METRICS FLOW SO FAR

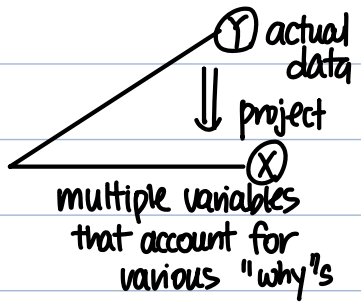
*** Projection Theorem**

- shortest distance between a "guess" and "actual data"
  → way to minimize error / expected loss (risk)

↓

*** OLS** (Bryan → "Least Sq. Fit") — sample

- minimizes Sample sq. error
- use proj thm for LSF



↓

*** Linear Regression** — pop.

- generalization : pop. analogue of LSF
- ※ linear regression = some fnc of joint dist. of X & Y
  ↳ the only feature of pop. that matters for linear reg.

※ (cf) Bryan emphasizes diff btw pop. & sample!

- identification → population : whether mapped well into structural parameters
- statistics → sample : how subsets (sample) approach population

*** Decision Theory**

- introducing Loss, Risk, Decision rule for the first time
  - risk: trade-off btw bias & variance
  - "admissibility" notion becomes important
- Point estimation of θ as a Bayesian
  ⇒ given a prior & likelihood, what is the posterior?
- LB, ML, LFB are simply all decision rules
  ⇒ what type of decision are we trying to make when we observe $X = x$?
- Avg risk : posterior mean = arg min Posterior Exp. Loss
  avg w/o prior

*** Prediction Problem**

- trying to predict values of Y given X
  = finding the conditional mean of $Y|X$
  - if we knew the joint distribution, wouldn't be a problem
  - but all we have is samples → LSF of Y
    ⇒ trying to reduce the volatility of our fit

  ✓ consistency
  ✓ asymptotic normality } ok
  ✓ but $\bar{N}$ → inadmissible

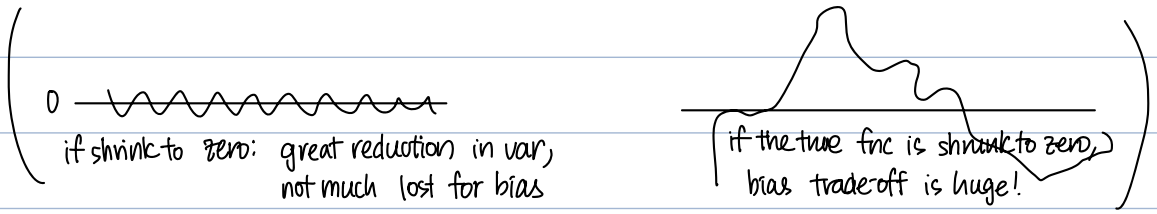
- Suppose $Y|X$ is well described by an 8th-order polynomial
  - ⇒ LSF: 8th-order polynomial

JS est

- Now we multiply unbiased LSF coefficients by 0.7 to reduce volatility
  - ⇒ then BIAS is GENERATED
    (b/c previously unbiased coeff. now biased towards 0)

Oracle est.

- instead of shrinking all of the coefficients uniformly,
  - coeff ≒ 0 → shrink to zero
  - coeff far from 0 → leave them to capture the wriggly parts



Econ 240A - (2)

Soobin

10/21 Monday

* Projection Theorem

- **Vector space $\mathcal{H}$** : an element of $\mathcal{H}$ is a vector
  - (ex) $\mathbb{R}^N$: Euclidean space → usu. (finite) sample space
    L^2 : (functions of) random variables (r.v.) w/ finite variance → usu. population space
  - null vector 0
    - $\mathbb{R}^N$: $\underline{0} = \begin{pmatrix} 0 \\ \vdots \\ 0 \end{pmatrix}$
    - L^2 : 0 = degenerate r.v. which always equals zero

- **Hilbert Space** : Vector Space + Inner Product
  - **Inner product**
    - [Def] $\langle \cdot, \cdot \rangle : \mathcal{H} \times \mathcal{H} \to \mathbb{R}$
      $h_1 \in \mathcal{H}, \ h_2 \in \mathcal{H} \quad \langle h_1, h_2 \rangle$
    - [Properties]
      - (i) Bi-linearity : $\langle a h_1 + b h_2, \ c h_3 + d h_4 \rangle$
        $= ac\langle h_1, h_3 \rangle + ad\langle h_1, h_4 \rangle + bc\langle h_2, h_3 \rangle + bd\langle h_2, h_4 \rangle$
      - (ii) Symmetry : $\langle h_1, h_2 \rangle = \langle h_2 \times h_1 \rangle$
      - (iii) Positivity : $\langle h_1, h_1 \rangle \geq 0$ (inner product with itself : strictly positive if $h_1 \neq \underline{0}$)
    - [Examples]
      - (1) Euclidean space $\mathbb{R}^N$:
        $X = (X_1, \cdots, X_N)', \ Y = (Y_1, \cdots, Y_N)'$ (X_i: years of completed schooling, Y_i = log earnings)
        $\langle X, Y \rangle = \frac{X'Y}{N} = \frac{1}{N}\sum_{i=1}^{N} X_i Y_i$ [check it satisfies i) bi-linearity, ii) symmetry] [cf] dot product
        (N: not by definition, but for the sake of statistical convenience)
      - (2) L^2 space :
        - X : years of completed schooling for a random draw from the target population
        - Y : earnings from random draw

        $(X, Y) \sim f_{X,Y}$ (population)
        $\langle X, Y \rangle = \mathbb{E}[XY]$ ($\mathbb{E}$: expectation)

        $Cov(X,Y) = E[X,Y] - E[X]E[Y]$ common to assume $E[r.v.] = 0$ in L^2
        [cf] covariance inner product

- **Norm** : associated w/ any Inner Product is a norm (measure of distance from the null vector)

[Def] $\|h\| = \langle h, h \rangle^{\frac{1}{2}}$

[Properties]
(i) $\|h\| = 0$ iff $h = 0$ (null vector)
(ii) $\|ah\| = |a| \cdot \|h\|$
(iii) Triangle Inequality (T.I.) : $\|h_1 + h_2\| \leq \|h_1\| + \|h_2\|$

[Pf] Use Lemma 1 to prove T.I.

