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
$L(\theta, \widehat{\theta}) = \int \log\left[\frac{f(x\mid\theta)}{f(x\mid\widehat{\theta})}\right] f(x\mid\theta)\, dx$	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

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

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

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

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,

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

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,

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

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

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

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 $\alpha = \beta = \sqrt{n}/2$,

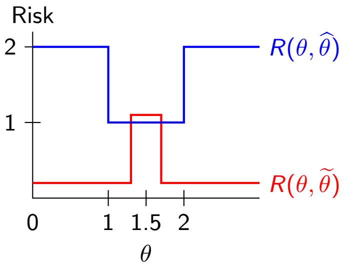
$$\widehat{p}(\mathbf{x}) = \frac{n\bar{X} + \sqrt{n}/2}{n + \sqrt{n}}$$

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 $R(\theta, \bar{X}) = \mathrm{Var}(\bar{X}) = 1/n$, 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

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

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 $p \geq 3$

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

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

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

  - $\widehat{\beta} = \text{OLS}$, $(x)_+ = \max(x, 0)$, $\widehat{\sigma}^2 =$ 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!