Econ 722 – Advanced Econometrics IV

Francis J. DiTraglia

University of Pennsylvania

Lecture #1 – Decision Theory

Statistical Decision Theory

The James-Stein Estimator

Decision Theoretic Preliminaries

Parameter $\theta \in \Theta$

Unknown state of nature, from parameter space Θ

Observed Data

Observe X with distribution F_θ from a sample space $\mathcal{X}$

Estimator $\widehat{\theta}$

An estimator (aka a decision rule) is a function from $\mathcal{X}$ to Θ

Loss Function $L(\theta, \widehat{\theta})$

A function from $\Theta \times \Theta$ to $\mathbb{R}$ that gives the cost we incur if we report $\widehat{\theta}$ when the true state of nature is θ.

Examples of Loss Functions

$L(\theta, \widehat{\theta}) = (\theta - \widehat{\theta})^2$	squared error loss
$L(\theta, \widehat{\theta}) = \lvert\theta - \widehat{\theta}\rvert$	absolute error loss
$L(\theta, \widehat{\theta}) = 0$ if $\theta = \widehat{\theta}$, 1 otherwise	zero-one loss
$L(\theta, \widehat{\theta}) = \int \log\left[\frac{f(x\lvert\theta)}{f(x\lvert\widehat{\theta})}\right] f(x\lvert\theta)\, dx$	Kullback–Leibler loss

(Frequentist) Risk of an Estimator $\widehat{\theta}$

$$\boxed{R(\theta, \widehat{\theta}) = \mathbb{E}_\theta \left[L(\theta, \widehat{\theta}) \right] = \int L\big(\theta, \widehat{\theta}(x)\big)\, dF_\theta(x)}$$

The frequentist decision theorist seeks to evaulate, for each θ, how much he would "expect" to lose if he used $\widehat{\theta}(X)$ repeatedly with varying X in the problem.

(Berger, 1985)

Example: Squared Error Loss

$$R(\theta, \widehat{\theta}) = \mathbb{E}_\theta \left[(\theta - \widehat{\theta})^2 \right] = \text{MSE} = \text{Var}(\widehat{\theta}) + \text{Bias}_\theta^2(\widehat{\theta})$$

Bayes Risk and Maximum Risk

Comparing Risk

$R(\theta, \widehat{\theta})$ is a *function* of θ rather than a single number. We want an estimator with low risk, but how can we compare?

Maximum Risk

$$\boxed{\bar{R}(\widehat{\theta}) = \sup_{\theta \in \Theta} R(\theta, \widehat{\theta})}$$

Bayes Risk

$$\boxed{r(\pi, \widehat{\theta}) = \mathbb{E}_{\pi}\left[R(\theta, \widehat{\theta})\right], \text{ where } \pi \text{ is a prior for } \theta}$$

Bayes and Minimax Rules

Minimize the Maximum or Bayes risk over all estimators $\widetilde{\theta}$

Minimax Rule/Estimator

$\widehat{\theta}$ is minimax if $\boxed{\sup_{\theta \in \Theta} R(\theta, \widehat{\theta}) = \inf_{\widetilde{\theta}} \sup_{\theta \in \Theta} R(\theta, \widetilde{\theta})}$

Bayes Rule/Estimator

$\widehat{\theta}$ is a Bayes rule with respect to prior π if $\boxed{r(\pi, \widehat{\theta}) = \inf_{\widetilde{\theta}} r(\pi, \widetilde{\theta})}$

Recall: Bayes' Theorem and Marginal Likelihood

Let π be a prior for θ. By Bayes' theorem, the posterior $\pi(\theta|\mathbf{x})$ is

$$\pi(\theta|\mathbf{x}) = \frac{f(\mathbf{x}|\theta)\pi(\theta)}{m(\mathbf{x})}$$

where the marginal likelihood $m(\mathbf{x})$ is given by

$$m(\mathbf{x}) = \int f(\mathbf{x}|\theta)\pi(\theta)\, d\theta$$

Posterior Expected Loss

Posterior Expected Loss

$$\rho\left(\pi(\theta|\mathbf{x}), \widehat{\theta}\right) = \int L\left(\theta, \widehat{\theta}\right)\pi(\theta|\mathbf{x})\, d\theta$$

Bayesian Decision Theory

Choose an estimator that minimizes posterior expected loss.

Easier Calculation

Since $m(\mathbf{x})$ does not depend on θ, to minimize $\rho\left(\pi(\theta|\mathbf{x}), \widehat{\theta}\right)$ it suffices to minimize $\int L(\theta, \widehat{\theta}) f(\mathbf{x}|\theta)\pi(\theta)\, d\theta$.