▶ **Lemma 1 : Cauchy-Schwarz Inequality**

if $h_1, h_2 \in \mathcal{H}$, $|\langle h_1, h_2 \rangle| \leq \|h_1\| \cdot \|h_2\|$

w/ equality iff $h_1 = \alpha h_2$ or $h_2 = 0$

$0 \leq \langle h_1 - \alpha h_2, h_1 - \alpha h_2 \rangle$ (positivity)

$= \langle h_1, h_1 \rangle - \alpha \langle h_1, h_2 \rangle - \alpha \langle h_2, h_1 \rangle + \alpha^2 \langle h_2, h_2 \rangle$ (bi-linearity)

$= \|h_1\|^2 - 2\alpha \langle h_1, h_2 \rangle + \alpha^2 \|h_2\|^2$ (symmetry, def. of norm)

Set $\alpha = \frac{\langle h_1, h_2 \rangle}{\|h_2\|^2}$

$0 \leq \|h_1\|^2 - \frac{\langle h_1, h_2 \rangle^2}{\|h_2\|^2} \iff \langle h_1, h_2 \rangle^2 \leq \|h_1\|^2 \cdot \|h_2\|^2$

[Ex] use CS inequality to show that correlation coefficient : $-1 \leq \rho \leq 1$.

(Two cases)
- $\rho = \frac{Cov(X,Y)}{V(X)^{\frac{1}{2}} \cdot V(Y)^{\frac{1}{2}}}$ ← use L^2, inner product
- sample correlation coeff. ← use $\mathbb{R}^N$, inner product

▶ [Lemma 2] $\|h_1 + h_2\| \leq \|h_1\| + \|h_2\|$

$\|h_1 + h_2\|^2 \leq \langle h_1 + h_2, h_1 + h_2 \rangle$

$= \|h_1\|^2 + 2\langle h_1, h_2 \rangle + \|h_2\|^2$

$\leq \|h_1\|^2 + 2|\langle h_1, h_2 \rangle| + \|h_2\|^2$

$\overset{CS}{\leq} \|h_1\|^2 + 2\|h_1\| \cdot \|h_2\| + \|h_2\|^2$

$= (\|h_1\| + \|h_2\|)^2$ ■

- **Orthogonal** : X, Y are orthogonal ($X \perp Y$) if their inner product is zero.

[cf] $X \perp\!\!\!\perp Y$: X is independent of Y

▶ [Thm 1 : Pythagorean Thm]

If $h_1 \perp h_2$, then $\|h_1\|^2 + \|h_2\|^2 = \|h_1 + h_2\|^2$

[Ex] Pf + connect it to geometric property : $\mathbb{R}^2$

• Other things to think about

- $\|h_1 - h_2\|$ is a measure of distance btw h_1 and h_2

  - $\mathbb{R}^2$: $h_1 = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}$, $h_2 = \begin{pmatrix} y_1 \\ y_2 \end{pmatrix}$

    $\|h_1 - h_2\| = \sqrt{(x_1-y_1)^2 + (x_2-y_2)^2}$

  - L^2 : $h_1 = X$, $h_2 = Y$

    $\|h_1 - h_2\| = E[(X-Y)^2]^{\frac{1}{2}}$ = Root Mean Squared Error (RMSE)

• Projection Thm

Let $\mathcal{L}$ be some linear subspace of $\mathcal{H}$.

Projection of $Y \in \mathcal{H}$ onto the subspace $\mathcal{L}$.

• Projection operator : $\Pi(\cdot|\mathcal{L}) : \mathcal{H} \to \mathcal{L}$

where $\Pi(Y|\mathcal{L})$ is the element of $\hat{Y} \in \mathcal{L}$ that achieves $\min_{\hat{Y} \in \mathcal{L}} \|Y - \hat{Y}\|$

• (Ex: $\mathbb{R}^N$) $\underset{N\times 1}{Y}$, $\underset{N\times 1}{X}$, $\mathcal{L}$ = linear span of $\underset{N\times 1}{1} = \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix} = 1_N$, X

↳ (i.e. vectors of the form $\alpha 1 + \beta X$)

$\min_{(\alpha,\beta)\in\mathbb{R}^2} \|Y - \alpha \cdot 1 - \beta X\|^2$

$= \min_{(\alpha,\beta)\in\mathbb{R}^2} \frac{1}{N} \sum_{i=1}^{N} (Y_i - \alpha - \beta X_i)^2$

→ corresponds to finding the OLS fit of Y onto a constant and X.

• (Ex: L^2) $(X, Y) \sim F_{X,Y}$, $\mathcal{L}$ all linear functions of X ; $\alpha + \beta X$

$\min_{(\alpha,\beta)\in\mathbb{R}^2} \|Y - \alpha - \beta X\|^2 = \min_{(\alpha,\beta)\in\mathbb{R}^2} \underbrace{E[(Y - \alpha - \beta X)^2]}_{\text{Mean Squared Error}}$ } population linear regression function

▶ [Thm 2 : Projection Thm]

Let $\mathcal{H}$ be a vector space with an inner product and associated norm and $\mathcal{L}$ a subspace of $\mathcal{H}$

Then for Y an arbitrary element of $\mathcal{H}$,

if there exists a vector $\hat{Y} \in \mathcal{L}$ s.t.

$$\|Y - \hat{Y}\| \leq \|Y - \tilde{Y}\| \quad \text{for } \forall \tilde{Y} \in \mathcal{L}$$

