

# 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 ( $AIC_c$ )

# 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

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

# 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

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

## 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  $\theta_0$

$$\hat{\theta}_{MLE} \xrightarrow{P} \theta_0 \equiv \arg \min_{\theta \in \Theta} \text{KL}(g; f_{\theta}) = \arg \max_{\theta \in \Theta} \mathbb{E}_G[\log f(Y|\theta)]$$

# KL Divergence and Mis-specified MLE

Pseudo-true Parameter Value  $\theta_0$

$$\hat{\theta}_{MLE} \xrightarrow{P} \theta_0 \equiv \arg \min_{\theta \in \Theta} \text{KL}(g; f_{\theta}) = \arg \max_{\theta \in \Theta} \mathbb{E}_G[\log f(Y|\theta)]$$

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  $\theta_0$

$$\hat{\theta}_{MLE} \xrightarrow{P} \theta_0 \equiv \arg \min_{\theta \in \Theta} \text{KL}(g; f_{\theta}) = \arg \max_{\theta \in \Theta} \mathbb{E}_G[\log f(Y|\theta)]$$

What if  $f_{\theta}$  is correctly specified?

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

Goal: Compare Mis-specified Models

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

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:  $Y_1, \dots, Y_n \sim \text{iid } g(y)$

Unbiased but Infeasible

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

# How to Estimate Expected Log Likelihood?

For simplicity:  $Y_1, \dots, Y_n \sim \text{iid } g(y)$

## Unbiased but Infeasible

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

## Biased but Feasible

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

# How to Estimate Expected Log Likelihood?

For simplicity:  $Y_1, \dots, Y_n \sim \text{iid } g(y)$

## Unbiased but Infeasible

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

## Biased but Feasible

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

## Intuition for the Bias

$T^{-1} \ell(\hat{\theta}_{MLE}) > T^{-1} \ell(\theta_0)$  unless  $\hat{\theta}_{MLE} = \theta_0$ . 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 ( $AIC_c$ ) 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 ( $AIC_c$ ) 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  $AIC_c$ .

# Recall: Asymptotics for Mis-specified ML Estimation

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

## Recall: Asymptotics for Mis-specified ML Estimation

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

### Fundamental Expansion

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

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



# Recall: Asymptotics for Mis-specified ML Estimation

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

## Fundamental Expansion

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

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

## Central Limit Theorem

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

$$\sqrt{T}(\hat{\theta} - \theta_0) \rightarrow_d J^{-1} U \sim N_p(0, J^{-1} K J^{-1})$$

# Recall: Asymptotics for Mis-specified ML Estimation

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

## Fundamental Expansion

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

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

## Central Limit Theorem

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

$$\sqrt{T}(\hat{\theta} - \theta_0) \rightarrow_d J^{-1} U \sim N_p(0, J^{-1} K J^{-1})$$

## Information Matrix Equality

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

# Bias Relative to Infeasible Plug-in Estimator

## Definition of Bias Term $B$

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

# Bias Relative to Infeasible Plug-in Estimator

## Definition of Bias Term $B$

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

## Question to Answer

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

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

# Derivation of AIC/TIC

## Step 1: Taylor Expansion

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

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

# Derivation of AIC/TIC

## Step 1: Taylor Expansion

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

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

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

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

# Derivation of AIC/TIC

## Step 1: Taylor Expansion

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

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

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

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

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

$$T(\hat{\theta} - \theta_0)' J(\hat{\theta} - \theta_0) \rightarrow_d U' J^{-1}U$$

## Derivation of AIC/TIC Continued...

$$\text{Step 3: } \sqrt{T}(\hat{\theta} - \theta_0) \rightarrow_d J^{-1}U$$

$$T(\hat{\theta} - \theta_0)'J(\hat{\theta} - \theta_0) \rightarrow_d U'J^{-1}U$$



## Derivation of AIC/TIC Continued...

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

$$T(\hat{\theta} - \theta_0)'J(\hat{\theta} - \theta_0) \rightarrow_d U'J^{-1}U$$

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

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

## Derivation of AIC/TIC Continued...

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

$$T(\hat{\theta} - \theta_0)'J(\hat{\theta} - \theta_0) \rightarrow_d U'J^{-1}U$$

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

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

Final Result:

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

# TIC and AIC

## Takeuchi's Information Criterion

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

# TIC and AIC

## Takeuchi's Information Criterion

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

## Akaike's Information Criterion

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

# TIC and AIC

## Takeuchi's Information Criterion

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

## Akaike's Information Criterion

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

## 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,  $\ell(\hat{\theta})$  dominates.

# Corrected AIC ( $AIC_c$ ) – Hurvich & Tsai (1989)

## Idea Behind $AIC_c$

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

## Corrected AIC ( $AIC_c$ ) – Hurvich & Tsai (1989)

### Idea Behind $AIC_c$

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

$$\mathbf{y} = \mathbf{X}\beta_0 + \varepsilon, \quad \varepsilon \sim N(\mathbf{0}, \sigma_0^2 \mathbf{I}_T), \quad k \text{ Regressors}$$

## Corrected AIC ( $AIC_c$ ) – Hurvich & Tsai (1989)

### Idea Behind $AIC_c$

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

$$\mathbf{y} = \mathbf{X}\beta_0 + \varepsilon, \quad \varepsilon \sim N(\mathbf{0}, \sigma_0^2 \mathbf{I}_T), \quad k \text{ Regressors}$$

Can Show That

$$KL(g, f) = \frac{T}{2} \left[ \frac{\sigma_0^2}{\sigma_1^2} - \log \left( \frac{\sigma_0^2}{\sigma_1^2} \right) - 1 \right] + \left( \frac{1}{2\sigma_1^2} \right) (\beta_0 - \beta_1)' \mathbf{X}' \mathbf{X} (\beta_0 - \beta_1)$$

Where  $f$  is a normal regression model with parameters  $(\beta_1, \sigma_1^2)$  that might not be the true parameters.



## But how can we use this?

$$KL(g, f) = \frac{T}{2} \left[ \frac{\sigma_0^2}{\sigma_1^2} - \log \left( \frac{\sigma_0^2}{\sigma_1^2} \right) - 1 \right] + \left( \frac{1}{2\sigma_1^2} \right) (\beta_0 - \beta_1)' \mathbf{X}' \mathbf{X} (\beta_0 - \beta_1)$$

1. Would need to know  $(\beta_1, \sigma_1^2)$  for **candidate model**.
  - ▶ Easy: just use MLE  $(\hat{\beta}_1, \hat{\sigma}_1^2)$
2. Would need to know  $(\beta_0, \sigma_0^2)$  for **true model**.
  - ▶ Very hard! The whole problem is that we don't know these!

## But how can we use this?

$$KL(g, f) = \frac{T}{2} \left[ \frac{\sigma_0^2}{\sigma_1^2} - \log \left( \frac{\sigma_0^2}{\sigma_1^2} \right) - 1 \right] + \left( \frac{1}{2\sigma_1^2} \right) (\beta_0 - \beta_1)' \mathbf{X}' \mathbf{X} (\beta_0 - \beta_1)$$

1. Would need to know  $(\beta_1, \sigma_1^2)$  for **candidate model**.
  - ▶ Easy: just use MLE  $(\hat{\beta}_1, \hat{\sigma}_1^2)$
2. Would need to know  $(\beta_0, \sigma_0^2)$  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  $(\hat{\beta}, \hat{\sigma}^2)$  is consistent for truth.

