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.

Definition

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

Properties

- *Not* symmetric: $KL(g; f) \neq KL(f; g)$

- By Jensen's Inequality: $KL(g; f) \geq 0$ (strict iff $g = f$ a.e.)

- Minimize KL $\iff$ Maximize Expected log-likelihood

KL Divergence and Mis-specified MLE

Pseudo-true Parameter Value θ_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_θ is correctly specified?

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

Goal: Compare Mis-specified Models

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

where θ_0 is the pseudo-true parameter value for f_θ and γ_0 is the pseudo-true parameter value for h_γ.

How to Estimate Expected Log Likelihood?

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

Unbiased but Infeasible

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

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)

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 θ_0. For simplicity $Y_1, \ldots, Y_T \sim$ iid $g(y)$.

Fundamental Expansion

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

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

Central Limit Theorem

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

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

Information Matrix Equality

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

Bias Relative to Infeasible Plug-in Estimator

Definition of Bias Term B

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

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

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

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

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

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

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

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

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 $J, K \Rightarrow$ 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 $J = K \Rightarrow \text{tr}\left\{J^{-1}K\right\} = p \Rightarrow$ 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 (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 (β_1, σ_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 (β_1, σ_1^2) for candidate model.
   - Easy: just use MLE $(\widehat{\beta}_1, \widehat{\sigma}_1^2)$

2. Would need to know (β_0, σ_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 $(\widehat{\beta}, \widehat{\sigma}^2)$ is consistent for truth.

Deriving the Corrected AIC

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

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

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

Lecture #2 – More on "Classical" Model Selection

Mallow's C_p

Bayesian Model Comparison

Laplace Approximation

Bayesian Information Criterion (BIC)

Motivation: Predict **y** from **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 $\boldsymbol{\beta}$ is known.

Notation

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

In-sample versus Out-of-sample Prediction Error

Why not compare RSS(m)?

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

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 $\mathbf{X}\widehat{\beta}_m$ relative to infeasible optimum $\mathbf{X}\beta$

Step 1: Algebra

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

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

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

Predictive MSE of $\mathbf{X}\widehat{\beta}_m$ relative to infeasible optimum $\mathbf{X}\beta$

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

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

where k_m denotes the number of regressors in $\mathbf{X}_m$ and
$\mathsf{tr}(\mathbf{P}_m) = \mathsf{tr}\left\{\mathbf{X}_m\left(\mathbf{X}_m'\mathbf{X}_m\right)^{-1}\mathbf{X}_m'\right\} = \mathsf{tr}\left\{\mathbf{X}_m'\mathbf{X}_m\left(\mathbf{X}_m'\mathbf{X}_m\right)^{-1}\right\} = \mathsf{tr}(\mathbf{I}_m)$

Now we know the MSE of a given model…

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

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 $\boldsymbol{\beta}$ and σ^2.

- Solution: Mallow's C_p constructs an unbiased estimator.

- Idea: what about plugging in $\widehat{\boldsymbol{\beta}}$ to estimate second term?

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

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

Notation

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

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

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

Substituting $\widehat{\boldsymbol{\beta}}$ doesn't work...

Step 5: Use "Trace Trick" on second term from Step 4

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

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

Bias of Plug-in Estimator

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

Putting Everything Together: Mallow's C_p

Want An Unbiased Estimator of This:

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

Previous Slide:

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

End Result:

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

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

Why is this different from the textbook formula?

Just algebra, but tedious. . .

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

Therefore:

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

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

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

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

Bayesian Model Comparison: Marginal Likelihoods

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

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

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

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

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

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. . .

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

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

Use to Approximate Marginal Likelihood

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

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

Laplace Approximation to Marginal Likelihood

Suppress model index m for simplicity.

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

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

Bayesian Information Criterion

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

Take Logs and Multiply by 2

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

The BIC

Assume uniform prior over models and ignore lower order terms:

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

large-sample Frequentist approx. to Bayesian marginal likelihood

Lecture #3 – Cross-Validation

Model selection via a Hold-out Sample

K-fold Cross-validation

Asymptotic Equivalence Between LOO-CV and TIC

Influence Functions

Model Selection using a Hold-out Sample

- The real problem is double use of the data: first for estimation, then for model comparison.
    - Maximized sample log-likelihood is an overly optimistic estimate of expected log-likelihood and hence KL-divergence
    - In-sample squared prediction error is an overly optimistic estimator of out-of-sample squared prediction error
- AIC/TIC, AIC_c, BIC, C_p penalize sample log-likelihood or RSS to compensate.
- Another idea: don't re-use the same data!

Hold-out Sample: Partition the Full Dataset

Unfortunately this is extremely wasteful of data...

K-fold Cross-Validation: "Pseudo-out-of-sample"

Step 1

Randomly partition full dataset into K folds of approx. equal size.

Step 2

Treat k^{th} fold as a hold-out sample and estimate model using all observations except those in fold k: yielding estimator $\widehat{\theta}(-k)$.

K-fold Cross-Validation: "Pseudo-out-of-sample"

Step 2

Treat k^{th} fold as a hold-out sample and estimate model using all observations except those in fold k: yielding estimator $\widehat{\theta}(-k)$.

Step 3

Repeat Step 2 for each $k = 1, \ldots, K$.

Step 4

For each t calculate the prediction $\widehat{y}_t^{-k(t)}$ of y_t based on $\widehat{\theta}(-k(t))$, the estimator that excluded observation t.

K-fold Cross-Validation: "Pseudo-out-of-sample"

Step 4

For each t calculate the prediction $\widehat{y}_t^{-k(t)}$ of y_t based on $\widehat{\theta}(-k(t))$, the estimator that excluded observation t.

Step 5

Define $\text{CV}_K = \frac{1}{T} \sum_{t=1}^{T} L\left(y_t, \widehat{y}_t^{-k(t)}\right)$ where L is a loss function.

Step 5

Repeat for each model & choose m to minimize $\text{CV}_K(m)$.

CV uses each observation for parameter estimation and model evaluation but never at the same time!

Cross-Validation (CV): Some Details

Which Loss Function?

- ▶ For regression squared error loss makes sense

- ▶ For classification (discrete prediction) could use zero-one loss.

- ▶ Can also use log-likelihood/KL-divergence as a loss function...

How Many Folds?

- ▶ One extreme: $K = 2$. Closest to Training/Test idea.

- ▶ Other extreme: $K = T$ Leave-one-out CV (LOO-CV).

- ▶ Computationally expensive model $\Rightarrow$ may prefer fewer folds.

- ▶ If your model is a linear smoother there's a computational trick that makes LOO-CV extremely fast. (Problem Set)

- ▶ Asymptotic properties are related to K...

Relationship between LOO-CV and TIC

Theorem

LOO-CV using KL-divergence as the loss function is asymptotically equivalent to TIC but doesn't require us to estimate the Hessian and variance of the score.

Large-sample Equivalence of LOO-CV and TIC

Notation and Assumptions

For simplicity let $Y_1, \ldots, Y_T \sim$ iid. Let $\widehat{\theta}_{(t)}$ be the maximum likelihood estimator based on all observations except t and $\widehat{\theta}$ be the full-sample estimator.

Log-likelihood as "Loss"

$CV_1 = \frac{1}{T} \sum_{t=1}^{T} \log f(y_t | \widehat{\theta}_{(t)})$ but since min. KL = max. log-like. we choose the model with highest $CV_1(m)$.

Overview of the Proof

First-Order Taylor Expansion of $\widehat{\theta}_{(t)}$ around $\widehat{\theta}$:

$$
\begin{aligned}
CV_1 &= \frac{1}{T} \sum_{t=1}^{T} \log f(y_t | \widehat{\theta}_{(t)}) \\
&= \frac{1}{T} \sum_{t=1}^{T} \left[\log f(y_t | \widehat{\theta}) + \frac{\partial \log f(y_t | \widehat{\theta})}{\partial \theta'} \left(\widehat{\theta}_{(t)} - \widehat{\theta} \right) \right] + o_p(1) \\
&= \frac{\ell(\widehat{\theta})}{T} + \frac{1}{T} \sum_{t=1}^{T} \frac{\partial \log f(y_t | \widehat{\theta})}{\partial \theta'} \left(\widehat{\theta}_{(t)} - \widehat{\theta} \right) + o_p(1)
\end{aligned}
$$

Crucial point: the first-order term is not zero in this case. (Why?)