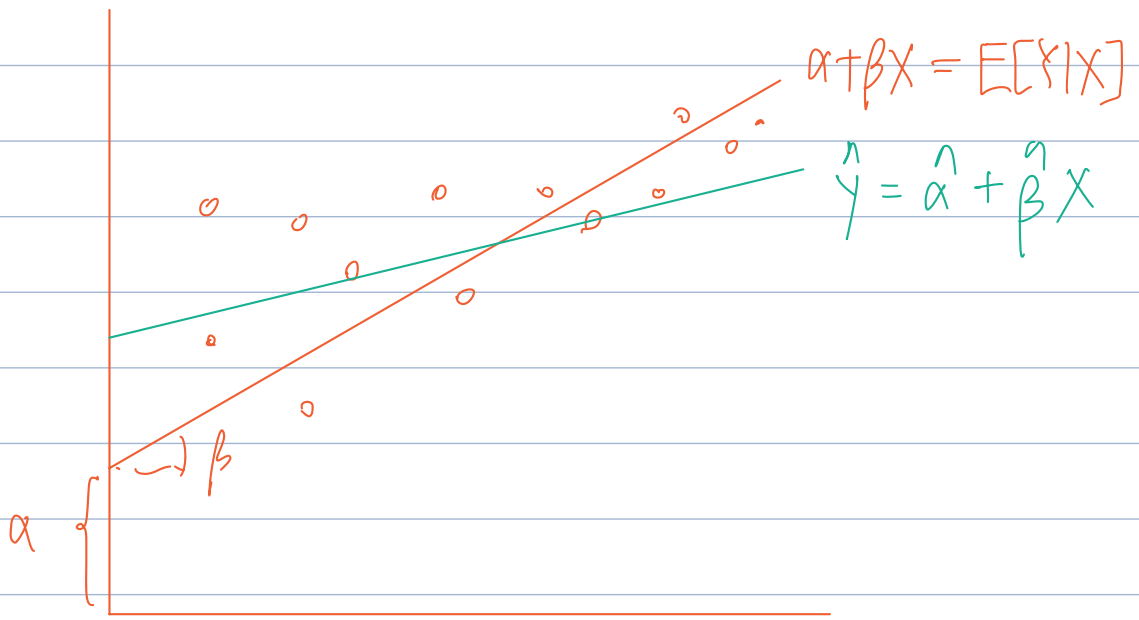
then (i) $\hat{Y} = \overset{\text{projection}}{\Pi}(Y|\mathcal{L})$ is unique

(ii) nec & suff condition for $\hat{Y}$ to be the solution:

$$\langle Y - \hat{Y}, \tilde{Y} \rangle = 0 \quad \text{for } \tilde{Y} \in \mathcal{L}$$

[or $Y - \Pi(Y|\mathcal{L}) \perp \tilde{Y}$ for $\forall \tilde{Y} \in \mathcal{L}$]

↑ a lot of the power of regression comes from the orthogonality



10/23 Lecture

* Projection Thm (cont'd)

·Ẋ· Recap

- $\mathcal{H}$: vector space w/ inner product + norm
- $\mathcal{L}$: subspace, $Y \in \mathcal{H}$, $\hat{Y} \in \mathcal{L}$ *
- If $\exists\ \hat{Y} \in \mathcal{L}$, $\|Y - \hat{Y}\| \leq \|Y - \tilde{Y}\|$ $\forall\ \tilde{Y} \in \mathcal{L}$
  - (i) $\hat{Y} = \Pi(Y|\mathcal{L})$ is unique
  - (ii) nec + suff : $\langle Y - \hat{Y}, \tilde{Y} \rangle = 0$ $\forall\ \tilde{Y} \in \mathcal{L}$

$Y - \Pi(Y|\mathcal{L}) \perp Y$: orthogonal

(Y : target, $\Pi(Y|\mathcal{L})$: approximated, $Y - \Pi(Y|\mathcal{L})$: approx. error)

[Pf] in the notes (* forthcoming)

• Properties of Projections

(i) Projections are Linear operators

[Set-up] (all vectors can be expressed in this form)

$X = \underbrace{\Pi(X|\mathcal{L})}_{\text{projection onto } \mathcal{L}} + \underbrace{u_X}_{\text{projection error}}$, $u_X \perp \mathcal{L}$, $u_X \perp \tilde{Y}$ $\forall\ \tilde{Y} \in \mathcal{L}$ (X : vector)

$Y = \Pi(Y|\mathcal{L}) + u_Y$, $u_Y \perp \mathcal{L}$

① $aX + bY = a\Pi(X|\mathcal{L}) + b\Pi(Y|\mathcal{L}) + au_X + bu_Y$ (a, b) scalar

Linear combination of two vectors:

② $W \in \mathcal{L}$

· by bilinearity $\Rightarrow \langle au_X + bu_Y, W \rangle = a\underbrace{\langle u_X, W \rangle}_{=0} + b\underbrace{\langle u_Y, W \rangle}_{=0}$ $W \in \mathcal{L}$

$= 0$

$\Pi(aX + bY) = a\Pi(X|\mathcal{L}) + b\Pi(Y|\mathcal{L})$ $\Rightarrow$ Projections are linear operators

[Following Property] $E[aX + bY \mid W] = a\mathbb{E}[X|W] + b\mathbb{E}[Y|W]$

$= a\int x \cdot f_{X|W}(x|W)\,dx + b\int y \cdot f_{Y|W}(y|W)\,dy$

·Ẋ· (notation)

- $Y \sim F_Y$: generic random draw $\rightarrow$ Y_i : i-th random draw
- $\mathbb{E}[Y] = \mathbb{E}[Y_i] = \mathbb{E}[Y_1] = \mathbb{E}[Y_{\forall i}]$: under sample random sampling
- $Y \in \mathcal{Y}$: support of r.v. Y

y : specific element of $\mathcal{Y}$ (sample space)