## Deriving the Corrected AIC

Since  $(\hat{\beta}, \hat{\sigma}^2)$  are random, look at  $\mathbb{E}[\widehat{KL}]$ , where

$$\widehat{KL} = \frac{T}{2} \left[ \frac{\sigma_0^2}{\hat{\sigma}^2} - \log \left( \frac{\sigma_0^2}{\hat{\sigma}^2} \right) - 1 \right] + \left( \frac{1}{2\hat{\sigma}^2} \right) (\hat{\beta} - \beta_0)' \mathbf{X}' \mathbf{X} (\hat{\beta} - \beta_0)$$

## Deriving the Corrected AIC

Since  $(\hat{\beta}, \hat{\sigma}^2)$  are random, look at  $\mathbb{E}[\widehat{KL}]$ , where

$$\widehat{KL} = \frac{T}{2} \left[ \frac{\sigma_0^2}{\hat{\sigma}^2} - \log \left( \frac{\sigma_0^2}{\hat{\sigma}^2} \right) - 1 \right] + \left( \frac{1}{2\hat{\sigma}^2} \right) (\hat{\beta} - \beta_0)' \mathbf{X}' \mathbf{X} (\hat{\beta} - \beta_0)$$

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

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

## Deriving the Corrected AIC

Since  $(\hat{\beta}, \hat{\sigma}^2)$  are random, look at  $\mathbb{E}[\widehat{KL}]$ , where

$$\widehat{KL} = \frac{T}{2} \left[ \frac{\sigma_0^2}{\hat{\sigma}^2} - \log \left( \frac{\sigma_0^2}{\hat{\sigma}^2} \right) - 1 \right] + \left( \frac{1}{2\hat{\sigma}^2} \right) (\hat{\beta} - \beta_0)' \mathbf{X}' \mathbf{X} (\hat{\beta} - \beta_0)$$

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

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

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

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

a finite-sample unbiased estimator of KL for model comparison

# Lecture #2 – More on “Classical” Model Selection

Mallow's  $C_p$

Bayesian Model Comparison

Laplace Approximation

Bayesian Information Criterion (BIC)

## Motivation: Predict $\mathbf{y}$ from $\mathbf{x}$ via Linear Regression

$$\underset{(T \times 1)}{\mathbf{y}} = \underset{(T \times K)}{\mathbf{X}} \underset{(K \times 1)}{\boldsymbol{\beta}} + \boldsymbol{\epsilon}$$

$$\mathbb{E}[\boldsymbol{\epsilon}|\mathbf{X}] = 0, \quad \text{Var}(\boldsymbol{\epsilon}|\mathbf{X}) = \sigma^2 \mathbf{I}$$

## Motivation: Predict $\mathbf{y}$ from $\mathbf{x}$ via Linear Regression

$$\underset{(T \times 1)}{\mathbf{y}} = \underset{(T \times K)}{\mathbf{X}} \underset{(K \times 1)}{\boldsymbol{\beta}} + \boldsymbol{\epsilon}$$

$$\mathbb{E}[\boldsymbol{\epsilon}|\mathbf{X}] = 0, \quad \text{Var}(\boldsymbol{\epsilon}|\mathbf{X}) = \sigma^2 \mathbf{I}$$

- If  $\boldsymbol{\beta}$  were known, could never achieve lower MSE than by using all regressors to predict.



## Motivation: Predict $\mathbf{y}$ from $\mathbf{x}$ via Linear Regression

$$\underset{(T \times 1)}{\mathbf{y}} = \underset{(T \times K)}{\mathbf{X}} \underset{(K \times 1)}{\boldsymbol{\beta}} + \boldsymbol{\epsilon}$$

$$\mathbb{E}[\boldsymbol{\epsilon}|\mathbf{X}] = 0, \quad \text{Var}(\boldsymbol{\epsilon}|\mathbf{X}) = \sigma^2 \mathbf{I}$$

- ▶ If  $\boldsymbol{\beta}$  were known, could never achieve lower MSE than by using all regressors to predict.
- ▶ But  $\boldsymbol{\beta}$  is unknown so we have to estimate it from data  $\Rightarrow$  bias-variance tradeoff.

## Motivation: Predict $\mathbf{y}$ from $\mathbf{x}$ via Linear Regression

$$\underset{(T \times 1)}{\mathbf{y}} = \underset{(T \times K)}{\mathbf{X}} \underset{(K \times 1)}{\boldsymbol{\beta}} + \boldsymbol{\epsilon}$$

$$\mathbb{E}[\boldsymbol{\epsilon}|\mathbf{X}] = 0, \quad \text{Var}(\boldsymbol{\epsilon}|\mathbf{X}) = \sigma^2 \mathbf{I}$$

- ▶ If  $\boldsymbol{\beta}$  were known, could never achieve lower MSE than by using all regressors to predict.
- ▶ But  $\boldsymbol{\beta}$  is unknown so we have to estimate it from data  $\Rightarrow$  bias-variance tradeoff.
- ▶ Could make sense to exclude regressors with small coefficients: add small bias but reduce variance.

# Operationalizing the Bias-Variance Tradeoff Idea

## Mallow's $C_p$

Approximate the predictive MSE of each model relative to the infeasible optimum in which  $\beta$  is known.

# Operationalizing the Bias-Variance Tradeoff Idea

## Mallow's $C_p$

Approximate the predictive MSE of each model relative to the infeasible optimum in which  $\beta$  is known.

## Notation

- ▶ Model index  $m$  and regressor matrix  $\mathbf{X}_m$
- ▶ Corresponding OLS estimator  $\hat{\beta}$  padded out with zeros
- ▶  $\mathbf{X}\hat{\beta}_m = \mathbf{X}_{(-m)}\mathbf{0} + \mathbf{X}_m [(\mathbf{X}'_m \mathbf{X}_m)^{-1} \mathbf{X}'_m \mathbf{y}] = \mathbf{P}_m \mathbf{y}$

# In-sample versus Out-of-sample Prediction Error

Why not compare  $RSS(m)$ ?

In-sample prediction error:  $RSS(m) = (\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}_m)'(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}_m)$

# In-sample versus Out-of-sample Prediction Error

Why not compare  $RSS(m)$ ?

In-sample prediction error:  $RSS(m) = (\mathbf{y} - \mathbf{X}\hat{\beta}_m)'(\mathbf{y} - \mathbf{X}\hat{\beta}_m)$

From your Problem Set

RSS cannot decrease even if we add irrelevant regressors. Thus in-sample prediction error is an **overly optimistic** estimate of out-of-sample prediction error.

# In-sample versus Out-of-sample Prediction Error

Why not compare  $RSS(m)$ ?

In-sample prediction error:  $RSS(m) = (\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}_m)'(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}_m)$

From your Problem Set

RSS cannot decrease even if we add irrelevant regressors. Thus in-sample prediction error is an **overly optimistic** estimate of out-of-sample prediction error.