Overview of Proof

From expansion on previous slide, we simply need to show that:

$$\frac{1}{T} \sum_{t=1}^{T} \frac{\partial \log f(y_t|\widehat{\theta})}{\partial \theta'} \left(\widehat{\theta}_{(t)} - \widehat{\theta} \right) = -\frac{1}{T} \text{tr} \left(\widehat{J}^{-1}\widehat{K} \right) + o_p(1)$$

$$\widehat{K} = \frac{1}{T} \sum_{t=1}^{T} \left(\frac{\partial \log f(y_t|\widehat{\theta})}{\partial \theta} \right) \left(\frac{\partial \log f(y_t|\widehat{\theta})}{\partial \theta} \right)'$$

$$\widehat{J} = -\frac{1}{T} \sum_{t=1}^{T} \frac{\partial \log f(y_t|\widehat{\theta})}{\partial \theta \partial \theta'}$$

Overview of Proof

By the definition of $\widehat{K}$ and the properties of the trace operator:

$$
\begin{aligned}
-\frac{1}{T}\text{tr}\left\{\widehat{J}^{-1}\widehat{K}\right\} &= -\frac{1}{T}\text{tr}\left\{\widehat{J}^{-1}\left[\frac{1}{T}\sum_{t=1}^{T}\left(\frac{\partial \log f(y_t|\widehat{\theta})}{\partial \theta}\right)\left(\frac{\partial \log f(y_t|\widehat{\theta})}{\partial \theta}\right)'\right]\right\} \\
&= \left[\frac{1}{T}\sum_{t=1}^{T}\text{tr}\left\{\frac{-\widehat{J}^{-1}}{T}\left(\frac{\partial \log f(y_t|\widehat{\theta})}{\partial \theta}\right)\left(\frac{\partial \log f(y_t|\widehat{\theta})}{\partial \theta}\right)'\right\}\right] \\
&= \frac{1}{T}\sum_{t=1}^{T}\frac{\partial \log f(y_t|\widehat{\theta})}{\partial \theta'}\left(-\frac{1}{T}\widehat{J}^{-1}\right)\frac{\partial \log f(y_t|\widehat{\theta})}{\partial \theta}
\end{aligned}
$$

So it suffices to show that

$$
\left(\widehat{\theta}_{(t)} - \widehat{\theta}\right) = -\frac{1}{T}\widehat{J}^{-1}\left[\frac{\partial \log f(y_t|\widehat{\theta})}{\partial \theta}\right] + o_p(1)
$$

Digression: Functionals and Influence Functions

(Statistical) Functional

$\mathbb{T} = \mathbb{T}(G)$ maps a CDF G to $\mathbb{R}^p$.

Example: ML Estimation

$$\theta_0 = \mathbb{T}(G) = \underset{\theta \in \Theta}{\arg \min} \; E_G \left[\log \left\{ \frac{g(Y)}{f(Y|\theta)} \right\} \right]$$

Influence Function

Let δ_y be a point mass at y: $\delta_y(y) = 1$, $\delta_y(y') = 0$ for $y' \neq y$.

Influence function = functional derivative: how does a small change in G affect $\mathbb{T}$?

$$\text{infl}(G, y) = \lim_{\epsilon \to 0} \frac{\mathbb{T}\left[(1 - \epsilon) \, G + \epsilon \delta_y\right] - \mathbb{T}(G)}{\epsilon}$$

Back to the Proof. . .

Step 1

The influence function for ML estimation turns out to be
$\text{infl}(G, y) = J^{-1} \frac{\partial}{\partial \theta} \log f(y|\theta_0)$.

Step 2

Let $\widehat{G}$ denote the empirical CDF based on $y_1, \ldots, y_T$. Then:

$$\left(\widehat{\theta}_{(t)} - \widehat{\theta} \right) = -\frac{1}{T} \text{infl}(\widehat{G}, y_t) + o_p(1)$$

Step 3

Evaluating Step 1 at $\widehat{G}$ and substituting into Step 2

$$\left(\widehat{\theta}_{(t)} - \widehat{\theta} \right) = -\frac{1}{T} \widehat{J}^{-1} \left[\frac{\partial \log f(y_t|\widehat{\theta})}{\partial \theta} \right] + o_p(1)$$

Lecture #4 – Asymptotic Properties

Overview

Weak Consistency

Consistency

Efficiency

AIC versus BIC in a Simple Example

Overview

- What happens as $T \to \infty$?

- Consistency: choose "best" model wpa 1

- Efficiency: procedure with good risk properties

- Can't have both at once.

- Large, fairly technical literature: only a brief overview today.

- More details: Sin and White (1992, 1996), Pötscher (1991), Leeb & Pötscher (2005), Yang (2005) and Yang (2007).

Penalizing the Likelihood

Examples we've seen:

$$TIC = 2\ell_T(\widehat{\theta}) - \text{trace}\left\{\widehat{J}^{-1}\widehat{K}\right\}$$

$$AIC = 2\ell_T(\widehat{\theta}) - 2\,\text{length}(\theta)$$

$$BIC = 2\ell_T(\widehat{\theta}) - \log(T)\,\text{length}(\theta)$$

Generic penalty $c_{T,k}$

$$IC(M_k) = 2\sum_{t=1}^{T}\log f_{k,t}(Y_t|\widehat{\theta}_k) - c_{T,k}$$

How does choice of $c_{T,k}$ affect behavior of the criterion?

Weak Consistency: Suppose M_{k_0} Uniquely Minimizes KL

Assumption

$$\liminf_{T \to \infty} \left(\min_{k \neq k_0} \frac{1}{T} \sum_{t=1}^{T} \{KL(g; f_{k,t}) - KL(g; f_{k_0,t})\} \right) > 0$$

Consequences

▶ Any criterion with $c_{T,k} > 0$ and $c_{T,k} = o_p(T)$ is weakly consistent: selects M_{k_0} wpa 1 in the limit.

▶ Weak consistency still holds if $c_{T,k}$ is zero for one of the models, so long as it is strictly positive for all the others.

Both AIC and BIC are Weakly Consistent

Both satisfy $T^{-1} c_{T,k} \xrightarrow{p} 0$.

$$\text{BIC Penalty:} \qquad c_{T,k} = \log(T) \times \text{length}(\theta_k)$$

$$\text{AIC Penalty:} \qquad c_{T,k} = 2 \times \text{length}(\theta_k)$$

Consistency: No Unique KL-minimizer

Example

If the truth is an AR(5) model then AR(6), AR(7), AR(8), etc. models all have zero KL-divergence.

Principle of Parsimony

Among the KL-minimizers, choose the simplest model, i.e. the one with the fewest parameters.

Notation

$\mathcal{J}$ = be the set of all models that attain minimum KL-divergence

$\mathcal{J}_0$ = subset with the minimum number of parameters.

Sufficient Conditions for Consistency

Consistency: Select Model from $\mathcal{J}_0$ wpa 1

$$\lim_{T \to \infty} \mathbb{P} \left\{ \min_{\ell \in \mathcal{J} \setminus \mathcal{J}_0} [IC(M_{j_0}) - IC(M_\ell)] > 0 \right\} = 1$$

Sufficient Conditions

(i) For all $k \neq \ell \in \mathcal{J}$

$$\sum_{t=1}^{T} [\log f_{k,t}(Y_t|\theta_k^*) - \log f_{\ell,t}(Y_t|\theta_\ell^*)] = O_p(1)$$

where θ_k^* and θ_ℓ^* are the KL minimizing parameter values.

(ii) For all $j_0 \in \mathcal{J}_0$ and $\ell \in (\mathcal{J} \setminus \mathcal{J}_0)$

$$P\left(c_{T,\ell} - c_{T,j_0} \to \infty\right) = 1$$

BIC is Consistent; AIC and TIC Are Not

- ► AIC and TIC *cannot* satisfy (ii) since $(c_{T,\ell} - c_{T,j_0})$ *does not depend on sample size*.

- ► It turns out that AIC and TIC are *not* consistent.

- ► BIC is consistent:

$$c_{T,\ell} - c_{T,j_0} = \log(T) \left\{ \text{length}(\theta_\ell) - \text{length}(\theta_{j_0}) \right\}$$

- ► Term in braces is *positive* since $\ell \in \mathcal{J} \backslash \mathcal{J}_0$, i.e. ℓ is not as parsimonious as j_0

- ► $\log(T) \to \infty$, so BIC always selects a model in $\mathcal{J}_0$ in the limit.

Efficiency

- ▶ Roughly speaking, a model selection criterion is called efficient if it performs "nearly as well" as the theoretical optimum relative to some loss function.

- ▶ More broadly, an efficient/conservative criterion is one that has "good risk properties."

- ▶ We don't have time to go into detail, so we'll look at a particular example. . .

Consistency versus Efficiency in a Simple Example

Information Criteria

Consider criteria of the form $IC_m = 2\ell(\theta) - d_T \times \text{length}(\theta)$.