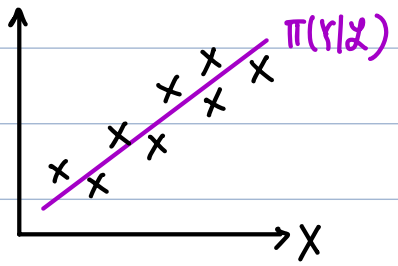
$\mathbb{E}[Y|X] = g(X)$ (r.v. : : function of r.v.) $\quad (X,Y) \sim F_{X,Y}$

$\mathbb{E}[Y|X=x] = g(x)$ for $x \in \mathcal{X}$

$\mathbb{E}[Y|x]$ (Goldberger notation)

(ii) Idempotency

Q. "The Projection of a projection is itself." (T/F question possible) $\Rightarrow$ (T)



[Pf] WTS $\pi(\pi(Y|\mathcal{L})|\mathcal{L}) = \pi(Y|\mathcal{L})$

$0 = \langle \pi(\pi(Y|\mathcal{L})|\mathcal{L}) - \pi(Y|\mathcal{L}), \tilde{Y} \rangle \quad \tilde{Y} \in \mathcal{L}$ by NSC of PT

($\pi(\pi(Y|\mathcal{L})|\mathcal{L})$: Y-Y; $\pi(Y|\mathcal{L})$: target)

$= \langle Y - \pi(Y|\mathcal{L}), \tilde{Y} \rangle - \langle Y - \pi(\pi(Y|\mathcal{L})|\mathcal{L}), \tilde{Y} \rangle$ by bi-linearity, uniqueness of PT.

($\langle Y - \pi(Y|\mathcal{L}), \tilde{Y} \rangle = 0$ P.T.; $\langle Y - \pi(\pi(Y|\mathcal{L})|\mathcal{L}), \tilde{Y} \rangle$: NSC of PT)

[Norm Reducing] Let $\mathbf{1}$ denote the constant vector

(a) $\mathcal{H} = \mathbb{R}^N$ Euclidean space ; $\mathbf{1} = \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix} = \mathbf{1}_N$

(b) $\mathcal{H} = L^2$; $\mathbf{1} = 1$

(a) $\pi(Y|\mathbf{1}) = \frac{1}{N} \cdot \sum_{i=1}^{N} Y_i = \bar{Y}$ $\quad \langle \underset{N\times 1}{X}, \underset{N\times 1}{Y} \rangle = \frac{1}{N} \Sigma X_i Y_i = \frac{1}{N} X'Y$

(b) $\pi(Y|1) = E[Y]$ $\quad \langle X, Y \rangle = E[XY]$

(a) $\| Y - \pi(Y|1) \|^2 = N \Sigma (Y_i - \bar{Y})^2$ (projection error)

(b) $\| Y - \pi(Y|1) \|^2 = \mathbb{V}(Y)$ $= \text{var}(Y)$

(iii) [Lemma 3: Analysis of Variance]

▶ If the linear subspace $\mathcal{L}$ contains a constant vector,

$$\| Y - \pi(Y|1) \|^2 = \| Y - \pi(Y|\mathcal{L}) \|^2 + \| \pi(Y|\mathcal{L}) \|^2 - \| \pi(Y|1) \|^2$$

($\| Y - \pi(Y|1) \|^2$: projection error from projecting Y onto a constant vector; $\mathcal{L}$ contains $\mathbf{1}$)

(cf)

TSS	=	RSS	+	ESS
$\mathbb{V}(Y)$	=	$\mathbb{E}[\mathbb{V}(Y\|X)]$	+	$\mathbb{V}(\mathbb{E}[Y\|X])$

- [Pf] $\|Y - \pi(Y|\mathcal{L})\|^2 = \|Y - \pi(Y|1) - [\pi(Y|\mathcal{L}) - \pi(Y|1)]\|^2$

  (by def. of norm $\|Y\| = \langle Y, Y\rangle^{\frac{1}{2}}$ & by bi-linearity) $= \|Y - \pi(Y|1)\|^2 - 2\langle Y - \pi(Y|1),\ \pi(Y,\mathcal{L}) - \pi(Y|1)\rangle_{(*)} + \|\pi(Y|\mathcal{L}) - \pi(Y|1)\|^2$

  (Show: $(*) = \|\pi(Y|\mathcal{L}) - \pi(Y|1)\|^2$ HW)

  $= \|Y - \pi(Y|1)\|^2 - \|\pi(Y|\mathcal{L}) - \pi(Y|1)\|^2$

- [Implication] Projections are norm-reducing:

  $\|Y - \pi(Y|\mathcal{L})\| \leq \|Y - \pi(Y|1)\|$

⇒ [Recap]
- (i) Projections are linear operators
- (ii) Idempotency
- (iii) ANOVA, Projections are norm-reducing

* OLS

- **Set-up**
  - $Y_{N\times 1}$: log earnings for a random sample of N adult males
  - $X_{N\times K}$: covariates — e.g. 1, Yrs Sch, AFQT, demographic controls ⋯
  - Euclidean space

  Assume : columns of $X_{N\times K} = \begin{pmatrix} 1 & \text{Yrs Sch}_1 & \text{AFQT}_1 \\ \vdots & \vdots & \vdots \\ 1 & \text{Yrs Sch}_N & \text{AFQT}_N \end{pmatrix}$ → assume linearly independent ∴ rank(X) = K

  $\mathcal{L} = C(X) := \{ \underbrace{X\beta}_{(N\times K)(K\times 1)} : \beta \text{ is a } K\times 1 \text{ vector of real numbers} \}$

  ($C(X)$: column space of X)