Bias-Variance Tradeoff

Out-of-sample performance of full model (using all regressors) could be very poor if there is a lot of estimation uncertainty associated with regressors that aren't very predictive.

# Predictive MSE of $\mathbf{X}\hat{\boldsymbol{\beta}}_m$ relative to infeasible optimum $\mathbf{X}\boldsymbol{\beta}$

Step 1: Algebra

$$\mathbf{X}\hat{\boldsymbol{\beta}}_m - \mathbf{X}\boldsymbol{\beta} = \mathbf{P}_m\mathbf{y} - \mathbf{X}\boldsymbol{\beta}$$



# Predictive MSE of $\mathbf{X}\hat{\boldsymbol{\beta}}_m$ relative to infeasible optimum $\mathbf{X}\boldsymbol{\beta}$

Step 1: Algebra

$$\mathbf{X}\hat{\boldsymbol{\beta}}_m - \mathbf{X}\boldsymbol{\beta} = \mathbf{P}_m\mathbf{y} - \mathbf{X}\boldsymbol{\beta} = \mathbf{P}_m(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) - (\mathbf{I} - \mathbf{P}_m)\mathbf{X}\boldsymbol{\beta}$$

# Predictive MSE of $\mathbf{X}\hat{\boldsymbol{\beta}}_m$ relative to infeasible optimum $\mathbf{X}\boldsymbol{\beta}$

Step 1: Algebra

$$\begin{aligned}\mathbf{X}\hat{\boldsymbol{\beta}}_m - \mathbf{X}\boldsymbol{\beta} &= \mathbf{P}_m\mathbf{y} - \mathbf{X}\boldsymbol{\beta} = \mathbf{P}_m(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) - (\mathbf{I} - \mathbf{P}_m)\mathbf{X}\boldsymbol{\beta} \\ &= \mathbf{P}_m\boldsymbol{\epsilon} - (\mathbf{I} - \mathbf{P}_m)\mathbf{X}\boldsymbol{\beta}\end{aligned}$$

# Predictive MSE of $\mathbf{X}\hat{\boldsymbol{\beta}}_m$ relative to infeasible optimum $\mathbf{X}\boldsymbol{\beta}$

Step 1: Algebra

$$\begin{aligned}\mathbf{X}\hat{\boldsymbol{\beta}}_m - \mathbf{X}\boldsymbol{\beta} &= \mathbf{P}_m\mathbf{y} - \mathbf{X}\boldsymbol{\beta} = \mathbf{P}_m(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) - (\mathbf{I} - \mathbf{P}_m)\mathbf{X}\boldsymbol{\beta} \\ &= \mathbf{P}_m\boldsymbol{\epsilon} - (\mathbf{I} - \mathbf{P}_m)\mathbf{X}\boldsymbol{\beta}\end{aligned}$$

Step 2:  $\mathbf{P}_m$  and  $(\mathbf{I} - \mathbf{P}_m)$  are symmetric, idempotent, and orthogonal

$$\left\| \mathbf{X}\hat{\boldsymbol{\beta}}_m - \mathbf{X}\boldsymbol{\beta} \right\|^2 = \{ \mathbf{P}_m\boldsymbol{\epsilon} - (\mathbf{I} - \mathbf{P}_m)\mathbf{X}\boldsymbol{\beta} \}' \{ \mathbf{P}_m\boldsymbol{\epsilon} + (\mathbf{I} - \mathbf{P}_m)\mathbf{X}\boldsymbol{\beta} \}$$

# Predictive MSE of $\mathbf{X}\hat{\boldsymbol{\beta}}_m$ relative to infeasible optimum $\mathbf{X}\boldsymbol{\beta}$

Step 1: Algebra

$$\begin{aligned}\mathbf{X}\hat{\boldsymbol{\beta}}_m - \mathbf{X}\boldsymbol{\beta} &= \mathbf{P}_m\mathbf{y} - \mathbf{X}\boldsymbol{\beta} = \mathbf{P}_m(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) - (\mathbf{I} - \mathbf{P}_m)\mathbf{X}\boldsymbol{\beta} \\ &= \mathbf{P}_m\boldsymbol{\epsilon} - (\mathbf{I} - \mathbf{P}_m)\mathbf{X}\boldsymbol{\beta}\end{aligned}$$

Step 2:  $\mathbf{P}_m$  and  $(\mathbf{I} - \mathbf{P}_m)$  are symmetric, idempotent, and orthogonal

$$\begin{aligned}\left\|\mathbf{X}\hat{\boldsymbol{\beta}}_m - \mathbf{X}\boldsymbol{\beta}\right\|^2 &= \{\mathbf{P}_m\boldsymbol{\epsilon} - (\mathbf{I} - \mathbf{P}_m)\mathbf{X}\boldsymbol{\beta}\}' \{\mathbf{P}_m\boldsymbol{\epsilon} + (\mathbf{I} - \mathbf{P}_m)\mathbf{X}\boldsymbol{\beta}\} \\ &= \boldsymbol{\epsilon}'\mathbf{P}_m'\mathbf{P}_m\boldsymbol{\epsilon} - \boldsymbol{\beta}'\mathbf{X}'(\mathbf{I} - \mathbf{P}_m)'\mathbf{P}_m\boldsymbol{\epsilon} - \boldsymbol{\epsilon}'\mathbf{P}_m'(\mathbf{I} - \mathbf{P}_m)\mathbf{X}\boldsymbol{\beta} \\ &\quad + \boldsymbol{\beta}'\mathbf{X}'(\mathbf{I} - \mathbf{P}_m)(\mathbf{I} - \mathbf{P}_m)\mathbf{X}\boldsymbol{\beta}\end{aligned}$$

# Predictive MSE of $\mathbf{X}\hat{\boldsymbol{\beta}}_m$ relative to infeasible optimum $\mathbf{X}\boldsymbol{\beta}$

Step 1: Algebra

$$\begin{aligned}\mathbf{X}\hat{\boldsymbol{\beta}}_m - \mathbf{X}\boldsymbol{\beta} &= \mathbf{P}_m\mathbf{y} - \mathbf{X}\boldsymbol{\beta} = \mathbf{P}_m(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) - (\mathbf{I} - \mathbf{P}_m)\mathbf{X}\boldsymbol{\beta} \\ &= \mathbf{P}_m\boldsymbol{\epsilon} - (\mathbf{I} - \mathbf{P}_m)\mathbf{X}\boldsymbol{\beta}\end{aligned}$$