True DGP

$Y_1, \ldots, Y_T \sim \text{iid } N(\mu, 1)$

Candidate Models

M_0 assumes $\mu = 0$, M_1 does not restrict μ. Only one parameter:

$$IC_0 = 2 \max_\mu \{\ell(\mu): M_0\}$$

$$IC_1 = 2 \max_\mu \{\ell(\mu): M_1\} - d_T$$

Log-Likelihood Function

Since $\sum_{t=1}^{T}(Y_t - \mu)^2 = T(\bar{Y} - \mu)^2 + T\hat{\sigma}^2$,

$$
\begin{aligned}
\ell_T(\mu) &= \sum_{t=1}^{T} \log\left(\frac{1}{2\pi} \exp\left\{-\frac{1}{2}(Y_t - \mu)^2\right\}\right) \\
&= -\frac{T}{2}\log(2\pi) - \frac{1}{2}\sum_{t=1}^{T}(Y_t - \mu)^2 \\
&= -\frac{T}{2}\log(2\pi) - \frac{T}{2}\hat{\sigma}^2 - \frac{T}{2}(\bar{Y} - \mu)^2 \\
&= \text{Constant} - \frac{T}{2}(\bar{Y} - \mu)^2
\end{aligned}
$$

Side Calculation: $\sum_{t=1}^{T}(Y_t - \mu)^2 = T(\bar{Y} - \mu)^2 + T\widehat{\sigma}^2$

$$
\begin{aligned}
T\widehat{\sigma}^2 &= \sum_{t=1}^{T}(Y_t - \bar{Y})^2 = \sum_{t=1}^{T}(Y_t - \mu + \mu - \bar{Y})^2 = \sum_{t=1}^{T}\left[(Y_t - \mu) - (\bar{Y} - \mu)\right]^2 \\
&= \sum_{t=1}^{T}(Y_t - \mu)^2 - \sum_{t=1}^{T}2(Y_t - \mu)(\bar{Y} - \mu) + \sum_{t=1}^{T}(\bar{Y} - \mu)^2 \\
&= \left[\sum_{t=1}^{T}(Y_t - \mu)^2\right] - 2(\bar{Y} - \mu)\left(\sum_{t=1}^{T}Y_t - \sum_{t=1}^{T}\mu\right) + T(\bar{Y} - \mu)^2 \\
&= \left[\sum_{t=1}^{T}(Y_t - \mu)^2\right] - 2(\bar{Y} - \mu)(T\bar{Y} - T\mu) + T(\bar{Y} - \mu)^2 \\
&= \left[\sum_{t=1}^{T}(Y_t - \mu)^2\right] - 2T(\bar{Y} - \mu)^2 + T(\bar{Y} - \mu)^2 \\
&= \left[\sum_{t=1}^{T}(Y_t - \mu)^2\right] - T(\bar{Y} - \mu)^2
\end{aligned}
$$

The Selected Model $\widehat{\mathsf{M}}$

Information Criteria

M_0 sets $\mu = 0$ while M_1 uses the MLE $\bar{Y}$, so we have

$$\mathsf{IC}_0 = 2 \max_{\mu} \{\ell(\mu)\colon \mathsf{M}_0\} = 2 \times \mathsf{Constant} - T\bar{Y}^2$$

$$\mathsf{IC}_1 = 2 \max_{\mu} \{\ell(\mu)\colon \mathsf{M}_1\} - d_T = 2 \times \mathsf{Constant} - d_T$$

Difference of Criteria

$$\mathsf{IC}_1 - \mathsf{IC}_0 = T\bar{Y}^2 - d_T$$

Selected Model

$$\widehat{\mathsf{M}} = \begin{cases} \mathsf{M}_1, & |\sqrt{T}\bar{Y}| \geq \sqrt{d_T} \\ \mathsf{M}_0, & |\sqrt{T}\bar{Y}| < \sqrt{d_T} \end{cases}$$

Case I: $\mu \neq 0$

Apply theory from earlier in lecture. . .

KL-Divergence of M_1

M_1 is the true DGP with minimized KL-divergence equal to zero.

KL-Divergence of M_0

► Truth: $g(y) = (2\pi)^{-1/2} \exp\left\{-(y-\mu)^2/2\right\}$

► M_0: $f(y) = (2\pi)^{-1/2} \exp\left\{-y^2/2\right\}$

► Hence: $\log g(y) - \log f(y) = -\frac{1}{2}(y-\mu)^2 + \frac{1}{2}y^2 = \mu\left(y - \frac{\mu}{2}\right)$

$$\mathsf{KL}(g; M_0) = \int_{\mathbb{R}} \mu(y - \mu/2)(2\pi)^{-1/2} \exp\left\{(y-\mu)^2/2\right\} \, \mathrm{d}y$$

$$= \mu(\mu - \mu/2) = \mu^2/2$$

Verifying Weak Consistency: $\mu \neq 0$

Condition on KL-Divergence

$$\liminf_{T \to \infty} \frac{1}{T} \sum_{t=1}^{T} \{KL(g; \mathsf{M}_0) - KL(g; \mathsf{M}_1)\} = \liminf_{n \to \infty} \frac{1}{T} \sum_{t=1}^{T} \left(\frac{\mu^2}{2} - 0 \right) > 0$$

Condition on Penalty

- Need $c_{T,k} = o_p(T)$, i.e. $c_{T,k}/T \xrightarrow{p} 0$.
- Both AIC and BIC satisfy this
- If $\mu \neq 0$, both AIC and BIC select M_1 wpa 1 as $T \to \infty$.

Case II: $\mu = 0$

What's different?

- ▶ Both M_1 and M_0 are true and minimize KL divergence at zero.

- ▶ Consistency says choose most parsimonious true model: M_0

Verifying Conditions for Consistency

- ▶ $N(0,1)$ model nested inside $N(\mu, 1)$ model

- ▶ Truth is $N(0,1)$ so LR-stat is asymptotically $\chi^2(1) = O_p(1)$.

- ▶ For penalty term, need $\mathbb{P}(c_{T,k} - c_{T,0}) \to \infty$

- ▶ BIC satisfies this but AIC doesn't.

Finite-Sample Selection Probabilities: AIC

AIC Sets $d_T = 2$

$$\widehat{M}_{AIC} = \left\{ \begin{array}{ll} M_1, & |\sqrt{T}\bar{Y}| \geq \sqrt{2} \\ M_0, & |\sqrt{T}\bar{Y}| < \sqrt{2} \end{array} \right.$$

$$
\begin{aligned}
P\left(\widehat{M}_{AIC} = M_1\right) &= P\left(\left|\sqrt{T}\bar{Y}\right| \geq \sqrt{2}\right) \\
&= P\left(\left|\sqrt{T}\mu + Z\right| \geq \sqrt{2}\right) \\
&= P\left(\sqrt{T}\mu + Z \leq -\sqrt{2}\right) + \left[1 - P\left(\sqrt{T}\mu + Z \leq \sqrt{2}\right)\right] \\
&= \Phi\left(-\sqrt{2} - \sqrt{T}\mu\right) + \left[1 - \Phi\left(\sqrt{2} - \sqrt{T}\mu\right)\right]
\end{aligned}
$$

where $Z \sim N(0,1)$ since $\bar{Y} \sim N(\mu, 1/T)$ because $Var(Y_t) = 1$.

Finite-Sample Selection Probabilities: BIC

BIC sets $d_T = \log(T)$

$$\widehat{M}_{BIC} = \left\{ \begin{array}{ll} M_1, & |\sqrt{T}\bar{Y}| \geq \sqrt{\log(T)} \\ M_0, & |\sqrt{T}\bar{Y}| < \sqrt{\log(T)} \end{array} \right.$$

Same steps as for the AIC except with $\sqrt{\log(T)}$ in the place of $\sqrt{2}$:

$$
\begin{aligned}
P\left(\widehat{M}_{BIC} = M_1\right) &= P\left(\left|\sqrt{T}\bar{Y}\right| \geq \sqrt{\log(T)}\right) \\
&= \Phi\left(-\sqrt{\log(T)} - \sqrt{T}\mu\right) + \left[1 - \Phi\left(\sqrt{\log(T)} - \sqrt{T}\mu\right)\right]
\end{aligned}
$$

Interactive Demo: AIC vs BIC

`https://fditraglia.shinyapps.io/CH_Figure_4_1/`

Probability of Over-fitting

- If $\mu = 0$ both models are true but M_0 is more parsimonious.

- Probability of over-fitting (Z denotes standard normal):

$$
\begin{aligned}
P\left(\widehat{M} = M_1\right) &= P\left(|\sqrt{T}\bar{Y}| \geq \sqrt{d_T}\right) = P(|Z| \geq \sqrt{d_T}) \\
&= P(Z^2 \geq d_T) = P(\chi_1^2 \geq d_T)
\end{aligned}
$$

- AIC: $d_T = 2$ and $P(\chi_1^2 \geq 2) \approx 0.157$.

- BIC: $d_T = \log(T)$ and $P(\chi_1^2 \geq \log T) \to 0$ as $T \to \infty$.

AIC has $\approx 16\%$ prob. of over-fitting; BIC does not over-fit in the limit.

