### Akaike's Information Criterion

If $g = f_\theta$ then $J^{-1} = K \Rightarrow \text{tr}\left\{J^{-1}K\right\} = p \Rightarrow \text{AIC} = 2\left[\ell(\widehat{\theta}) - p\right]$

# TIC and AIC

### Takeuchi's Information Criterion

Multiply by $2T$, estimate $J, K \Rightarrow \text{TIC} = 2\left[\ell(\widehat{\theta}) - \text{tr}\left\{\widehat{J}^{-1}\widehat{K}\right\}\right]$

### Akaike's Information Criterion

If $g = f_\theta$ then $J^{-1} = K \Rightarrow \text{tr}\left\{J^{-1}K\right\} = p \Rightarrow \text{AIC} = 2\left[\ell(\widehat{\theta}) - p\right]$

### Contrasting AIC and TIC

Technically, AIC requires that all models under consideration are at least correctly specified while TIC doesn't. But $J^{-1}K$ is hard to estimate, and if a model is badly mis-specified, $\ell(\widehat{\theta})$ dominates.

# Corrected AIC ($\text{AIC}_c$) – Hurvich & Tsai (1989)

## Idea Behind $\text{AIC}_c$

Asymptotic approximation used for AIC/TIC works poorly if $p$ is too large relative to $T$. Try exact, finite-sample approach instead.

# Corrected AIC ($\text{AIC}_c$) – Hurvich & Tsai (1989)

## Idea Behind $\text{AIC}_c$

Asymptotic approximation used for AIC/TIC works poorly if $p$ is too large relative to $T$. Try exact, finite-sample approach instead.

## Assumption: True DGP

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta}_0 + \boldsymbol{\varepsilon}, \quad \boldsymbol{\varepsilon} \sim N(\mathbf{0}, \sigma_0^2 \mathbf{I}_T)$$

# Corrected AIC ($\text{AIC}_c$) – Hurvich & Tsai (1989)

## Idea Behind $\text{AIC}_c$

Asymptotic approximation used for AIC/TIC works poorly if $p$ is too large relative to $T$. Try exact, finite-sample approach instead.

## Assumption: True DGP

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta}_0 + \boldsymbol{\varepsilon}, \quad \boldsymbol{\varepsilon} \sim N(\mathbf{0}, \sigma_0^2 \mathbf{I}_T)$$

## Can Show That

$$KL(g, f) = \frac{T}{2}\left[\frac{\sigma_0^2}{\sigma_1^2} - \log\left(\frac{\sigma_0^2}{\sigma_1^2}\right) - 1\right] + (\boldsymbol{\beta}_0 - \boldsymbol{\beta}_1)'\mathbf{X}'\mathbf{X}(\boldsymbol{\beta}_0 - \boldsymbol{\beta}_1)$$

Where $f$ is a normal regression model with parameters $(\boldsymbol{\beta}_1, \sigma_1^2)$ that might not be the true parameters.

# But how can we use this?

$$KL(g,f) = \frac{T}{2}\left[\frac{\sigma_0^2}{\sigma_1^2} - \log\left(\frac{\sigma_0^2}{\sigma_1^2}\right) - 1\right] + (\boldsymbol{\beta}_0 - \boldsymbol{\beta}_1)'\mathbf{X}'\mathbf{X}(\boldsymbol{\beta}_0 - \boldsymbol{\beta}_1)$$

1. Would need to know $(\beta_1, \sigma_1^2)$ for candidate model.
   - Easy: just use MLE $(\widehat{\boldsymbol{\beta}}_1, \widehat{\sigma}_1^2)$
2. Would need to know $(\boldsymbol{\beta}_0, \sigma_0^2)$ for true model.
   - Very hard! The whole problem is that we don't know these!

# But how can we use this?

$$KL(g,f) = \frac{T}{2}\left[\frac{\sigma_0^2}{\sigma_1^2} - \log\left(\frac{\sigma_0^2}{\sigma_1^2}\right) - 1\right] + (\boldsymbol{\beta}_0 - \boldsymbol{\beta}_1)'\mathbf{X}'\mathbf{X}(\boldsymbol{\beta}_0 - \boldsymbol{\beta}_1)$$

1. Would need to know $(\beta_1, \sigma_1^2)$ for candidate model.
   - Easy: just use MLE $(\widehat{\boldsymbol{\beta}}_1, \widehat{\sigma}_1^2)$
2. Would need to know $(\boldsymbol{\beta}_0, \sigma_0^2)$ for true model.
   - Very hard! The whole problem is that we don't know these!

## Hurvich & Tsai (1989) Assume:

- Every candidate model is at least correctly specified
- Implies any candidate estimator $(\widehat{\boldsymbol{\beta}}, \widehat{\sigma}^2)$ is consistent for truth.

# Deriving the Corrected AIC

Since $(\widehat{\boldsymbol{\beta}}, \widehat{\sigma}^2)$ are random, look at expectation of estimated KL:

$$\mathbb{E}[\widehat{KL}] = \frac{T}{2}\left\{\mathbb{E}\left[\frac{\sigma_0^2}{\widehat{\sigma}^2}\right] - \mathbb{E}\left[\log\left(\frac{\sigma_0^2}{\sigma_1^2}\right)\right] - 1\right\} + \mathbb{E}\left[(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)'\mathbf{X}'\mathbf{X}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)\right]$$

# Deriving the Corrected AIC

Since $(\widehat{\boldsymbol{\beta}}, \widehat{\sigma}^2)$ are random, look at expectation of estimated KL:

$$\mathbb{E}[\widehat{KL}] = \frac{T}{2}\left\{\mathbb{E}\left[\frac{\sigma_0^2}{\widehat{\sigma}^2}\right] - \mathbb{E}\left[\log\left(\frac{\sigma_0^2}{\sigma_1^2}\right)\right] - 1\right\} + \mathbb{E}\left[(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)'\mathbf{X}'\mathbf{X}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)\right]$$

Finite-sample theory for correctly spec. normal regression model:

$$\mathbb{E}\left[\widehat{KL}\right] = \frac{T}{2}\left\{\frac{T+k}{T-k-2} - \log(\sigma_0^2) + \mathbb{E}[\log\widehat{\sigma}^2] - 1\right\}$$

# Deriving the Corrected AIC

Since $(\widehat{\boldsymbol{\beta}}, \widehat{\sigma}^2)$ are random, look at expectation of estimated KL:

$$\mathbb{E}[\widehat{KL}] = \frac{T}{2}\left\{\mathbb{E}\left[\frac{\sigma_0^2}{\widehat{\sigma}^2}\right] - \mathbb{E}\left[\log\left(\frac{\sigma_0^2}{\sigma_1^2}\right)\right] - 1\right\} + \mathbb{E}\left[(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)'\mathbf{X}'\mathbf{X}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)\right]$$

Finite-sample theory for correctly spec. normal regression model:

$$\mathbb{E}\left[\widehat{KL}\right] = \frac{T}{2}\left\{\frac{T+k}{T-k-2} - \log(\sigma_0^2) + \mathbb{E}[\log\widehat{\sigma}^2] - 1\right\}$$

Eliminate constants and scaling, unbiased estimator of $\mathbb{E}[\log\widehat{\sigma}^2]$:

$$\text{AIC}_c = \log\widehat{\sigma}^2 + \frac{T+k}{T-k-2}$$

a finite-sample unbiased estimator of KL for model comparison