Question

Is there a relationship between Bayes risk, $r(\pi, \widehat{\theta}) \equiv \mathbb{E}_\pi[R(\theta, \widehat{\theta})]$, and posterior expected loss?

Bayes Risk vs. Posterior Expected Loss

Theorem

$$r(\pi, \widehat{\theta}) = \int \rho\big(\pi(\theta|\mathbf{x}), \widehat{\theta}(\mathbf{x})\big)\, m(\mathbf{x})\, d\mathbf{x}$$

Proof

$$\begin{aligned}
r(\pi, \widehat{\theta}) &= \int R(\theta, \widehat{\theta})\pi(\theta)\, d\theta = \int \left[\int L\big(\theta, \widehat{\theta}(\mathbf{x})\big) f(\mathbf{x}|\theta)\, d\mathbf{x}\right] \pi(\theta)\, d\theta \\
&= \int\int L\big(\theta, \widehat{\theta}(\mathbf{x})\big)\, [f(\mathbf{x}|\theta)\pi(\theta)]\; d\mathbf{x} d\theta \\
&= \int\int L\big(\theta, \widehat{\theta}(\mathbf{x})\big)\, [\pi(\theta|\mathbf{x})m(\mathbf{x})]\; d\mathbf{x} d\theta \\
&= \int \left[\int L\big(\theta, \widehat{\theta}(\mathbf{x})\big)\pi(\theta|\mathbf{x})\, d\theta\right] m(\mathbf{x})\, d\mathbf{x} \\
&= \int \rho\big(\pi(\theta|\mathbf{x}), \widehat{\theta}(\mathbf{x})\big)\, m(\mathbf{x})\, d\mathbf{x}
\end{aligned}$$

Finding a Bayes Estimator

Hard Problem

Find the function $\widehat{\theta}(\mathbf{x})$ that minimizes $r(\pi, \widehat{\theta})$.

Easy Problem

Find the number $\widehat{\theta}$ that minimizes $\rho\big(\pi(\theta|\mathbf{x}), \widehat{\theta}\big)$

Punchline

Since $r(\pi, \widehat{\theta}) = \int \rho\big(\pi(\theta|\mathbf{x}), \widehat{\theta}(\mathbf{x})\big)\, m(\mathbf{x})\, d\mathbf{x}$, to minimize $r(\pi, \widehat{\theta})$ we can set $\widehat{\theta}(\mathbf{x})$ to be the value $\widehat{\theta}$ that minimizes $\rho\big(\pi(\theta|\mathbf{x}), \widehat{\theta}\big)$.

Bayes Estimators for Common Loss Functions

Zero-one Loss

For zero-one loss, the Bayes estimator is the posterior mode.

Absolute Error Loss: $L(\theta, \widehat{\theta}) = |\theta - \widehat{\theta}|$

For absolute error loss, the Bayes estimator is the posterior median.

Squared Error Loss: $L(\theta, \widehat{\theta}) = (\theta - \widehat{\theta})^2$

For squared error loss, the Bayes estimator is the posterior mean.

Derivation of Bayes Estimator for Squared Error Loss

By definition,

$$\widehat{\theta} \equiv \arg\min_{a \in \Theta} \int (\theta - a)^2 \pi(\theta|\mathbf{x})\, d\theta$$

Differentiating with respect to a, we have

$$\begin{aligned} 2 \int (\theta - a)\pi(\theta|\mathbf{x})\, d\theta &= 0 \\ \int \theta \pi(\theta|\mathbf{x})\, d\theta &= a \end{aligned}$$

Example: Bayes Estimator for a Normal Mean

Suppose $X \sim N(\mu, 1)$ and π is a $N(a, b^2)$ prior. Then,

$$\begin{aligned}
\pi(\mu|x) &\propto f(x|\mu) \times \pi(\mu) \\
&\propto \exp\left\{-\frac{1}{2}\left[(x-\mu)^2 + \frac{1}{b^2}(\mu - a)^2\right]\right\} \\
&\propto \exp\left\{-\frac{1}{2}\left[\left(1+\frac{1}{b^2}\right)\mu^2 - 2\left(x + \frac{a}{b^2}\right)\mu\right]\right\} \\
&\propto \exp\left\{-\frac{1}{2}\left(\frac{b^2+1}{b^2}\right)\left[\mu - \left(\frac{b^2 x + a}{b^2+1}\right)\right]^2\right\}
\end{aligned}$$

So $\pi(\mu|x)$ is $N(m, \omega^2)$ with $\omega^2 = \frac{b^2}{1+b^2}$ and $m = \omega^2 x + (1-\omega^2)a$.