Risk of the Post-Selection Estimator

The Post-Selection Estimator

$$\widehat{\mu} = \left\{ \begin{array}{ll} \bar{Y}, & |\sqrt{T}\bar{Y}| \geq \sqrt{d_T} \\ 0, & |\sqrt{T}\bar{Y}| < \sqrt{d_T} \end{array} \right.$$

Recall from above

Recall from above that $\sqrt{T}\bar{Y} = \sqrt{T}\mu + Z$ where $Z \sim N(0,1)$

Risk Function

MSE risk times T since Var. of well-behaved estimator $= O(1/T)$

$$R_T(\mu) = T \cdot \mathbb{E}\left[(\widehat{\mu} - \mu)^2\right] = \mathbb{E}\left[\left(\sqrt{T}\widehat{\mu} - \sqrt{T}\mu\right)^2\right]$$

Simplifying the MSE Risk Function

$\sqrt{T}\bar{Y} = \sqrt{T}\mu + Z$ where $Z \sim N(0,1)$

Let $X = \mathbf{1}\{A\}$ where $A = \left\{ |\sqrt{T}\mu + Z| \geq \sqrt{d_T} \right\}$

$$
\begin{aligned}
R_T(\mu) &= \mathbb{E}\left[\left(\sqrt{T}\hat{\mu} - \sqrt{T}\mu \right)^2 \right] \\
&= \mathbb{E}\left\{ \left[\left(\sqrt{T}\mu + Z \right) X - \sqrt{T}\mu \right]^2 \right\} \\
&= \mathbb{P}(A)\, \mathbb{E}\left\{ \left[\left(\sqrt{T}\mu + Z \right) - \sqrt{T}\mu \right]^2 \middle| X = 1 \right\} + [1 - \mathbb{P}(A)] \left(\sqrt{T}\mu \right)^2 \\
&= \mathbb{P}(A)\, \mathbb{E}\left[Z^2 | X = 1 \right] + [1 - \mathbb{P}(A)]\, T\mu^2
\end{aligned}
$$

So we need to calculate $\mathbb{P}(A)\, \mathbb{E}[Z^2|X=1]$ and $\mathbb{P}(A)$.

Calculating $\mathbb{P}(A)$

Define $a = (-\sqrt{d_T} - \sqrt{T}\mu)$ and $b = (\sqrt{d_T} - \sqrt{T}\mu)$

$$
\begin{aligned}
\mathbb{P}(A) &= \mathbb{P}\left(|\sqrt{T}\mu + Z| \geq \sqrt{d_T}\right) \\
&= \mathbb{P}\left(\sqrt{T}\mu + Z \geq \sqrt{d_T}\right) + \mathbb{P}\left(\sqrt{T}\mu + Z \leq -\sqrt{d_T}\right) \\
&= \mathbb{P}(Z \geq b) + \mathbb{P}(Z \leq a) \\
&= 1 - \Phi(b) + \Phi(a)
\end{aligned}
$$

And hence:

$$
1 - \mathbb{P}(A) = \Phi(b) - \Phi(a)
$$

Calculating $\mathbb{P}(A)\,\mathbb{E}[Z^2|X=1]$ – Step 1

Conditional Density of $Z|X = 1$

$$f(z|x = 1) = \frac{\mathbf{1}(A)\varphi(z)}{\mathbb{P}(A)} \quad \text{where } \varphi \text{ is the } N(0,1) \text{ density}$$

Therefore:

$$
\begin{aligned}
\mathbb{P}(A)\,\mathbb{E}[Z^2|X=1] &= \mathbb{P}(A) \int_{\mathbb{R}} z^2 \left[\frac{\mathbf{1}(A)\varphi(z)}{\mathbb{P}(A)}\right] \, \mathrm{d}z \\
&= \int_{-\infty}^{a} z^2 \varphi(z) \, \mathrm{d}z + \int_{b}^{\infty} z^2 \varphi(z) \, \mathrm{d}z
\end{aligned}
$$

Calculating $\mathbb{P}(A)\,\mathbb{E}[Z^2|X=1]$ – Step 2

Unconditional Expectation: $\mathbb{E}[Z^2]$

$$1 = \mathbb{E}[Z^2] = \int_{-\infty}^{a} z^2 \varphi(z)\,\mathrm{d}z + \int_{a}^{b} z^2 \varphi(z)\,\mathrm{d}z + \int_{b}^{\infty} z^2 \varphi(z)\,\mathrm{d}z$$

Therefore:

$$
\begin{aligned}
\mathbb{P}(A)\,\mathbb{E}[Z^2|X=1] &= \int_{-\infty}^{a} z^2 \varphi(z)\,\mathrm{d}z + \int_{b}^{\infty} z^2 \varphi(z)\,\mathrm{d}z \\
&= 1 - \int_{a}^{b} z^2 \varphi(z)\,\mathrm{d}z
\end{aligned}
$$

Calculating $\mathbb{P}(A)\,\mathbb{E}[Z^2|X=1]$ – Step 3

Integration By Parts

Take $u = -z$ and $dv = -z\exp\{-z^2/2\}$ since

$$\frac{d}{dz}\left(\exp\left\{-z^2/2\right\}\right) = -z\exp\left\{-z^2/2\right\}$$

Thus, $v = \exp\{-z^2/2\}$, $du = -1$ and

$$
\begin{aligned}
\int_a^b z^2\phi(z)\, dz &= (2\pi)^{-1/2}\int_a^b z^2\exp\left\{-z^2/2\right\}\, dz \\
&= (2\pi)^{-1/2}\left[-z\exp\left\{-z^2/2\right\}\Big|_a^b + \int_a^b \exp\left\{-\frac{z^2}{2}\right\}\, dz\right] \\
&= a\phi(a) - b\phi(b) + \Phi(b) - \Phi(a)
\end{aligned}
$$

The Simplifed MSE Risk Function

$$
\begin{aligned}
R_T(\mu) &= 1 - [a\phi(a) - b\phi(b) + \Phi(b) - \Phi(a)] + T\mu^2[\Phi(b) - \Phi(a)] \\
&= 1 + [b\phi(b) - a\phi(a)] + (T\mu^2 - 1)[\Phi(b) - \Phi(a)]
\end{aligned}
$$

where

$$
\begin{aligned}
a &= -\sqrt{d_T} - \sqrt{T}\mu \\
b &= \sqrt{d_T} - \sqrt{T}\mu
\end{aligned}
$$

`https://fditraglia.shinyapps.io/CH_Figure_4_2/`

Punchline: Risk of the Post-Selection Estimator

- AIC: bounded worst-case risk

- BIC: low risk in a neighborhood of $\mu = 0$ in exhange for unbounded worst-case risk as sample size grows

- General phenomenon: consistency and efficiency are mutually exclusive: consistent criteria have unbounded worst-case risk.

- For more details, see Yang (2007, ET)

Lecture #7 – Focused Moment Selection

DiTraglia (2016, JoE)

Focused Moment Selection Criterion (FMSC)
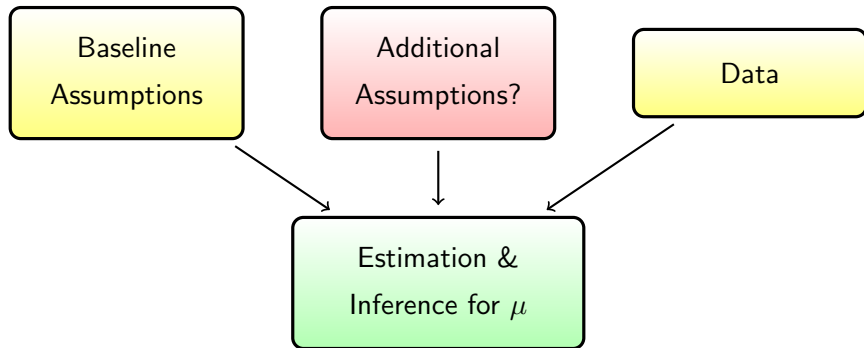
1. Choose False Assumptions on Purpose
2. Focused Choice of Assumptions
3. Local mis-specification
4. Averaging, Inference post-selection