Step 2:  $\mathbf{P}_m$  and  $(\mathbf{I} - \mathbf{P}_m)$  are symmetric, idempotent, and orthogonal

$$\begin{aligned}\left\|\mathbf{X}\hat{\boldsymbol{\beta}}_m - \mathbf{X}\boldsymbol{\beta}\right\|^2 &= \{\mathbf{P}_m\boldsymbol{\epsilon} - (\mathbf{I} - \mathbf{P}_m)\mathbf{X}\boldsymbol{\beta}\}' \{\mathbf{P}_m\boldsymbol{\epsilon} + (\mathbf{I} - \mathbf{P}_m)\mathbf{X}\boldsymbol{\beta}\} \\ &= \boldsymbol{\epsilon}'\mathbf{P}_m'\mathbf{P}_m\boldsymbol{\epsilon} - \boldsymbol{\beta}'\mathbf{X}'(\mathbf{I} - \mathbf{P}_m)'\mathbf{P}_m\boldsymbol{\epsilon} - \boldsymbol{\epsilon}'\mathbf{P}_m'(\mathbf{I} - \mathbf{P}_m)\mathbf{X}\boldsymbol{\beta} \\ &\quad + \boldsymbol{\beta}'\mathbf{X}'(\mathbf{I} - \mathbf{P}_m)(\mathbf{I} - \mathbf{P}_m)\mathbf{X}\boldsymbol{\beta} \\ &= \boldsymbol{\epsilon}'\mathbf{P}_m\boldsymbol{\epsilon} + \boldsymbol{\beta}'\mathbf{X}'(\mathbf{I} - \mathbf{P}_m)\mathbf{X}\boldsymbol{\beta}\end{aligned}$$

# Predictive MSE of $\mathbf{X}\hat{\boldsymbol{\beta}}_m$ relative to infeasible optimum $\mathbf{X}\boldsymbol{\beta}$

Step 3: Expectation of Step 2 conditional on  $\mathbf{X}$

$$\begin{aligned}\text{MSE}(m|\mathbf{X}) &= \mathbb{E} \left[ (\mathbf{X}\hat{\boldsymbol{\beta}}_m - \mathbf{X}\boldsymbol{\beta})'(\mathbf{X}\hat{\boldsymbol{\beta}}_m - \mathbf{X}\boldsymbol{\beta}) | \mathbf{X} \right] \\ &= \mathbb{E} [\boldsymbol{\epsilon}'\mathbf{P}_m\boldsymbol{\epsilon} | \mathbf{X}] + \mathbb{E} [\boldsymbol{\beta}'\mathbf{X}'(\mathbf{I} - \mathbf{P}_m)\mathbf{X}\boldsymbol{\beta} | \mathbf{X}]\end{aligned}$$

# Predictive MSE of $\mathbf{X}\hat{\boldsymbol{\beta}}_m$ relative to infeasible optimum $\mathbf{X}\boldsymbol{\beta}$

Step 3: Expectation of Step 2 conditional on  $\mathbf{X}$

$$\begin{aligned}\text{MSE}(m|\mathbf{X}) &= \mathbb{E} \left[ (\mathbf{X}\hat{\boldsymbol{\beta}}_m - \mathbf{X}\boldsymbol{\beta})' (\mathbf{X}\hat{\boldsymbol{\beta}}_m - \mathbf{X}\boldsymbol{\beta}) | \mathbf{X} \right] \\ &= \mathbb{E} [\boldsymbol{\epsilon}' \mathbf{P}_m \boldsymbol{\epsilon} | \mathbf{X}] + \mathbb{E} [\boldsymbol{\beta}' \mathbf{X}' (\mathbf{I} - \mathbf{P}_m) \mathbf{X} \boldsymbol{\beta} | \mathbf{X}] \\ &= \mathbb{E} [\text{tr} \{ \boldsymbol{\epsilon}' \mathbf{P}_m \boldsymbol{\epsilon} \} | \mathbf{X}] + \boldsymbol{\beta}' \mathbf{X}' (\mathbf{I} - \mathbf{P}_m) \mathbf{X} \boldsymbol{\beta}\end{aligned}$$

## Predictive MSE of $\mathbf{X}\hat{\boldsymbol{\beta}}_m$ relative to infeasible optimum $\mathbf{X}\boldsymbol{\beta}$

Step 3: Expectation of Step 2 conditional on  $\mathbf{X}$

$$\begin{aligned}\text{MSE}(m|\mathbf{X}) &= \mathbb{E} \left[ (\mathbf{X}\hat{\boldsymbol{\beta}}_m - \mathbf{X}\boldsymbol{\beta})'(\mathbf{X}\hat{\boldsymbol{\beta}}_m - \mathbf{X}\boldsymbol{\beta}) | \mathbf{X} \right] \\ &= \mathbb{E} [\boldsymbol{\epsilon}'\mathbf{P}_m\boldsymbol{\epsilon} | \mathbf{X}] + \mathbb{E} [\boldsymbol{\beta}'\mathbf{X}'(\mathbf{I} - \mathbf{P}_m)\mathbf{X}\boldsymbol{\beta} | \mathbf{X}] \\ &= \mathbb{E} [\text{tr} \{ \boldsymbol{\epsilon}'\mathbf{P}_m\boldsymbol{\epsilon} \} | \mathbf{X}] + \boldsymbol{\beta}'\mathbf{X}'(\mathbf{I} - \mathbf{P}_m)\mathbf{X}\boldsymbol{\beta} \\ &= \text{tr} \{ \mathbb{E}[\boldsymbol{\epsilon}\boldsymbol{\epsilon}' | \mathbf{X}]\mathbf{P}_m \} + \boldsymbol{\beta}'\mathbf{X}'(\mathbf{I} - \mathbf{P}_m)\mathbf{X}\boldsymbol{\beta}\end{aligned}$$



## Predictive MSE of $\mathbf{X}\hat{\boldsymbol{\beta}}_m$ relative to infeasible optimum $\mathbf{X}\boldsymbol{\beta}$

Step 3: Expectation of Step 2 conditional on  $\mathbf{X}$

$$\begin{aligned}\text{MSE}(m|\mathbf{X}) &= \mathbb{E} \left[ (\mathbf{X}\hat{\boldsymbol{\beta}}_m - \mathbf{X}\boldsymbol{\beta})'(\mathbf{X}\hat{\boldsymbol{\beta}}_m - \mathbf{X}\boldsymbol{\beta}) | \mathbf{X} \right] \\ &= \mathbb{E} [\boldsymbol{\epsilon}'\mathbf{P}_m\boldsymbol{\epsilon} | \mathbf{X}] + \mathbb{E} [\boldsymbol{\beta}'\mathbf{X}'(\mathbf{I} - \mathbf{P}_m)\mathbf{X}\boldsymbol{\beta} | \mathbf{X}] \\ &= \mathbb{E} [\text{tr} \{ \boldsymbol{\epsilon}'\mathbf{P}_m\boldsymbol{\epsilon} \} | \mathbf{X}] + \boldsymbol{\beta}'\mathbf{X}'(\mathbf{I} - \mathbf{P}_m)\mathbf{X}\boldsymbol{\beta} \\ &= \text{tr} \{ \mathbb{E}[\boldsymbol{\epsilon}\boldsymbol{\epsilon}' | \mathbf{X}]\mathbf{P}_m \} + \boldsymbol{\beta}'\mathbf{X}'(\mathbf{I} - \mathbf{P}_m)\mathbf{X}\boldsymbol{\beta} \\ &= \text{tr} \{ \sigma^2\mathbf{P}_m \} + \boldsymbol{\beta}'\mathbf{X}'(\mathbf{I} - \mathbf{P}_m)\mathbf{X}\boldsymbol{\beta}\end{aligned}$$