- **Find: Projection of $Y_{N\times 1}$ onto the column space of $X_{N\times K}$ (i.e. onto $\mathcal{L} = C(X)$)**

  Use the Projection Thm:

  (NSC) $\langle Y - \underbrace{\pi(Y|\mathcal{L})}_{= X\hat{\beta}},\ \underbrace{X\beta}_{\text{an element of } \mathcal{L}=C(X)} \rangle = 0$

  (#) $\langle \underbrace{Y - X\hat{\beta}}_{\text{projection error}},\ X\beta \rangle = 0$

  (#) $\frac{1}{N}\sum^{N}(Y_i - X_i'\hat{\beta})X_i'\beta = 0$

$$\underset{N\times K}{\mathbb{X}} = \begin{pmatrix} 1 & X_{11} & \cdots & X_{1K-1} \\ 1 & X_{21} & \cdots & X_{2K-1} \\ \vdots & \vdots & \ddots & \\ 1 & X_{N1} & \cdots & X_{NK-1} \end{pmatrix} \qquad \underset{K\times 1}{X_i} = \begin{pmatrix} 1 \\ X_{i1} \\ \vdots \\ X_{iK-1} \end{pmatrix} \Rightarrow \underset{N\times K}{\mathbb{X}} = \begin{pmatrix} X_1' \\ X_2' \\ \vdots \\ X_N' \end{pmatrix}$$

$$\underset{\underbrace{(N\times K)(K\times 1)}_{N\times 1}}{\mathbb{X}\beta} = \begin{pmatrix} X_1' \\ X_2' \\ \vdots \\ X_N' \end{pmatrix}\beta = \begin{pmatrix} X_1'\beta \\ X_2'\beta \\ \vdots \\ X_N'\beta \end{pmatrix}$$

$X_i'\beta = (1 \quad X_{i1}, \cdots, X_{iK-1}) \begin{pmatrix} \beta_0 \\ \vdots \\ \beta_{K-1} \end{pmatrix} \rightarrow (X_{i1}, \cdots, X_{iK}) \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_K \end{pmatrix}$ where $X_1 = 1$ (constant)

$$= \underbrace{\sum_{k=1}^{K} X_{ik}\beta_k}$$

$$\Rightarrow (\#)\ \frac{1}{N}\sum_{i=1}^{N}\sum_{k=1}^{K}\beta_k X_{ik}\,(Y_i - X_i'\hat{\beta}) = 0 \quad \text{for } \forall \beta \in \mathbb{R}^K \quad (\text{NSC of PT})$$

- Next, set $\begin{cases} \beta_k = 1 \\ \beta_j = 0 \quad \text{for } j \neq k \end{cases}$

$$\Rightarrow \frac{1}{N}\sum_{i=1}^{N} X_{ik}\,(Y_i - X_i'\hat{\beta}) = 0 \quad \text{for } k = 1, \cdots, K$$

- K equations & K unknowns: $\hat{\beta} = \begin{pmatrix} \hat{\beta}_1 \\ \vdots \\ \hat{\beta}_K \end{pmatrix}$

- Stacking the K equations: $\frac{1}{N}\sum_{i=1}^{N} \underset{(K\times 1)}{X_i}(Y_i - X_i'\hat{\beta}) = \begin{pmatrix} 0 \\ \vdots \\ 0 \end{pmatrix}$ [FOC]

$$= \frac{1}{N}\sum_{i=1}^{N} X_i Y_i - \left[\frac{1}{N}\sum_{i=1}^{N} X_i X_i'\right]\hat{\beta} = 0$$

$$\Rightarrow \underbrace{\left[\frac{1}{N}\sum_{i=1}^{N} X_i X_i'\right]}_{K\times K}\underbrace{\hat{\beta}}_{K\times 1} = \underbrace{\left[\frac{1}{N}\sum_{i=1}^{N} X_i Y_i\right]}_{K\times 1}$$

$$\Rightarrow \hat{\beta} = \left[\frac{1}{N}\sum_{i=1}^{N} X_i X_i'\right]^{-1} \cdot \left[\frac{1}{N}\sum_{i=1}^{N} X_i Y_i\right]$$

$$= (\mathbb{X}'\mathbb{X})^{-1}\mathbb{X}'\mathbb{Y} \quad \text{where } \underset{N\times K}{\mathbb{X}}\ \underset{N\times 1}{\mathbb{Y}}$$

$\begin{cases} X\hat{\beta} \text{ unique (in this class)} \\ \hat{\beta} \text{ unique?} \end{cases}$

- Hence, the projection of $\underset{N\times 1}{\mathbb{Y}}$ onto the column space of $\underset{N\times K}{\mathbb{X}}$ is:

$$\Pi(\mathbb{Y}|\mathbb{X}) = \mathbb{X}\hat{\beta} = \mathbb{X}\underbrace{(\mathbb{X}'\mathbb{X})^{-1}\mathbb{X}'\mathbb{Y}}_{\hat{\beta}}$$

: the Ordinary Least Squares fit of Y_i onto X_i for $i = 1, \cdots, N$

- More on OLS coefficient $\hat{\beta}$

$$\hat{\beta} = \arg\min_{\beta \in \mathbb{R}^k} \frac{1}{N} \sum (Y_i - X_i'\beta)^2$$

$$= \arg\min_{\beta \in \mathbb{R}^k} \| \underset{(N\times 1)}{Y} - \underset{(N\times k)(k\times 1)}{X\beta} \|^2$$

since $\|Y - \hat{Y}\| \leq \|Y - \tilde{Y}\| \quad \forall \tilde{Y} \in C(X)$

$\|Y - X\hat{\beta}\| \leq \|Y - X\beta\| \quad \forall \beta \in \mathbb{R}^k$

10/28 Lecture

* Linear Regression

- **Set-up**

  $(X, Y)_{(K\times 1)\ (1\times 1)} \sim F_{X,Y}$

  $d = \{X'\beta - \beta \text{ is a } (K\times 1) \text{ vector of reals}\}$

  $\|X\|_d = \left[\sum_{k=1}^{K} X_k^2\right]^{1/2}$: Euclidean norm

- **Assumption on $F_{X,Y}$ ("regularity conditions")** (→ conditions on moments)