GMM Framework
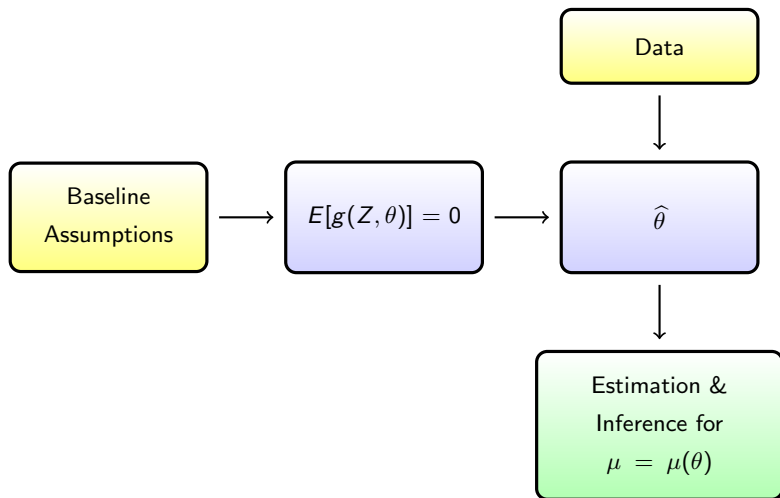
Adding Moment Conditions
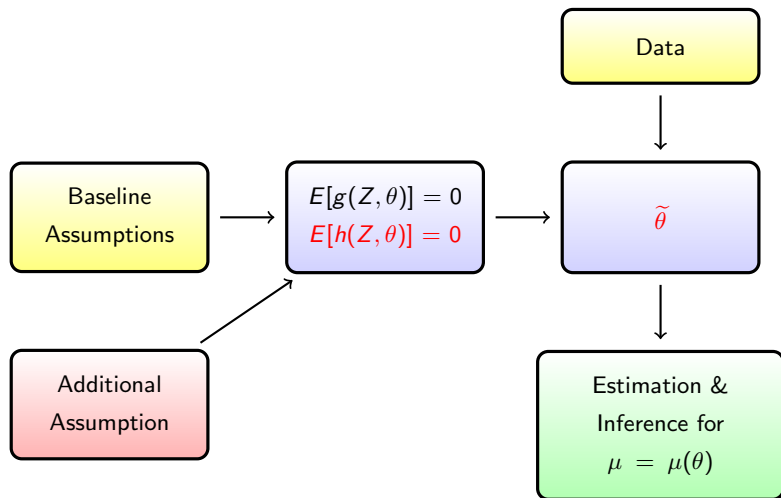
Ordinary versus Two-Stage Least Squares

$$y_i = \beta x_i + \epsilon_i$$
$$x_i = \mathbf{z}_i' \boldsymbol{\pi} + v_i$$

$$E[\mathbf{z}_i \epsilon_i] = 0$$
$$E[x_i \epsilon_i] = ?$$

Choosing Instrumental Variables

$$y_i = \beta x_i + \epsilon_i$$
$$x_i = \Pi_1' \mathbf{z}_i^{(1)} + \Pi_2' \mathbf{z}_i^{(2)} + v_i$$

$$E[\mathbf{z}_i^{(1)} \epsilon_i] = 0$$
$$E[\mathbf{z}_i^{(2)} \epsilon_i] = ?$$

FMSC Asymptotics – Local Mis-Specification
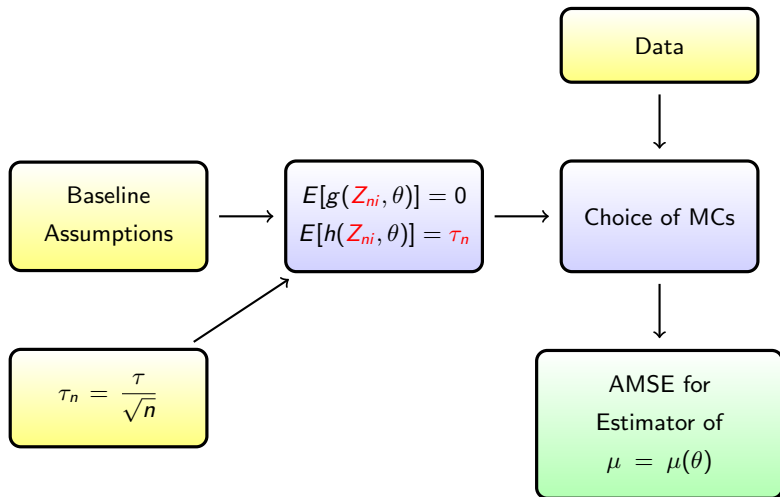
Local Mis-Specification for OLS versus TSLS

$$y_i = \beta x_i + \epsilon_i$$
$$x_i = \mathbf{z}_i'\boldsymbol{\pi} + v_i$$

$$E[\mathbf{z}_i \epsilon_i] = 0$$
$$E[x_i \epsilon_i] = \tau/\sqrt{n}$$

Local Mis-Specification for Choosing IVs

$$y_i = \beta x_i + \epsilon_i$$
$$x_i = \Pi_1' \mathbf{z}_i^{(1)} + \Pi_2' \mathbf{z}_i^{(2)} + v_i$$

$$E[\mathbf{z}_i^{(1)} \epsilon_i] = 0$$
$$E[\mathbf{z}_i^{(1)} \epsilon_i] = \tau/\sqrt{n}$$

Local Mis-Specification

Triangular Array $\{Z_{ni} : 1 \leq i \leq n, n = 1, 2, \ldots\}$ with

(a) $E[g(Z_{ni}, \theta_0)] = 0$

(b) $E[h(Z_{ni}, \theta_0)] = n^{-1/2}\tau$

(c) $\{f(Z_{ni}, \theta_0) : 1 \leq i \leq n, n = 1, 2, \ldots\}$ uniformly integrable

(d) $Z_{ni} \to_d Z_i$, where the Z_i are identically distributed.

Shorthand: Write Z for Z_i

Candidate GMM Estimator

$$\widehat{\theta}_S = \arg\min_{\theta \in \Theta} \; [\Xi_S f_n(\theta)]' \; \widetilde{W_S} \; [\Xi_S f_n(\theta)]$$

$$\Xi_S = \text{Selection Matrix (ones and zeros)}$$

$$\widetilde{W_S} = \text{Weight Matrix (p.s.d.)}$$

$$f_n(\theta) = \begin{bmatrix} g_n(\theta) \\ h_n(\theta) \end{bmatrix} = \begin{bmatrix} n^{-1} \sum_{i=1}^{n} g(Z_{ni}, \theta) \\ n^{-1} \sum_{i=1}^{n} h(Z_{ni}, \theta) \end{bmatrix}$$

Notation: Limit Quantities

$$G = E\left[\nabla_\theta \, g(Z, \theta_0)\right], \quad H = E\left[\nabla_\theta \, h(Z, \theta_0)\right], \quad F = \left[\begin{array}{c} G \\ H \end{array}\right]$$

$$\Omega = Var\left[f(Z, \theta_0)\right] = \left[\begin{array}{cc} \Omega_{gg} & \Omega_{gh} \\ \Omega_{hg} & \Omega_{hh} \end{array}\right]$$

$$\widetilde{W}_S \rightarrow_p W_S \ (\text{p.d.})$$

Local Mis-Specification + Standard Regularity Conditions

Every candidate estimator is consistent for θ_0 and

$$\sqrt{n}(\widehat{\theta}_S - \theta_0) \to_d -K_S \Xi_S \left(\left[\begin{array}{c} M_g \\ M_h \end{array} \right] + \left[\begin{array}{c} 0 \\ \tau \end{array} \right] \right)$$

$$K_S = [F_S' W_S F_S]^{-1} F_S' W_S$$
$$M = (M_g', M_h')'$$
$$M \sim N(0, \Omega)$$

Scalar Target Parameter μ

$$\mu = \mu(\theta) \quad \text{Z-a.s. continuous function}$$

$$\mu_0 = \mu(\theta_0) \quad \text{true value}$$

$$\widehat{\mu} = \mu(\widehat{\theta}_S) \quad \text{estimator}$$

Delta Method

$$\sqrt{n}\left(\widehat{\mu}_S - \mu_0\right) \to_d -\nabla_\theta \mu(\theta_0)' K_S \Xi_S \left(M + \begin{bmatrix} 0 \\ \tau \end{bmatrix} \right)$$

FMSC: Estimate AMSE($\widehat{\mu}_S$) and minimize over S

$$\text{AMSE}(\widehat{\mu}_S) = \nabla_\theta \mu(\theta_0)' K_S \Xi_S \left\{ \begin{bmatrix} 0 & 0 \\ 0 & \tau\tau' \end{bmatrix} + \Omega \right\} \Xi_S' K_S' \nabla_\theta \mu(\theta_0)$$

Estimating the unknowns

No consistent estimator of τ exists! (But everything else is easy)

A Plug-in Estimator of τ
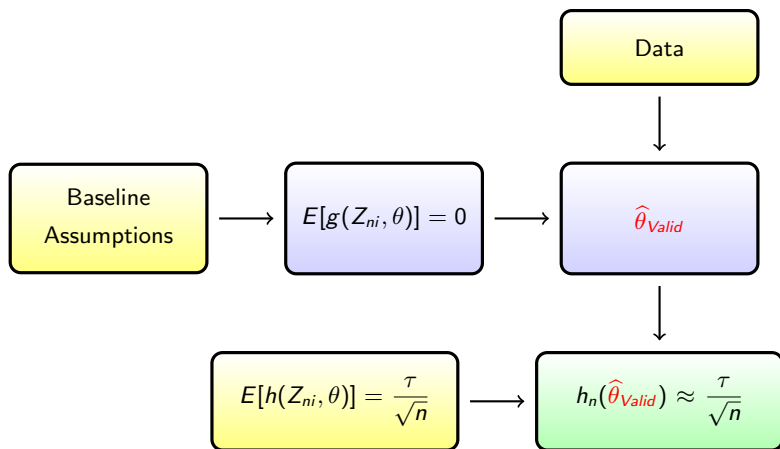
An Asymptotically Unbiased Estimator of $\tau\tau'$

$$\sqrt{n}h_n(\widehat{\theta}_v) = \widehat{\tau} \to_d (\Psi M + \tau) \sim N_q(\tau, \Psi\Omega\Psi')$$

$$\Psi = \left[\begin{array}{cc} -HK_v & \mathbf{I}_q \end{array} \right]$$

$\widehat{\tau}\widehat{\tau}' - \widehat{\Psi}\widehat{\Omega}\widehat{\Psi}$ is an asymptotically unbiased estimator of $\tau\tau'$.