# Predictive MSE of $\mathbf{X}\hat{\boldsymbol{\beta}}_m$ relative to infeasible optimum $\mathbf{X}\boldsymbol{\beta}$

Step 3: Expectation of Step 2 conditional on  $\mathbf{X}$

$$\begin{aligned}\text{MSE}(m|\mathbf{X}) &= \mathbb{E} \left[ (\mathbf{X}\hat{\boldsymbol{\beta}}_m - \mathbf{X}\boldsymbol{\beta})' (\mathbf{X}\hat{\boldsymbol{\beta}}_m - \mathbf{X}\boldsymbol{\beta}) | \mathbf{X} \right] \\ &= \mathbb{E} [\boldsymbol{\epsilon}' \mathbf{P}_m \boldsymbol{\epsilon} | \mathbf{X}] + \mathbb{E} [\boldsymbol{\beta}' \mathbf{X}' (\mathbf{I} - \mathbf{P}_m) \mathbf{X} \boldsymbol{\beta} | \mathbf{X}] \\ &= \mathbb{E} [\text{tr} \{ \boldsymbol{\epsilon}' \mathbf{P}_m \boldsymbol{\epsilon} \} | \mathbf{X}] + \boldsymbol{\beta}' \mathbf{X}' (\mathbf{I} - \mathbf{P}_m) \mathbf{X} \boldsymbol{\beta} \\ &= \text{tr} \{ \mathbb{E} [\boldsymbol{\epsilon} \boldsymbol{\epsilon}' | \mathbf{X}] \mathbf{P}_m \} + \boldsymbol{\beta}' \mathbf{X}' (\mathbf{I} - \mathbf{P}_m) \mathbf{X} \boldsymbol{\beta} \\ &= \text{tr} \{ \sigma^2 \mathbf{P}_m \} + \boldsymbol{\beta}' \mathbf{X}' (\mathbf{I} - \mathbf{P}_m) \mathbf{X} \boldsymbol{\beta} \\ &= \sigma^2 k_m + \boldsymbol{\beta}' \mathbf{X}' (\mathbf{I} - \mathbf{P}_m) \mathbf{X} \boldsymbol{\beta}\end{aligned}$$

where  $k_m$  denotes the number of regressors in  $\mathbf{X}_m$  and

$$\text{tr}(\mathbf{P}_m) = \text{tr} \left\{ \mathbf{X}_m (\mathbf{X}_m' \mathbf{X}_m)^{-1} \mathbf{X}_m' \right\} = \text{tr} \left\{ \mathbf{X}_m' \mathbf{X}_m (\mathbf{X}_m' \mathbf{X}_m)^{-1} \right\} = \text{tr}(\mathbf{I}_m)$$

Now we know the MSE of a given model. . .

$$\text{MSE}(m|\mathbf{X}) = \sigma^2 k_m + \beta' \mathbf{X}' (\mathbf{I} - \mathbf{P}_m) \mathbf{X} \beta$$

Bias-Variance Tradeoff

Now we know the MSE of a given model. . .

$$\text{MSE}(m|\mathbf{X}) = \sigma^2 k_m + \beta' \mathbf{X}' (\mathbf{I} - \mathbf{P}_m) \mathbf{X} \beta$$

### Bias-Variance Tradeoff

- ▶ Smaller Model  $\Rightarrow \sigma^2 k_m$  smaller: less estimation uncertainty.

Now we know the MSE of a given model. . .

$$\text{MSE}(m|\mathbf{X}) = \sigma^2 k_m + \beta' \mathbf{X}' (\mathbf{I} - \mathbf{P}_m) \mathbf{X} \beta$$

### Bias-Variance Tradeoff

- ▶ Smaller Model  $\Rightarrow \sigma^2 k_m$  smaller: less estimation uncertainty.
- ▶ Bigger Model  $\Rightarrow \mathbf{X}'(\mathbf{I} - \mathbf{P}_m)\mathbf{X} = ||(\mathbf{I} - \mathbf{P}_m)\mathbf{X}||^2$  is in general smaller: less (squared) bias.

Now we know the MSE of a given model. . .

$$\text{MSE}(m|\mathbf{X}) = \sigma^2 k_m + \beta' \mathbf{X}' (\mathbf{I} - \mathbf{P}_m) \mathbf{X} \beta$$

### Bias-Variance Tradeoff

- ▶ Smaller Model  $\Rightarrow \sigma^2 k_m$  smaller: less estimation uncertainty.
- ▶ Bigger Model  $\Rightarrow \mathbf{X}' (\mathbf{I} - \mathbf{P}_m) \mathbf{X} = \|(\mathbf{I} - \mathbf{P}_m) \mathbf{X}\|^2$  is in general smaller: less (squared) bias.

### Mallow's $C_p$

- ▶ Problem: MSE formula is infeasible since it involves  $\beta$  and  $\sigma^2$ .

Now we know the MSE of a given model. . .

$$\text{MSE}(m|\mathbf{X}) = \sigma^2 k_m + \beta' \mathbf{X}' (\mathbf{I} - \mathbf{P}_m) \mathbf{X} \beta$$

### Bias-Variance Tradeoff

- ▶ Smaller Model  $\Rightarrow \sigma^2 k_m$  smaller: less estimation uncertainty.
- ▶ Bigger Model  $\Rightarrow \mathbf{X}' (\mathbf{I} - \mathbf{P}_m) \mathbf{X} = ||(\mathbf{I} - \mathbf{P}_m) \mathbf{X}||^2$  is in general smaller: less (squared) bias.

### Mallow's $C_p$

- ▶ Problem: MSE formula is infeasible since it involves  $\beta$  and  $\sigma^2$ .
- ▶ Solution: Mallow's  $C_p$  constructs an unbiased estimator.

Now we know the MSE of a given model...

$$\text{MSE}(m|\mathbf{X}) = \sigma^2 k_m + \beta' \mathbf{X}' (\mathbf{I} - \mathbf{P}_m) \mathbf{X} \beta$$

### Bias-Variance Tradeoff

- ▶ Smaller Model  $\Rightarrow \sigma^2 k_m$  smaller: less estimation uncertainty.
- ▶ Bigger Model  $\Rightarrow \mathbf{X}' (\mathbf{I} - \mathbf{P}_m) \mathbf{X} = ||(\mathbf{I} - \mathbf{P}_m) \mathbf{X}||^2$  is in general smaller: less (squared) bias.

### Mallow's $C_p$

- ▶ Problem: MSE formula is infeasible since it involves  $\beta$  and  $\sigma^2$ .
- ▶ Solution: Mallow's  $C_p$  constructs an unbiased estimator.
- ▶ Idea: what about plugging in  $\hat{\beta}$  to estimate second term?



What if we plug in  $\hat{\beta}$  to estimate the second term?

For the missing algebra in Step 4, see the lecture notes.

## Notation

Let  $\hat{\beta}$  denote the full model estimator and  $\mathbf{P}$  be the corresponding projection matrix:  $\mathbf{X}\hat{\beta} = \mathbf{P}\mathbf{y}$ .

## What if we plug in $\hat{\beta}$ to estimate the second term?

For the missing algebra in Step 4, see the lecture notes.