Hence the Bayes estimator for μ under squared error loss is

$$\boxed{\widehat{\theta}(X) = \frac{b^2 X + a}{1 + b^2}}$$

Minimax Analysis

Wasserman (2004)

The advantage of using maximum risk, despite its problems, is that it does not require one to choose a prior.

Berger (1986)

Perhaps the greatest use of the minimax principle is in situations for which no prior information is available . . . but two notes of caution should be sounded. First, the minimax principle can lead to bad decision rules. . . Second, the minimax approach can be devilishly hard to implement.

Methods for Finding a Minimax Estimator

1. Direct Calculation
2. Guess a "Least Favorable" Prior
3. Search for an "Equalizer Rule"

Method 1 rarely applicable so focus on 2 and 3. . .

The Bayes Rule for a Least Favorable Prior is Minimax

Theorem

Let $\widehat{\theta}$ be a Bayes rule with respect to π and suppose that for all $\theta \in \Theta$ we have $R(\theta, \widehat{\theta}) \leq r(\pi, \widehat{\theta})$. Then $\widehat{\theta}$ is a **minimax estimator**, and π is called a **least favorable prior**.

Proof

Suppose that $\widehat{\theta}$ is not minimax. Then there exists another estimator $\widetilde{\theta}$ with $\sup_{\theta \in \Theta} R(\theta, \widetilde{\theta}) < \sup_{\theta \in \Theta} R(\theta, \widehat{\theta})$. But since

$$r(\pi, \widetilde{\theta}) \equiv \mathbb{E}_\pi \left[R(\theta, \widetilde{\theta}) \right] \leq \mathbb{E}_\pi \left[\sup_{\theta \in \Theta} R(\theta, \widetilde{\theta}) \right] = \sup_{\theta \in \Theta} R(\theta, \widetilde{\theta})$$

but this implies that $\widehat{\theta}$ is *not* Bayes with respect to π since

$$r(\pi, \widetilde{\theta}) \leq \sup_{\theta \in \Theta} R(\theta, \widetilde{\theta}) < \sup_{\theta \in \Theta} R(\theta, \widehat{\theta}) \leq r(\pi, \widehat{\theta})$$

Example of Least Favorable Prior

Bounded Normal Mean

- $X \sim N(\theta, 1)$
- Squared error loss
- $\Theta = [-m, m]$ for $0 < m < 1$

Least Favorable Prior

$\pi(\theta) = 1/2$ for $\theta \in \{-m, m\}$, zero otherwise.

Resulting Bayes Rule is Minimax

$$\widehat{\theta}(X) = m \tanh(mX) = m \left[\frac{\exp\{mX\} - \exp\{-mX\}}{\exp\{mX\} + \exp\{-mX\}} \right]$$

Equalizer Rules

Definition

An estimator $\widehat{\theta}$ is called an **equalizer rule** if its risk function is constant: $R(\theta, \widehat{\theta}) = C$ for some C.

Theorem

If $\widehat{\theta}$ is an equalizer rule and is Bayes with respect to π, then $\widehat{\theta}$ is **minimax** and π is **least favorable**.

Proof

$$r(\pi, \widehat{\theta}) = \int R(\theta, \widehat{\theta})\pi(\theta)\, d\theta = \int C\pi(\theta)\, d\theta = C$$

Hence, $R(\theta, \widehat{\theta}) \leq r(\pi, \widehat{\theta})$ for all θ so we can apply the preceding theorem.

Example: $X_1, \ldots, X_n \sim$ iid Bernoulli(p)

Under a Beta(α, β) prior with $\alpha = \beta = \sqrt{n}/2$,

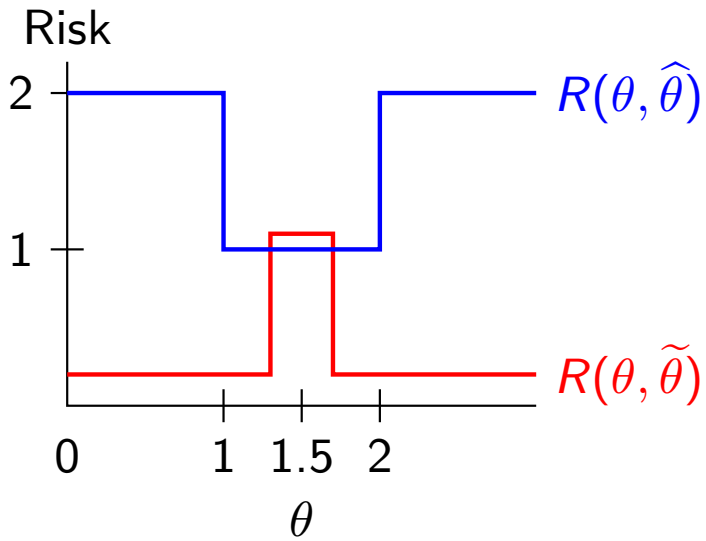
$$\widehat{p}(\mathbf{x}) = \frac{n\bar{X} + \sqrt{n}/2}{n + \sqrt{n}}$$