FMSC: Asymptotically Unbiased Estimator of AMSE

$$\mathsf{FMSC}_n(S) = \nabla_\theta \mu(\widehat{\theta})' \widehat{K}_S \Xi_S \left\{ \begin{bmatrix} 0 & 0 \\ 0 & \widehat{B} \end{bmatrix} + \widehat{\Omega} \right\} \Xi_S' \widehat{K}_S' \nabla_\theta \mu(\widehat{\theta})$$

$$\widehat{B} = \widehat{\tau}\widehat{\tau}' - \widehat{\Psi}\widehat{\Omega}\widehat{\Psi}'$$

Choose S to minimize $\mathsf{FMSC}_n(S)$ over the set of candidates $\mathscr{S}$.

A (Very) Special Case of the FMSC

Under homoskedasticity, FMSC selection in the OLS versus TSLS example is *identical* to a Durbin-Hausman-Wu test with $\alpha \approx 0.16$

$$\widehat{\tau} = n^{-1/2}\mathbf{x}'(\mathbf{y} - \mathbf{x}\widetilde{\beta}_{TSLS})$$

OLS gets benefit of the doubt, but not as much as $\alpha = 0.05, \, 0.1$

Limit Distribution of FMSC

$FMSC_n(S) \to_d FMSC_S$, where

$$\text{FMSC}_S \;=\; \nabla_\theta \mu(\theta_0)' K_S \Xi_S \left\{ \begin{bmatrix} 0 & 0 \\ 0 & B \end{bmatrix} + \Omega \right\} \Xi_S' K_S' \nabla_\theta \mu(\theta_0)$$

$$B \;=\; (\Psi M + \tau)(\Psi M + \tau)' - \Psi \Omega \Psi'$$

Conservative criterion: random even in the limit.

Moment Average Estimators

$$\widehat{\mu} = \sum_{S \in \mathscr{S}} \widehat{\omega}_S \widehat{\mu}_S$$

Additional Notation

$\widehat{\mu}$ Moment-average Estimator

$\widehat{\mu}_S$ Estimator of target parameter under moment set S

$\widehat{\omega}_S$ Data-dependent weight function

$\mathscr{S}$ Collection of moment sets under consideration

Examples of Moment-Averaging Weights

Post-Moment Selection Weights

$\widehat{\omega}_S = \mathbf{1}\left\{\mathsf{MSC}_n(S) = \min_{S' \in \mathscr{S}} \mathsf{MSC}_n(S')\right\}$

Exponential Weights

$$\widehat{\omega}_S = \exp\left\{-\frac{\kappa}{2}\mathsf{MSC}(S)\right\} \Big/ \sum_{S' \in \mathscr{S}} \exp\left\{-\frac{\kappa}{2}\mathsf{MSC}(S')\right\}$$

Minimum-AMSE Weights...

Minimum AMSE-Averaging Estimator: OLS vs. TSLS

$$\widetilde{\beta}(\omega) = \omega\widehat{\beta}_{OLS} + (1 - \omega)\widetilde{\beta}_{TSLS}$$

Under homoskedasticity:

$$\omega^* = \left[1 + \frac{\text{ABIAS(OLS)}^2}{\text{AVAR(TSLS)} - \text{AVAR(OLS)}}\right]^{-1}$$

Estimate by:

$$\widehat{\omega}^* = \left[1 + \frac{\max\left\{0, \ \left(\widehat{\tau}^2 - \widehat{\sigma}_\epsilon^2\widehat{\sigma}_x^2\left(\widehat{\sigma}_x^2/\widehat{\gamma}^2 - 1\right)\right) / \widehat{\sigma}_x^4\right\}}{\widehat{\sigma}_\epsilon^2(1/\widehat{\gamma}^2 - 1/\widehat{\sigma}_x^2)}\right]^{-1}$$

Where $\widehat{\gamma}^2 = n^{-1}\mathbf{x}'Z(Z'Z)^{-1}Z'\mathbf{x}$

Limit Distribution of Moment-Average Estimators

$$\widehat{\mu} = \sum_{S \in \mathscr{S}} \widehat{\omega}_S \widehat{\mu}_S$$

(i) $\sum_{S \in \mathscr{S}} \widehat{\omega}_S = 1$ a.s.

(ii) $\widehat{\omega}(S) \to_d \varphi_S(\tau, M)$ a.s.-continuous function of τ, M and consistently-estimable constants only

$$\sqrt{n}\,(\widehat{\mu} - \mu_0) \to_d \Lambda(\tau)$$

$$\Lambda(\tau) = -\nabla_\theta \mu(\theta_0)' \left[\sum_{S \in \mathscr{S}} \varphi_S(\tau, M) K_S \Xi_S \right] \left(M + \begin{bmatrix} 0 \\ \tau \end{bmatrix} \right)$$

Simulating from the Limit Experiment

Suppose τ Known, Consistent Estimators of Everything Else

1. for $j \in \{1, 2, \ldots, J\}$
   (i) $M_j \overset{iid}{\sim} N_{p+q}\left(0, \widehat{\Omega}\right)$
   (ii) $\Lambda_j(\tau) = -\nabla_\theta \mu(\widehat{\theta})' \left[\sum_{S \in \mathscr{S}} \widehat{\varphi}_S (M_j + \tau) \widehat{K}_S \Xi_S \right] (M_j + \tau)$

2. Using $\{\Lambda_j(\tau)\}_{j=1}^{J}$ calculate $\widehat{a}(\tau)$, $\widehat{b}(\tau)$ such that
$$P\left[\widehat{a}(\tau) \leq \Lambda(\tau) \leq \widehat{b}(\tau)\right] = 1 - \alpha$$

3. $P\left[\widehat{\mu} - \widehat{b}(\tau)/\sqrt{n} \leq \mu_0 \leq \widehat{\mu} - \widehat{a}(\tau)/\sqrt{n}\right] \approx 1 - \alpha$

Two-step Procedure for Conservative Intervals

1. Construct $1 - \delta$ confidence region $\mathscr{T}(\widehat{\tau}, \delta)$ for τ

2. For each $\tau^* \in \mathscr{T}(\widehat{\tau}, \delta)$ calculate $1 - \alpha$ confidence interval $\left[\widehat{a}(\tau^*), \widehat{b}(\tau^*)\right]$ for $\Lambda(\tau^*)$ as descibed on previous slide.

3. Take the lower and upper bound over the resulting intervals:
$$\widehat{a}_{min}(\widehat{\tau}) = \min_{\tau^* \in \mathscr{T}} \widehat{a}(\tau^*), \quad \widehat{b}_{max}(\widehat{\tau^*}) = \max_{\tau^* \in \mathscr{T}} \widehat{b}(\tau)$$

4. The interval
$$\text{CI}_{sim} = \left[\widehat{\mu} - \frac{\widehat{b}_{max}(\widehat{\tau})}{\sqrt{n}}, \quad \widehat{\mu} - \frac{\widehat{a}_{min}(\widehat{\tau})}{\sqrt{n}}\right]$$

has asymptotic coverage of at least $1 - (\alpha + \delta)$

OLS versus TSLS Simulation

$$y_i = 0.5x_i + \epsilon_i$$
$$x_i = \pi(z_{1i} + z_{2i} + z_{3i}) + v_i$$

$(\epsilon_i, v_i, z_{1i}, z_{2i}, z_{3i}) \sim$ iid $N(0, \mathcal{S})$

$$\mathcal{S} = \begin{bmatrix} 1 & \rho & 0 & 0 & 0 \\ \rho & 1 - \pi^2 & 0 & 0 & 0 \\ 0 & 0 & 1/3 & 0 & 0 \\ 0 & 0 & 0 & 1/3 & 0 \\ 0 & 0 & 0 & 0 & 1/3 \end{bmatrix}$$