  (A.1)
  - (i) $\mathbb{E}[Y^2] < \infty$ → vector in L^2 space
  - (ii) $\mathbb{E}[\|X\|_d^i] < \infty$
  - (iii) $\mathbb{E}[(\alpha'X)^2] > 0$ (α is a non-zero vector of constants) ⇒ ensures uniqueness of coefficient vector indexing the projection (∵ ensures $\mathbb{E}[XX']^{-1}$ exists)

    ↳ implies: $\mathbb{E}[(\alpha'X)^2] = \alpha' \underbrace{\mathbb{E}[XX']}_{\text{positive def.}} \alpha > 0$

    $\Rightarrow \mathbb{E}[XX']^{-1}_{(K\times K)}$ is well-defined (*)

- **Goal**: Find Projection of Y onto $\mathcal{L}$

  $\min_{\beta\in\mathbb{R}^K} \|Y - X'\beta\|^2 = \min_{\beta\in\mathbb{R}^K} \mathbb{E}[(Y - X'\beta)^2]$

  (= cov. inner product; mean squared error (mse) (heteroskedasticity))

  in L^2: $\langle X, Y\rangle = \mathbb{E}[XY]$, $\|Y\| = \mathbb{E}[Y^2]^{\frac{1}{2}}$

  Use the NSC of Projection Thm:

  $\langle Y - X'\beta_0, X'\beta\rangle = 0 \quad \forall \beta \in \mathbb{R}^K$

  (β_0: coefficient matrix indexing $\hat{Y}$)

  where $\langle Y - \hat{Y}, \tilde{Y}\rangle = 0$ for $\tilde{Y}\in\mathcal{L}$ ($\hat{Y} = X'\beta_0$, $\tilde{Y} = X'\beta$)

  $\mathbb{E}[(\underbrace{Y - X'\beta_0}_{\text{r.v.}})\cdot \underbrace{X'\beta}_{\text{r.v.}}] = 0 \iff \mathbb{E}[(Y - X'\beta_0)X] = 0$

  ($= U$; $\Rightarrow$; $(K\times 1)$; $(K\times 1)$)

  (∵ zero covariance condition)

  $\mathbb{E}[XY] - \mathbb{E}[XX']\beta_0 = 0$ (by (*) from A1-iii)

  $\mathbb{E}[XY] = \mathbb{E}[XX']\beta_0$

  $\Rightarrow \beta_0 = \mathbb{E}[XX']^{-1}\mathbb{E}[XY]$

  Linear Regression: $\pi(Y|\mathcal{L}) = \underbrace{\mathbb{E}^*[Y|X]}_{\text{linear reg. of } Y \text{ on } X} = X'\beta_0$

  (not Expectation operator)

  $\beta_0 = \mathbb{E}[XX']^{-1}\cdot\mathbb{E}[XY]$

∴ (Note) $X \sim F_X$:
- X is a random draw from F_X
- x is a specific configuration
- $\mathcal{X}$ is the support of r.v. X

* Decision Theory <Laplace>

- **Notation / Introduction**

$\{P_\theta : \theta \in \Theta\}$
- P_θ: statistical model
- θ: parameter
- Θ: parameter space (all states of nature)

- Z: a r.v. w/ distribution P_θ for some $\theta \in \Theta$.
- $\mathcal{Z}$: sample space
- Experiment: the process that generates Z
- After observing Z, the econometrician takes an action a in some action space $\mathcal{A}$.
- Our payoff depends on our action a, and the "true" state-of-nature θ.

  Loss function $\mathcal{L}(\theta, a)$ is real-valued on $\mathcal{A} \times \Theta$.
- Let $d: \mathcal{Z} \to \mathcal{A}$: after observing $Z = z$, we choose action

  $d(z) \in \mathcal{A}$, $d(Z)$ = decision rule
- Risk (- Expected Utility) : $R(\theta, d) = \mathbb{E}_\theta[L(\theta, d(Z))]$

  decision-rule: fully-specified plan of what to do for $\forall z$

- **Estimate a Bernoulli Success problem**

(i) Set-up

$Z = (Z_1, \cdots, Z_N)'$: a sequence of independent Bernoulli draws w/ an unknown success probability of θ.

(ii) [Goal] Construct a point-estimate of θ based on Z.

expected loss ⇒ risk

$L(\theta, a) = (\theta - a)^2$ → cannot choose a to min $\mathcal{L}$ since θ unknown

$R(\theta, d) = \mathbb{E}_\theta[L(\theta, d(Z))]$

$= \mathbb{E}_\theta[(\theta - d(Z))^2]$

$= \mathbb{E}_\theta\{([\theta - \mathbb{E}_\theta[d(Z)]] - [d(Z) - \mathbb{E}_\theta[d(Z)]])^2\}$

$= (\theta - \mathbb{E}_\theta[d(Z)])^2 - 2[\theta - \mathbb{E}_\theta[d(Z)]]\underbrace{[\mathbb{E}_\theta[d(Z)] - \mathbb{E}_\theta[d(Z)]]}_{=0}$

$+ \mathbb{E}_\theta\{(d(Z) - \mathbb{E}_\theta[d(Z)])^2\}$

$\Rightarrow R(\theta, d) = \underbrace{(\theta - E_\theta[d(\mathbf{Z}_1)])^2}_{\text{bias squared}} + \underbrace{V_\theta(d(\mathbf{Z}_1))}_{\text{variance}}$