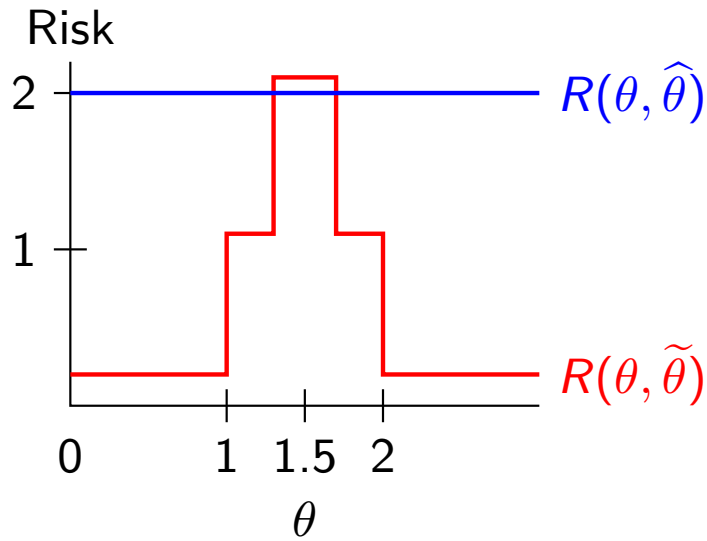
is the Bayesian posterior mean, hence the Bayes rule under squared error loss. The risk function of $\widehat{p}$ is,

$$R(p, \widehat{p}) = \frac{n}{4(n + \sqrt{n})^2}$$

which is constant in p. Hence, $\widehat{p}$ is an equalizer rule, and by the preceding theorem is minimax.

Problems with the Minimax Principle





In the left panel, $\widetilde{\theta}$ is preferred by the minimax principle; in the right panel $\widehat{\theta}$ is preferred. But the only difference between them is that the right panel adds an additional *fixed* loss of 1 for $1 \leq \theta \leq 2$.

Problems with the Minimax Principle

Suppose that $\Theta = \{\theta_1, \theta_2\}$, $\mathcal{A} = \{a_1, a_2\}$ and the loss function is:

	a_1	a_2
θ_1	10	10.01
θ_2	8	-8

- Minimax principle: choose a_1
- Bayes: Choose a_2 unless $\pi(\theta_1) > 0.9994$

Minimax ignores the fact that under θ_1 we can never do better than a loss of 10, and tries to prevent us from incurring a tiny additional loss of 0.01

Dominance and Admissibility

Dominance

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

Admissibility

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

Inadmissiblility

$\widehat{\theta}$ is **inadmissible** if there is an estimator that dominates it.

Example of an Admissible Estimator

Say we want to estimate θ from $X \sim N(\theta, 1)$ under squared error loss. Is the estimator $\widehat{\theta}(X) = 3$ admissible?

If not, then there is a $\widetilde{\theta}$ with $R(\theta, \widetilde{\theta}) \leq R(\theta, \widehat{\theta})$ for all θ. Hence:

$$R(3, \widetilde{\theta}) \leq R(3, \widehat{\theta}) = \left\{ \mathbb{E}\left[\widehat{\theta} - 3 \right] \right\}^2 + \text{Var}(\widehat{\theta}) = 0$$

Since R cannot be negative for squared error loss,

$$0 = R(3, \widetilde{\theta}) = \left\{ \mathbb{E}\left[\widetilde{\theta} - 3 \right] \right\}^2 + \text{Var}(\widetilde{\theta})$$

Therefore $\widehat{\theta} = \widetilde{\theta}$, so $\widehat{\theta}$ is admissible, although very silly!

Bayes Rules are Admissible

Theorem A-1

Suppose that Θ is a discrete set and π gives strictly positive probability to each element of Θ. Then, if $\widehat{\theta}$ is a Bayes rule with respect to π, it is admissible.

Theorem A-2

If a Bayes rule is unique, it is admissible.

Theorem A-3

Suppose that $R(\theta, \widehat{\theta})$ is continuous in θ for all $\widehat{\theta}$ and that π gives strictly positive probability to any open subset of Θ. Then if $\widehat{\theta}$ is a Bayes rule with respect to π, it is admissible.