### Notation

Let  $\hat{\beta}$  denote the full model estimator and  $\mathbf{P}$  be the corresponding projection matrix:  $\mathbf{X}\hat{\beta} = \mathbf{P}\mathbf{y}$ .

### Crucial Fact

$\text{span}(\mathbf{X}_m)$  is a subspace of  $\text{span}(\mathbf{X})$ , so  $\mathbf{P}_m\mathbf{P} = \mathbf{P}\mathbf{P}_m = \mathbf{P}_m$ .

# What if we plug in $\hat{\beta}$ to estimate the second term?

For the missing algebra in Step 4, see the lecture notes.

## Notation

Let  $\hat{\beta}$  denote the full model estimator and  $\mathbf{P}$  be the corresponding projection matrix:  $\mathbf{X}\hat{\beta} = \mathbf{P}\mathbf{y}$ .

## Crucial Fact

$\text{span}(\mathbf{X}_m)$  is a subspace of  $\text{span}(\mathbf{X})$ , so  $\mathbf{P}_m\mathbf{P} = \mathbf{P}\mathbf{P}_m = \mathbf{P}_m$ .

## Step 4: Algebra using the preceding fact

$$\mathbb{E} \left[ \hat{\beta}' \mathbf{X}' (\mathbf{I} - \mathbf{P}_m) \mathbf{X} \hat{\beta} | \mathbf{X} \right] = \dots = \beta' \mathbf{X}' (\mathbf{I} - \mathbf{P}_m) \mathbf{X} \beta + \mathbb{E} \left[ \epsilon' (\mathbf{P} - \mathbf{P}_m) \epsilon | \mathbf{X} \right]$$

Substituting  $\hat{\beta}$  doesn't work...

Step 5: Use “Trace Trick” on second term from Step 4

$$\mathbb{E}[\epsilon'(\mathbf{P} - \mathbf{P}_m)\epsilon | \mathbf{X}] = \mathbb{E}[\text{tr} \{ \epsilon'(\mathbf{P} - \mathbf{P}_m)\epsilon \} | \mathbf{X}]$$

Substituting  $\hat{\beta}$  doesn't work...

Step 5: Use “Trace Trick” on second term from Step 4

$$\begin{aligned}\mathbb{E}[\epsilon'(\mathbf{P} - \mathbf{P}_m)\epsilon|\mathbf{X}] &= \mathbb{E}[\text{tr}\{\epsilon'(\mathbf{P} - \mathbf{P}_m)\epsilon\}|\mathbf{X}] \\ &= \text{tr}\{\mathbb{E}[\epsilon\epsilon'|\mathbf{X}](\mathbf{P} - \mathbf{P}_m)\}\end{aligned}$$

Substituting  $\hat{\beta}$  doesn't work...

Step 5: Use “Trace Trick” on second term from Step 4

$$\begin{aligned}\mathbb{E}[\epsilon'(\mathbf{P} - \mathbf{P}_m)\epsilon|\mathbf{X}] &= \mathbb{E}[\text{tr}\{\epsilon'(\mathbf{P} - \mathbf{P}_m)\epsilon\}|\mathbf{X}] \\ &= \text{tr}\{\mathbb{E}[\epsilon\epsilon'|\mathbf{X}](\mathbf{P} - \mathbf{P}_m)\} \\ &= \text{tr}\{\sigma^2(\mathbf{P} - \mathbf{P}_m)\}\end{aligned}$$

Substituting  $\hat{\beta}$  doesn't work...

Step 5: Use “Trace Trick” on second term from Step 4

$$\begin{aligned}\mathbb{E}[\epsilon'(\mathbf{P} - \mathbf{P}_m)\epsilon|\mathbf{X}] &= \mathbb{E}[\text{tr}\{\epsilon'(\mathbf{P} - \mathbf{P}_m)\epsilon\}|\mathbf{X}] \\ &= \text{tr}\{\mathbb{E}[\epsilon\epsilon'|\mathbf{X}](\mathbf{P} - \mathbf{P}_m)\} \\ &= \text{tr}\{\sigma^2(\mathbf{P} - \mathbf{P}_m)\} \\ &= \sigma^2(\text{trace}\{\mathbf{P}\} - \text{trace}\{\mathbf{P}_m\})\end{aligned}$$

Substituting  $\hat{\beta}$  doesn't work...

Step 5: Use “Trace Trick” on second term from Step 4

$$\begin{aligned}\mathbb{E}[\epsilon'(\mathbf{P} - \mathbf{P}_m)\epsilon|\mathbf{X}] &= \mathbb{E}[\text{tr}\{\epsilon'(\mathbf{P} - \mathbf{P}_m)\epsilon\}|\mathbf{X}] \\ &= \text{tr}\{\mathbb{E}[\epsilon\epsilon'|\mathbf{X}](\mathbf{P} - \mathbf{P}_m)\} \\ &= \text{tr}\{\sigma^2(\mathbf{P} - \mathbf{P}_m)\} \\ &= \sigma^2(\text{trace}\{\mathbf{P}\} - \text{trace}\{\mathbf{P}_m\}) \\ &= \sigma^2(K - k_m)\end{aligned}$$

where  $K$  is the total number of regressors in  $\mathbf{X}$



## Substituting $\hat{\beta}$ doesn't work...

Step 5: Use “Trace Trick” on second term from Step 4

$$\begin{aligned}\mathbb{E}[\epsilon'(\mathbf{P} - \mathbf{P}_m)\epsilon|\mathbf{X}] &= \mathbb{E}[\text{tr}\{\epsilon'(\mathbf{P} - \mathbf{P}_m)\epsilon\}|\mathbf{X}] \\&= \text{tr}\{\mathbb{E}[\epsilon\epsilon'|\mathbf{X}](\mathbf{P} - \mathbf{P}_m)\} \\&= \text{tr}\{\sigma^2(\mathbf{P} - \mathbf{P}_m)\} \\&= \sigma^2(\text{trace}\{\mathbf{P}\} - \text{trace}\{\mathbf{P}_m\}) \\&= \sigma^2(K - k_m)\end{aligned}$$

where  $K$  is the total number of regressors in  $\mathbf{X}$

### Bias of Plug-in Estimator

$$\mathbb{E}\left[\hat{\beta}'\mathbf{X}'(\mathbf{I} - \mathbf{P}_m)\mathbf{X}\hat{\beta}|\mathbf{X}\right] = \underbrace{\beta'\mathbf{X}'(\mathbf{I} - \mathbf{P}_m)\mathbf{X}\beta}_{\text{Truth}} + \underbrace{\sigma^2(K - k_m)}_{\text{Bias}}$$

## Putting Everything Together: Mallows's $C_p$

Want An Unbiased Estimator of This:

$$\text{MSE}(m|\mathbf{X}) = \sigma^2 k_m + \boldsymbol{\beta}' \mathbf{X}' (\mathbf{I} - \mathbf{P}_m) \mathbf{X} \boldsymbol{\beta}$$

Previous Slide:

$$\mathbb{E} \left[ \widehat{\boldsymbol{\beta}}' \mathbf{X}' (\mathbf{I} - \mathbf{P}_m) \mathbf{X} \widehat{\boldsymbol{\beta}} | \mathbf{X} \right] = \boldsymbol{\beta}' \mathbf{X}' (\mathbf{I} - \mathbf{P}_m) \mathbf{X} \boldsymbol{\beta} + \sigma^2 (K - k_m)$$