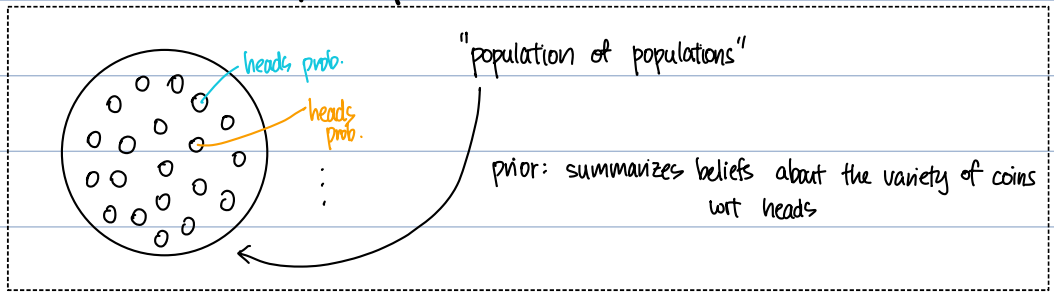
: low-risk decision rules involve a variance trade-off

(iii) [Goal] Find a decision rule $d(\mathbf{Z}_1)$ that "predicts" θ "well"

$\Rightarrow$ use Bayes' Rule

Bayes' Rule

- Prior distribution: summarizes our pre-experiment beliefs about θ



$\Rightarrow$ Prior: $\pi(\theta) = \begin{cases} 1 & 0 < \theta < 1 \\ 0 & \text{otherwise} \end{cases}$ ← controversial b/c consequential

- Likelihood: $l(\mathbf{Z}_1|\theta) = \prod_{i=1}^{N} \theta^{z_i}(1-\theta)^{1-z_i}$

$= \theta^{S_N}(1-\theta)^{N-S_N}$ $\left(S_N = \sum_{i=1}^{N} z_i : \text{sum of success of heads}\right)$

$(\mathbf{Z}_1 = (0, 1, 0, 0, \cdots, 1))$

- Posterior Distribution $(\mathbf{Z}_1 = \mathbf{z} = (0,1,1,0,0,0,\cdots)$

$$\pi(\theta|\mathbf{Z}_1) = \frac{l(\mathbf{z}|\theta)\cdot\pi(\theta)}{\int_{t=0}^{t=1} l(\mathbf{z}|t)\,\pi(t)\,dt} = \frac{\theta^{S_N}(1-\theta)^{N-S_N}}{\underbrace{\int_{t=0}^{t=1} t^{S_N}(1-t)^{N-S_N}\,dt}_{\text{Beta fnc: } \beta(S_N+1,\ N-S_N+1)}}$$

$\left(\because \beta(\alpha_1,\alpha_2) = \int_{t=0}^{t=1} t^{\alpha_1-1}(1-t)^{\alpha_2-1}\,dt\right)$

$\Rightarrow \theta|\mathbf{Z}_1 = \mathbf{z} \sim \text{Beta}(S_N+1,\ N-S_N+1)$: posterior dist.

∵ Note: median of a Beta(α_1, α_2) distributed r.v. $= \dfrac{\alpha_1 - 1/3}{\alpha_1 + \alpha_2 - 2/3}$

- I believe there is a 50% chance that θ is less than $\dfrac{S_N + 2/3}{N + 4/3}$

or { subjective prob. Statements
degree-of-belief statements

- Posterior mean & variance :

$$\begin{cases} \mathbb{E}[\theta \mid Z_1 = z] = \frac{S_N + 1}{N+2} \\ \mathbb{V}[\theta \mid Z_1 = z] = \frac{(S_N+1)(N-S_N+1)}{(N+2)^2(N+3)} \end{cases}$$

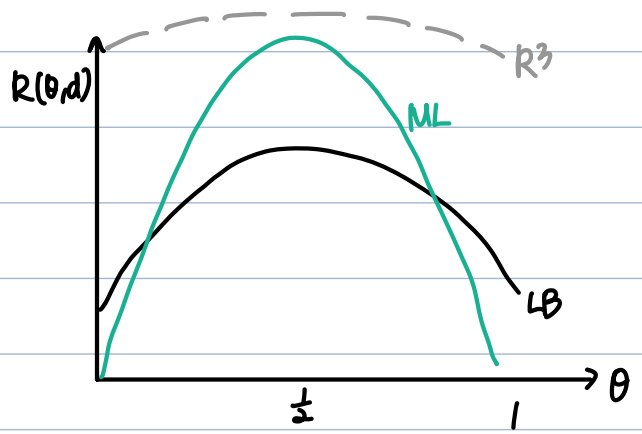
- Decision rule of "best guess" for θ

① $d_{LB}(Z_1) = \frac{S_N+1}{N+2}$ (↑ fix which depends on Z_1)

$$\begin{cases} \mathbb{E}_\theta[d_{LB}(Z_1)] = \frac{N\theta+1}{N+2} \\ \mathbb{V}_\theta[d_{LB}(Z_1)] = \frac{N\theta(1-\theta)}{(N+2)^2} \end{cases}$$

$$\Rightarrow R(\theta, d_{LB}) = \underbrace{\left(\frac{2\theta-1}{d+2}\right)^2}_{\text{bias sq.}} + \underbrace{\frac{N\theta(1-\theta)}{(N+2)^2}}_{\text{var.}}$$

② $d_{ML}(Z_1) = \frac{S_N}{N}$, $R(\theta, d_{ML}) = \frac{\theta(1-\theta)}{N}$



: not admissible → worse than another rule in every state of world (strictly dominated)

→ both rules are admissible (cross!)

if true value ≒ 0.5 : ML

≒ 0 or 1 : LB

(∵ neither dominates the other)

- Another example:

(Prior ①) $\theta \sim \text{Beta}(\alpha_1, \alpha_2)$

$\Rightarrow \theta \mid Z_1 = z \sim \text{Beta}(S_N + \alpha_1, N - S_N - \alpha_2)$