Admissible Equalizer Rules are Minimax

Theorem

Let $\widehat{\theta}$ be an equalizer rule. Then if $\widehat{\theta}$ is admissible, it is minimax.

Proof

Since $\widehat{\theta}$ is an equalizer rule, $R(\theta, \widehat{\theta}) = C$. Suppose that $\widehat{\theta}$ is not minimax. Then there is a $\widetilde{\theta}$ such that

$$\sup_{\theta \in \Theta} R(\theta, \widetilde{\theta}) < \sup_{\theta \in \Theta} R(\theta, \widehat{\theta}) = C$$

But for any θ, $R(\theta, \widetilde{\theta}) \leq \sup_{\theta \in \Theta} R(\theta, \widetilde{\theta})$. Thus we have shown that $\widetilde{\theta}$ dominates $\widehat{\theta}$, so that $\widehat{\theta}$ cannot be admissible.

Minimax Implies "Nearly" Admissible

Strong Inadmissibility

We say that $\widehat{\theta}$ is **strongly inadmissible** if there exists an estimator $\widetilde{\theta}$ and an $\varepsilon > 0$ such that $R(\theta, \widetilde{\theta}) < R(\theta, \widehat{\theta}) - \varepsilon$ for all θ.

Theorem

If $\widehat{\theta}$ is minimax, then it is **not** strongly inadmissible.

Example: Sample Mean, Unbounded Parameter Space

Theorem

Suppose that $X_1, \ldots, X_n \sim N(\theta, 1)$ with $\Theta = \mathbb{R}$. Under squared error loss, one can show that $\widehat{\theta} = \bar{X}$ is admissible.

Intuition

The proof is complicated, but effectively we view this estimator as a **limit** of a of Bayes estimator with prior $N(a, b^2)$, as $b^2 \to \infty$.

Minimaxity

Since $R(\theta, \bar{X}) = \text{Var}(\bar{X}) = 1/n$, we see that $\bar{X}$ is an equalizer rule. Since it is admissible, it is therefore minimax.

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 and squared error loss?

Multiple Normal Means: $X \sim N(\theta, I)$

Goal

Estimate the p-vector θ using X with $L(\theta, \widehat{\theta}) = ||\widehat{\theta} - \theta||^2$.

Maximum Likelihood Estimator $\widehat{\theta}$

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

Risk of $\widehat{\theta}$

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

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

Multiple Normal Means: $X \sim N(\theta, I)$

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 ($\widehat{\theta}'\widehat{\theta}$ small) gives more shrinkage

MSE of James-Stein Estimator

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

Using fact that $R(\theta, \widehat{\theta}) = p$

Simplifying the Second Term

Writing Numerator as a Sum

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

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 $p \geq 3$

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

- $\mathbb{E}[1/(X'X)]$ 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!

Lecture #2 – Model Selection I

Kullback-Leibler Divergence

Bias of Maximized Sample Log-Likelihood

Review of Asymptotics for Mis-specified MLE

Deriving AIC and TIC

Corrected AIC (AIC_c)

Mallow's C_p

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

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

Properties

- *Not* symmetric: $\mathrm{KL}(g;f) \neq \mathrm{KL}(f;g)$
- By Jensen's Inequality: $\mathrm{KL}(g;f) \geq 0$ (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

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

What if f_θ is correctly specified?

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

Goal: Compare Mis-specified Models

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

where 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

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

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

Information Matrix Equality

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

Bias Relative to Infeasible Plug-in Estimator

Definition of Bias Term B

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

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

Akaike's Information Criterion

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

Contrasting AIC and TIC

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

Corrected AIC (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?

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

1. Would need to know (β_1, σ_1^2) for candidate model.
   - Easy: just use MLE $(\widehat{\boldsymbol{\beta}}_1, \widehat{\sigma}_1^2)$
2. Would need to know $(\boldsymbol{\beta}_0, \sigma_0^2)$ for true model.
   - Very hard! The whole problem is that we don't know these!

Hurvich & Tsai (1989) Assume:

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

Deriving the Corrected AIC

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

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 $\boldsymbol{\beta}$ 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: $\text{RSS}(m) = (\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}_m)'(\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}_m)$

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

Step 1: Algebra

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

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

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

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

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

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

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

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

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

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

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 $\widehat{\beta}$ 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: $\mathbf{X}\widehat{\beta} = \mathbf{P}\mathbf{y}$.

Crucial Fact

span($\mathbf{X}_m$) is a subspace of 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 $\widehat{\beta}$ doesn't work...

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

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

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

Bias of Plug-in Estimator

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

Putting Everything Together: 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...

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

Therefore:

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

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

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

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