$Var(x) = 1,$ $\qquad \rho = Cor(x, \epsilon),$ $\qquad \pi^2 =$ First-Stage R^2
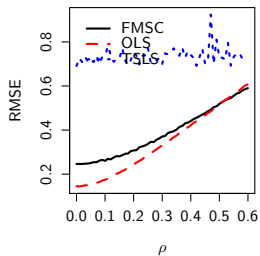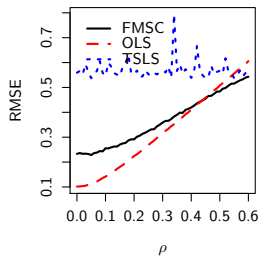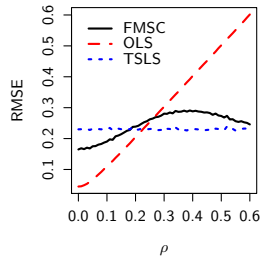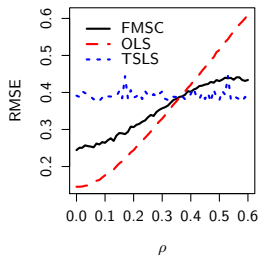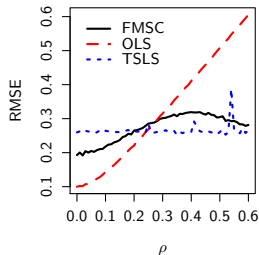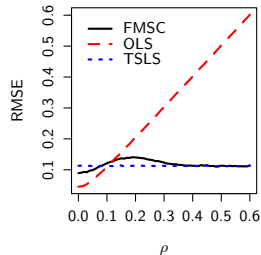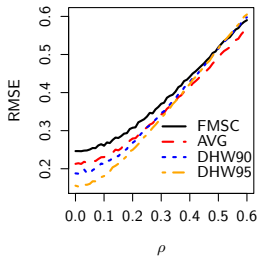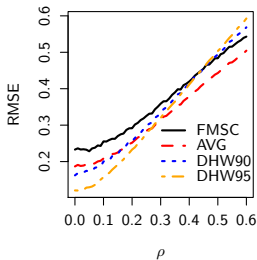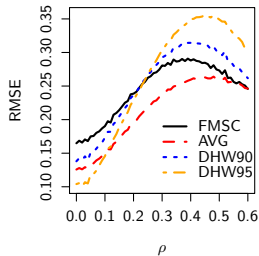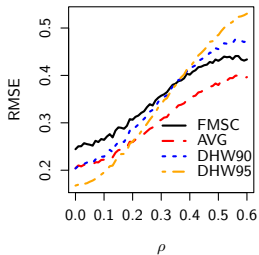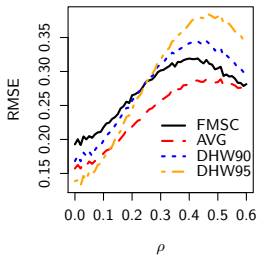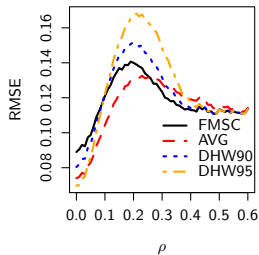
Choosing Instrumental Variables Simulation

$$y_i = 0.5x_i + \epsilon_i$$

$$x_i = (z_{1i} + z_{2i} + z_{3i})/3 + \gamma w_i + v_i$$

$(\epsilon_i, v_i, w_i, z_{i1}, z_{2i}, z_{3i})' \sim$ iid $N(0, \mathcal{V})$

$$\mathcal{V} = \begin{bmatrix} 1 & (0.5 - \gamma\rho) & \rho & 0 & 0 & 0 \\ (0.5 - \gamma\rho) & (8/9 - \gamma^2) & 0 & 0 & 0 & 0 \\ \rho & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1/3 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1/3 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1/3 \end{bmatrix}$$

$\gamma = Cor(x, w), \quad \rho = Cor(w, \epsilon), \quad$ First-Stage $R^2 = 1/9 + \gamma^2$

$$Var(x) = 1, \quad Cor(x, \epsilon) = 0.5$$
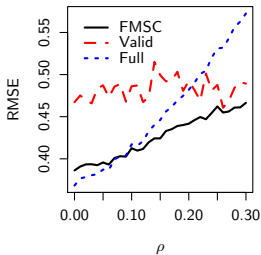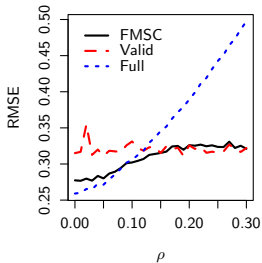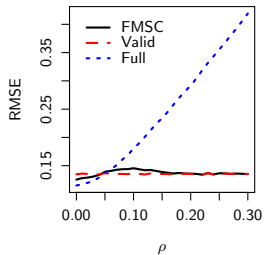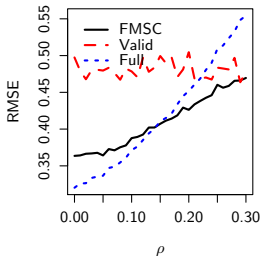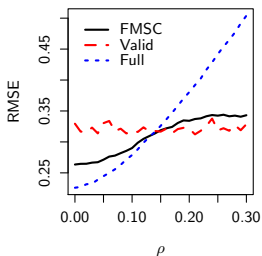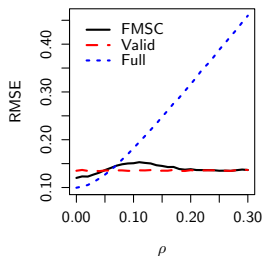
Alternative Moment Selection Procedures

Downward J-test

Use Full instrument set unless J-test rejects.

Andrews (1999) – GMM Moment Selection Criteria

$\text{GMM-MSC}(S) = J_n(S) - \text{Bonus}$

Hall & Peixe (2003) – Canonical Correlations Info. Criterion

$\text{CCIC}(S) = n \log \left[1 - R_n^2(S) \right] + \textit{Penalty}$

Penalty/Bonus Terms
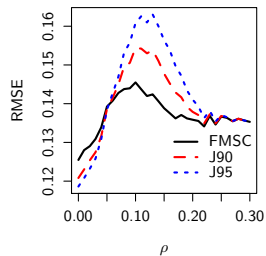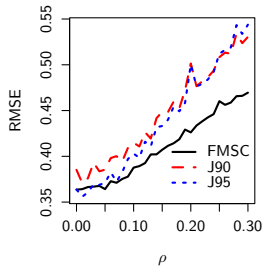
Analogies to AIC, BIC, and Hannan-Quinn
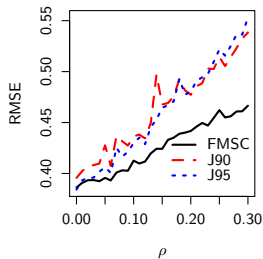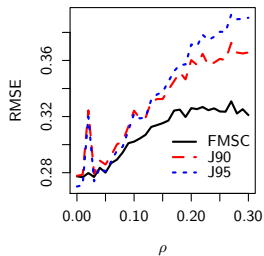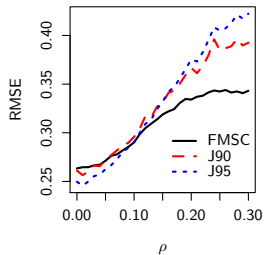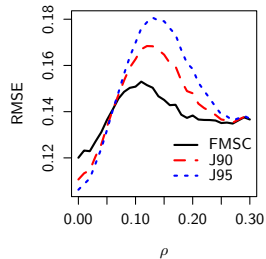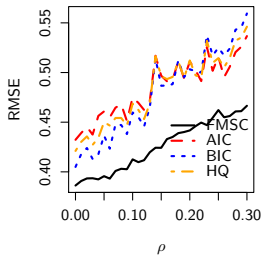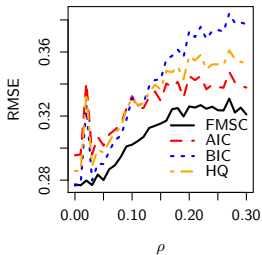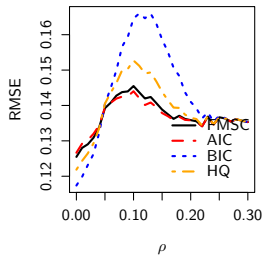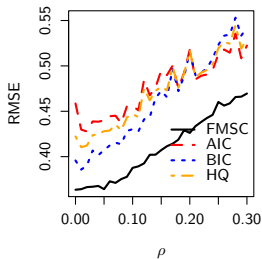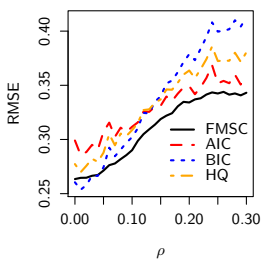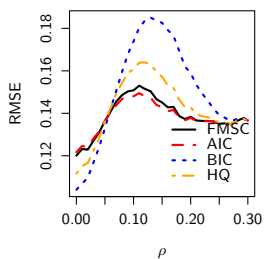
Empirical Example: Geography or Institutions?

Institutions Rule

Acemoglu et al. (2001), Rodrik et al. (2004), Easterly & Levine (2003) – zero or negligible effects of "tropics, germs, and crops" in income per capita, controlling for institutions.

Institutions *Don't* Rule

Sachs (2003) – Large negative direct effect of malaria transmission on income.

Carstensen & Gundlach (2006)

How robust is Sachs's result?