## Putting Everything Together: Mallows's $C_p$

Want An Unbiased Estimator of This:

$$\text{MSE}(m|\mathbf{X}) = \sigma^2 k_m + \boldsymbol{\beta}' \mathbf{X}' (\mathbf{I} - \mathbf{P}_m) \mathbf{X} \boldsymbol{\beta}$$

Previous Slide:

$$\mathbb{E} \left[ \hat{\boldsymbol{\beta}}' \mathbf{X}' (\mathbf{I} - \mathbf{P}_m) \mathbf{X} \hat{\boldsymbol{\beta}} | \mathbf{X} \right] = \boldsymbol{\beta}' \mathbf{X}' (\mathbf{I} - \mathbf{P}_m) \mathbf{X} \boldsymbol{\beta} + \sigma^2 (K - k_m)$$

End Result:

$$\begin{aligned} \text{MC}(m) &= \hat{\sigma}^2 k_m + \left[ \hat{\boldsymbol{\beta}}' \mathbf{X}' (\mathbf{I} - \mathbf{P}_m) \mathbf{X} \hat{\boldsymbol{\beta}} - \hat{\sigma}^2 (K - k_m) \right] \\ &= \hat{\boldsymbol{\beta}}' \mathbf{X}' (\mathbf{I} - \mathbf{P}_m) \mathbf{X} \hat{\boldsymbol{\beta}} + \hat{\sigma}^2 (2k_m - K) \end{aligned}$$

is an unbiased estimator of MSE, with  $\hat{\sigma}^2 = \mathbf{y}'(\mathbf{I} - \mathbf{P})\mathbf{y}/(T - K)$

## Why is this different from the textbook formula?

Just algebra, but tedious. . .

$$\begin{aligned}\text{MC}(m) - 2\hat{\sigma}^2 k_m &= \hat{\beta}' X' (\mathbf{I} - P_M) X \hat{\beta} - K \hat{\sigma}^2 \\ &\vdots \\ &= \mathbf{y}' (\mathbf{I} - P_M) \mathbf{y} - T \hat{\sigma}^2 \\ &= \text{RSS}(m) - T \hat{\sigma}^2\end{aligned}$$

## Why is this different from the textbook formula?

Just algebra, but tedious. . .

$$\begin{aligned}\text{MC}(m) - 2\hat{\sigma}^2 k_m &= \hat{\beta}' X' (\mathbf{I} - P_M) X \hat{\beta} - K \hat{\sigma}^2 \\ &\vdots \\ &= \mathbf{y}' (\mathbf{I} - P_M) \mathbf{y} - T \hat{\sigma}^2 \\ &= \text{RSS}(m) - T \hat{\sigma}^2\end{aligned}$$

Therefore:

$$\text{MC}(m) = \text{RSS}(m) + \hat{\sigma}^2(2k_m - T)$$

## Why is this different from the textbook formula?

Just algebra, but tedious. . .

$$\begin{aligned}\text{MC}(m) - 2\hat{\sigma}^2 k_m &= \hat{\beta}' X' (\mathbf{I} - P_M) X \hat{\beta} - K \hat{\sigma}^2 \\ &\vdots \\ &= \mathbf{y}' (\mathbf{I} - P_M) \mathbf{y} - T \hat{\sigma}^2 \\ &= \text{RSS}(m) - T \hat{\sigma}^2\end{aligned}$$

Therefore:

$$\text{MC}(m) = \text{RSS}(m) + \hat{\sigma}^2(2k_m - T)$$

Divide Through by  $\hat{\sigma}^2$ :

$$C_p(m) = \frac{\text{RSS}(m)}{\hat{\sigma}^2} + 2k_m - T$$

Tells us how to adjust RSS for number of regressors. . .

# Bayesian Model Comparison: Marginal Likelihoods

Bayes' Rule for Model  $m \in \mathcal{M}$

$$\underbrace{\pi(\boldsymbol{\theta}|\mathbf{y}, m)}_{\text{Posterior}} \propto \underbrace{\pi(\boldsymbol{\theta}|m)}_{\text{Prior}} \underbrace{f(\mathbf{y}|\boldsymbol{\theta}, m)}_{\text{Likelihood}}$$

$$\underbrace{f(\mathbf{y}|m)}_{\text{Marginal Likelihood}} = \int_{\Theta} \pi(\boldsymbol{\theta}|m) f(\mathbf{y}|\boldsymbol{\theta}, m) \, d\boldsymbol{\theta}$$

# Bayesian Model Comparison: Marginal Likelihoods

Bayes' Rule for Model  $m \in \mathcal{M}$

$$\underbrace{\pi(\boldsymbol{\theta}|\mathbf{y}, m)}_{\text{Posterior}} \propto \underbrace{\pi(\boldsymbol{\theta}|m)}_{\text{Prior}} \underbrace{f(\mathbf{y}|\boldsymbol{\theta}, m)}_{\text{Likelihood}}$$
$$\underbrace{f(\mathbf{y}|m)}_{\text{Marginal Likelihood}} = \int_{\Theta} \pi(\boldsymbol{\theta}|m) f(\mathbf{y}|\boldsymbol{\theta}, m) \, d\boldsymbol{\theta}$$

Posterior Model Probability for  $m \in \mathcal{M}$

$$P(m|\mathbf{y}) = \frac{P(m)f(\mathbf{y}|m)}{f(\mathbf{y})} =$$



# Bayesian Model Comparison: Marginal Likelihoods

Bayes' Rule for Model  $m \in \mathcal{M}$

$$\underbrace{\pi(\boldsymbol{\theta}|\mathbf{y}, m)}_{\text{Posterior}} \propto \underbrace{\pi(\boldsymbol{\theta}|m)}_{\text{Prior}} \underbrace{f(\mathbf{y}|\boldsymbol{\theta}, m)}_{\text{Likelihood}}$$
$$\underbrace{f(\mathbf{y}|m)}_{\text{Marginal Likelihood}} = \int_{\Theta} \pi(\boldsymbol{\theta}|m) f(\mathbf{y}|\boldsymbol{\theta}, m) \, d\boldsymbol{\theta}$$

Posterior Model Probability for  $m \in \mathcal{M}$

$$P(m|\mathbf{y}) = \frac{P(m)f(\mathbf{y}|m)}{f(\mathbf{y})} = \frac{\int_{\Theta} P(m)f(\mathbf{y}, \boldsymbol{\theta}|m) \, d\boldsymbol{\theta}}{f(\mathbf{y})} =$$

# Bayesian Model Comparison: Marginal Likelihoods

## Bayes' Rule for Model $m \in \mathcal{M}$

$$\underbrace{\pi(\boldsymbol{\theta}|\mathbf{y}, m)}_{\text{Posterior}} \propto \underbrace{\pi(\boldsymbol{\theta}|m)}_{\text{Prior}} \underbrace{f(\mathbf{y}|\boldsymbol{\theta}, m)}_{\text{Likelihood}}$$
$$\underbrace{f(\mathbf{y}|m)}_{\text{Marginal Likelihood}} = \int_{\Theta} \pi(\boldsymbol{\theta}|m) f(\mathbf{y}|\boldsymbol{\theta}, m) \, d\boldsymbol{\theta}$$

## Posterior Model Probability for $m \in \mathcal{M}$

$$P(m|\mathbf{y}) = \frac{P(m)f(\mathbf{y}|m)}{f(\mathbf{y})} = \frac{\int_{\Theta} P(m)f(\mathbf{y}, \boldsymbol{\theta}|m) \, d\boldsymbol{\theta}}{f(\mathbf{y})} = \frac{P(m)}{f(\mathbf{y})} \int_{\Theta} \pi(\boldsymbol{\theta}|m)f(\mathbf{y}|\boldsymbol{\theta}, m) \, d\boldsymbol{\theta}$$

where  $P(m)$  is the **prior model probability** and  $f(\mathbf{y})$  is constant across models.