Carstensen & Gundlach (2006)

Both Regressors Endogenous

$$\ln GDPC_i = \beta_1 + \beta_2 \cdot INSTITUTIONS_i + \beta_3 \cdot MALARIA_i + \epsilon_i$$

Robustness

- Various measures of *INSTITUTIONS*, *MALARIA*
- Various instrument sets
- β_3 remains large, negative and significant.

2SLS for All Results That Follow

Expand on Instrument Selection Exercise

FMSC and Corrected Confidence Intervals

1. FMSC – which instruments to estimate effect of malaria?

2. Correct CIs for Instrument Selection – effect of malaria still negative and significant?

Measures of *INSTITUTIONS* and *MALARIA*

▶ *rule* – Average governance indicator (Kaufmann, Kray and Mastruzzi; 2004)

▶ *malfal* – Proportion of population at risk of malaria transmission in 1994 (Sachs, 2001)

Instrument Sets

Baseline Instruments – Assumed Valid

- ▶ ln*mort* – Log settler mortality (per 1000), early 19th century

- ▶ *maleco* – Index of stability of malaria transmission

Further Instrument Blocks

Climate   *frost*, *humid*, *latitude*

Europe   *eurfrac*, *engfrac*

Openness   *coast*, *trade*

	$\mu = \textit{malfal}$			$\mu = \textit{rule}$		
	FMSC	posFMSC	$\widehat{\mu}$	FMSC	posFMSC	$\widehat{\mu}$
(1) Valid	3.0	3.0	−1.0	1.3	1.3	0.9
(2) Climate	3.1	3.1	−0.9	1.0	1.0	1.0
(3) Open	2.3	2.4	−1.1	1.2	1.2	0.8
(4) Eur	1.8	2.2	−1.1	0.5	0.7	0.9
(5) Climate, Eur	0.9	2.0	−1.0	0.3	0.6	0.9
(6) Climate, Open	1.9	2.3	−1.0	0.5	0.8	0.9
(7) Open, Eur	1.6	1.8	−1.2	0.8	0.8	0.8
(8) Full	0.5	1.7	−1.1	0.2	0.6	0.8
> 90% CI FMSC		$(-1.6, -0.6)$			$(0.5, 1.2)$	
> 90% CI posFMSC		$(-1.6, -0.6)$			$(0.6, 1.3)$	

Lecture #8 – High-Dimensional Regression I

The James-Stein Estimator

Recall: Gauss-Markov Theorem

Linear Regression Model

$$\mathbf{y} = X\beta + \boldsymbol{\epsilon}, \quad \mathbb{E}[\boldsymbol{\epsilon}|X] = \mathbf{0}$$

Best Linear Unbiased Estimator

- $\text{Var}(\epsilon|X) = \sigma^2 I \Rightarrow$ then OLS has lowest variance among linear, unbiased estimators of β.

- $\text{Var}(\varepsilon|X) \neq \sigma^2 I \Rightarrow$ then GLS gives a lower variance estimator.

What if we consider biased estimators?

Dominance and Admissibility

Notation

Let R be a risk function, e.g. MSE, and $\widehat{\theta}$ and $\widetilde{\theta}$ be estimators of θ.

Dominance

We say that $\widehat{\theta}$ **dominates** $\widetilde{\theta}$ with respect to R if $R(\widehat{\theta}, \theta) \leq R(\widetilde{\theta}, \theta)$ for all $\theta \in \Theta$ and the inequality is strict for at least one value of θ.

Admissibility

We say that $\widehat{\theta}$ is **admissible** if no other estimator dominates it.

Inadmissiblility

To prove that an estimator $\widetilde{\theta}$ is **inadmissible** it suffices to find an estimator $\widehat{\theta}$ that dominates it.

A Very Simple Example: $X \sim N(\theta, I)$

Goal

Estimate the p-vector of unknown parameters θ using X.

Maximum Likelihood Estimator $\widehat{\theta}$

MLE = sample mean, but only one observation: $\hat{\theta} = X$.

MSE of $\widehat{\theta}$

$$\left(\hat{\theta} - \theta\right)' \left(\hat{\theta} - \theta\right) = (X - \theta)'(X - \theta) = \sum_{i=1}^{p} (X_i - \theta_i)^2 \sim \chi_p^2$$

Since $\mathbb{E}[\chi_p^2] = p$, we have $MSE(\hat{\theta}) = p$.

A Very Simple Example: $X \sim N(\theta, I)$

James-Stein Estimator

$$\hat{\theta}^{JS} = \hat{\theta}\left(1 - \frac{p-2}{\hat{\theta}'\hat{\theta}}\right) = X - \frac{(p-2)X}{X'X}$$

- ▶ Shrinks components of sample mean vector towards zero
- ▶ More elements in $\theta \Rightarrow$ more shrinkage
- ▶ MLE close to zero ($\widehat{\theta'\theta}$ small) gives more shrinkage

MSE of James-Stein Estimator

$$
\begin{aligned}
MSE\left(\hat{\theta}^{JS}\right) &= \mathbb{E}\left[\left(\hat{\theta}^{JS} - \theta\right)'\left(\hat{\theta}^{JS} - \theta\right)\right] \\
&= \mathbb{E}\left[\left\{(X - \theta) - \frac{(p-2)X}{X'X}\right\}'\left\{(X - \theta) - \frac{(p-2)X}{X'X}\right\}\right] \\
&= \mathbb{E}\left[(X - \theta)'(X - \theta)\right] - 2(p-2)\mathbb{E}\left[\frac{X'(X - \theta)}{X'X}\right] \\
&\quad + (p-2)^2\,\mathbb{E}\left[\frac{1}{X'X}\right] \\
&= p - 2(p-2)\mathbb{E}\left[\frac{X'(X - \theta)}{X'X}\right] + (p-2)^2\,\mathbb{E}\left[\frac{1}{X'X}\right]
\end{aligned}
$$

Using fact that $MSE(\hat{\theta}) = p$

Simplifying the Second Term

Writing Numerator as a Sum

$$\mathbb{E}\left[\frac{X'(X-\theta)}{X'X}\right] = \mathbb{E}\left[\frac{\sum_{i=1}^{p} X_i\left(X_i - \theta_i\right)}{X'X}\right] = \sum_{i=1}^{p} \mathbb{E}\left[\frac{X_i(X_i - \theta_i)}{X'X}\right]$$

For $i = 1, \ldots, p$

$$\mathbb{E}\left[\frac{X_i(X_i - \theta_i)}{X'X}\right] = \mathbb{E}\left[\frac{X'X - 2X_i^2}{(X'X)^2}\right]$$

Not obvious: integration by parts, expectation as a p-fold integral, $X \sim N(\theta, I)$

Combining

$$
\begin{aligned}
\mathbb{E}\left[\frac{X'(X-\theta)}{X'X}\right] &= \sum_{i=1}^{p} \mathbb{E}\left[\frac{X'X - 2X_i^2}{(X'X)^2}\right] = p\mathbb{E}\left[\frac{1}{X'X}\right] - 2\mathbb{E}\left[\frac{\sum_{i=1}^{p} X_i^2}{(X'X)^2}\right] \\
&= p\mathbb{E}\left[\frac{1}{X'X}\right] - 2\mathbb{E}\left[\frac{X'X}{(X'X)^2}\right] = (p-2)\,\mathbb{E}\left[\frac{1}{X'X}\right]
\end{aligned}
$$

The MLE is Inadmissible when $p \geq 3$

$$
\begin{aligned}
MSE\left(\hat{\theta}^{JS}\right) &= p - 2(p-2)\left\{(p-2)\,\mathbb{E}\left[\frac{1}{X'X}\right]\right\} + (p-2)^2\,\mathbb{E}\left[\frac{1}{X'X}\right] \\
&= p - (p-2)^2\,\mathbb{E}\left[\frac{1}{X'X}\right]
\end{aligned}
$$

- $\mathbb{E}[1/(X'X)]$ exists and is positive whenever $p \geq 3$

- $(p-2)^2$ is always positive

- Hence, second term in the MSE expression is *negative*

- First term is MSE of the MLE

Therefore James-Stein strictly dominates MLE whenever $p \geq 3$!

James-Stein More Generally

- ▶ Our example was specific, but the result is general:
  - ▶ MLE is inadmissible under quadratic loss in regression model with at least three regressors.
  - ▶ Note, however, that this is MSE for the *full parameter vector*
- ▶ James-Stein estimator is also inadmissible!
  - ▶ Dominated by "positive-part" James-Stein estimator:

$$\widehat{\beta}^{JS} = \widehat{\beta} \left[1 - \frac{(p-2)\widehat{\sigma}^2}{\widehat{\beta}'X'X\widehat{\beta}} \right]_+$$

  - ▶ $\widehat{\beta} = $ OLS, $(x)_+ = \max(x, 0)$, $\widehat{\sigma}^2 = $ usual OLS-based estimator
  - ▶ Stops us us from shrinking *past* zero to get a negative estimate for an element of β with a small OLS estimate.
  - ▶ Positive-part James-Stein isn't admissible either!