# Laplace (aka Saddlepoint) Approximation

Suppress model index  $m$  for simplicity.

General Case: for  $T$  large...

$$\int_{\Theta} g(\boldsymbol{\theta}) \exp\{T \cdot h(\boldsymbol{\theta})\} d\boldsymbol{\theta} \approx \left(\frac{2\pi}{T}\right)^{p/2} \exp\{T \cdot h(\boldsymbol{\theta}_0)\} g(\boldsymbol{\theta}_0) |H(\boldsymbol{\theta}_0)|^{-1/2}$$

$$p = \dim(\boldsymbol{\theta}), \quad \boldsymbol{\theta}_0 = \arg \max_{\boldsymbol{\theta} \in \Theta} h(\boldsymbol{\theta}), \quad H(\boldsymbol{\theta}_0) = -\frac{\partial^2 h(\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} \Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0}$$

# Laplace (aka Saddlepoint) Approximation

Suppress model index  $m$  for simplicity.

General Case: for  $T$  large...

$$\int_{\Theta} g(\boldsymbol{\theta}) \exp\{T \cdot h(\boldsymbol{\theta})\} d\boldsymbol{\theta} \approx \left(\frac{2\pi}{T}\right)^{p/2} \exp\{T \cdot h(\boldsymbol{\theta}_0)\} g(\boldsymbol{\theta}_0) |H(\boldsymbol{\theta}_0)|^{-1/2}$$

$$p = \dim(\boldsymbol{\theta}), \quad \boldsymbol{\theta}_0 = \arg \max_{\boldsymbol{\theta} \in \Theta} h(\boldsymbol{\theta}), \quad H(\boldsymbol{\theta}_0) = -\frac{\partial^2 h(\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} \bigg|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0}$$

Use to Approximate Marginal Likelihood

$$h(\boldsymbol{\theta}) = \frac{\ell(\boldsymbol{\theta})}{T} = \frac{1}{T} \sum_{t=1}^T \log f(Y_t | \boldsymbol{\theta}), \quad H(\boldsymbol{\theta}) = J_T(\boldsymbol{\theta}) = -\frac{1}{T} \sum_{t=1}^T \frac{\partial^2 \log f(Y_t | \boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'}, \quad g(\boldsymbol{\theta}) = \pi(\boldsymbol{\theta})$$

and substitute  $\hat{\boldsymbol{\theta}}_{MLE}$  for  $\boldsymbol{\theta}_0$

# Laplace Approximation to Marginal Likelihood

Suppress model index  $m$  for simplicity.

$$\int_{\Theta} \pi(\boldsymbol{\theta}) f(\mathbf{y}|\boldsymbol{\theta}) \, d\boldsymbol{\theta} \approx \left(\frac{2\pi}{T}\right)^{p/2} \exp\left\{\ell(\hat{\boldsymbol{\theta}}_{MLE})\right\} \pi(\hat{\boldsymbol{\theta}}_{MLE}) \left|J_T(\hat{\boldsymbol{\theta}}_{MLE})\right|^{-1/2}$$

$$\ell(\boldsymbol{\theta}) = \sum_{t=1}^T \log f(Y_t|\boldsymbol{\theta}), \quad H(\boldsymbol{\theta}) = J_T(\boldsymbol{\theta}) = -\frac{1}{T} \sum_{t=1}^T \frac{\partial^2 \log f(Y_t|\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'}$$

## Bayesian Information Criterion

$$\int_{\Theta} \pi(\boldsymbol{\theta}) f(\mathbf{y}|\boldsymbol{\theta}) \, d\boldsymbol{\theta} \approx \left( \frac{2\pi}{T} \right)^{p/2} \exp \left\{ \ell(\hat{\boldsymbol{\theta}}_{MLE}) \right\} \pi(\hat{\boldsymbol{\theta}}_{MLE}) \left| J_T(\hat{\boldsymbol{\theta}}_{MLE}) \right|^{-1/2}$$

## Bayesian Information Criterion

$$\int_{\Theta} \pi(\boldsymbol{\theta}) f(\mathbf{y}|\boldsymbol{\theta}) \, d\boldsymbol{\theta} \approx \left( \frac{2\pi}{T} \right)^{p/2} \exp \left\{ \ell(\hat{\boldsymbol{\theta}}_{MLE}) \right\} \pi(\hat{\boldsymbol{\theta}}_{MLE}) \left| J_T(\hat{\boldsymbol{\theta}}_{MLE}) \right|^{-1/2}$$

Take Logs and Multiply by 2

$$2 \log f(\mathbf{y}|\boldsymbol{\theta}) \approx \underbrace{2\ell(\hat{\boldsymbol{\theta}}_{MLE})}_{O_p(T)} - \underbrace{p \log(T)}_{O(\log T)} + \underbrace{p \log(2\pi) + \log \pi(\hat{\boldsymbol{\theta}}) - \log |J_T(\hat{\boldsymbol{\theta}})|}_{O_p(1)}$$

# Bayesian Information Criterion

$$\int_{\Theta} \pi(\boldsymbol{\theta}) f(\mathbf{y}|\boldsymbol{\theta}) \, d\boldsymbol{\theta} \approx \left( \frac{2\pi}{T} \right)^{p/2} \exp \left\{ \ell(\hat{\boldsymbol{\theta}}_{MLE}) \right\} \pi(\hat{\boldsymbol{\theta}}_{MLE}) \left| J_T(\hat{\boldsymbol{\theta}}_{MLE}) \right|^{-1/2}$$

Take Logs and Multiply by 2

$$2 \log f(\mathbf{y}|\boldsymbol{\theta}) \approx \underbrace{2\ell(\hat{\boldsymbol{\theta}}_{MLE})}_{O_p(T)} - \underbrace{p \log(T)}_{O(\log T)} + \underbrace{p \log(2\pi) + \log \pi(\hat{\boldsymbol{\theta}}) - \log |J_T(\hat{\boldsymbol{\theta}})|}_{O_p(1)}$$

The BIC

Assume uniform prior over **models** and ignore lower order terms:

$$\text{BIC}(m) = 2 \log f(\mathbf{y}|\hat{\boldsymbol{\theta}}, m) - p_m \log(T)$$

large-sample Frequentist approx. to Bayesian marginal likelihood