Imagine that you observed $\alpha_1 + \alpha_2$ coin flips of the coin in hand already and of these α_1 were heads

$\Rightarrow \alpha_1, \alpha_2$: pre-data "pseudo-counts"

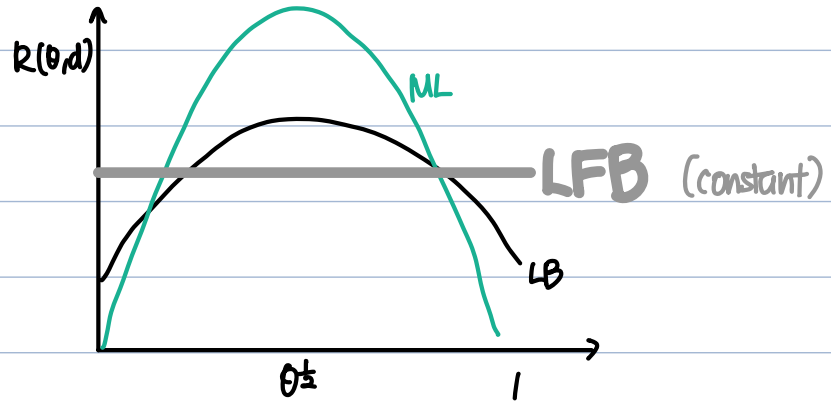
$\theta \sim U[0,1] = \text{Beta}(1,1)$ (Laplace-Bayes)

($\theta \sim \text{Beta}(\alpha_1, \alpha_2)$ as $\alpha_1, \alpha_2 \to 0$: improper prior, non-informative prior)

(Prior ②) Least Favorable Bayes :

$\theta \sim \text{Beta}(\alpha_1, \alpha_2)$, $\alpha_1 = \alpha_2 = \sqrt{\frac{N}{8}}$

$$\begin{cases} d_{LFB}(\mathbf{Z}) = \dfrac{S_N + \frac{1}{2}\sqrt{N}}{N + \sqrt{N}} \\ R(\theta, d_{LFB}) = \dfrac{N}{4(N+\sqrt{N})^2} \end{cases}$$



- Under a general $\theta \sim \text{Beta}(\alpha_1, \alpha_2)$, the posterior mean decision rule

$$d_{BB}(\mathbf{Z}) = \frac{S_N + \alpha_1}{N + \alpha_1 + \alpha_2}$$

$$= \frac{N}{N+\alpha_1+\alpha_2}\left(\frac{S_N}{N}\right) + \frac{\alpha_1+\alpha_2}{N+\alpha_1+\alpha_2}\left(\frac{\alpha_1}{\alpha_1+\alpha_2}\right)$$

$\frac{S_N}{N}$: MLE / sample mean; $\frac{\alpha_1}{\alpha_1+\alpha_2}$: prior mean; ∴ weighted avg btw sample mean & prior mean

(var of BB < var of MLE, but biased)

Q. Why is R highest for $\hat{\theta}_{ML}$ around $\theta = \frac{1}{2}$?

- risk = mean sq. error = bias² + variance
- MLE unbiased → risk entirely in variance only

⇒ coin flip 50%. (50% H, 50% T): much more uncertainty in the middle → huge variance

(cf) if Heads = 95%. → next flip almost surely H → almost no uncertainty

Q. What is LFB est. and why is R_{LFB} constant?

- generally, R changes w/ θ.
- LFB is "designed" s.t. variance − bias² tradeoff remains constant.

10/20 Lecture

Recap

- [Likelihood] $Z_i \mid \theta \overset{iid}{\sim} Ber(\theta)$, $i = 1, \cdots, N$, $Z = (Z_1, \cdots, Z_N)'$
- [Prior Dist.] $\theta \sim Beta(\alpha_1, \alpha_2)$
  - pulling a coin → Beta dist.
  - start flipping the coin → Bernoulli dist.
- [Posterior Dist] $\theta \mid Z \sim Beta(S_N + \alpha_1, N - S_N + \alpha_2)$ $\left(S_N = \sum_{i=1}^{N} Z_i\right)$

⇒ 2 ways of thinking about this:

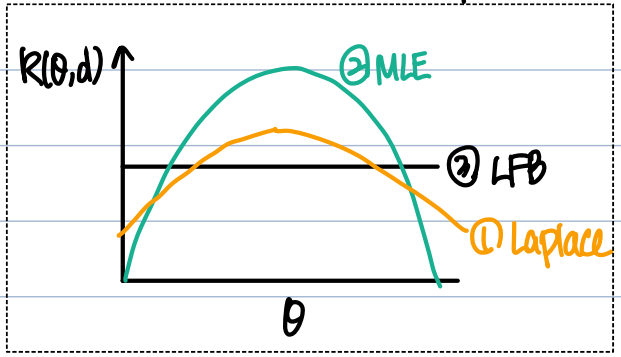
(i) Subjective Bayesian (Savage): Posterior as belief summary (believe ≈95% chance of θ being btw 0.45 ~ 0.55)

(ii) Frequentist
- don't like making probability statements about θ ∵ θ is a fixed population parameter (θ is in an interval either w/ prob. 1 or 0)
- (but might still use this procedure) want to use Bayesian apparatus to construct procedures w/ good frequentist properties

* Point Estimation as a Bayesian

- [Goal] What is our best guess of θ as a Bayesian?
- [Idea]
  - (i) we need a loss function (this gives content to "best")
  - (ii) Recap:
    1. $\theta \sim Beta(1,1) = U[0,1]$ (Laplace)
    2. $\theta \sim Beta(0,0)$ (improper prior) ⇒ "proper posterior" $\theta \mid Z \sim Beta(S_N, N - S_N)$
    3. $\theta \sim Beta\left(\sqrt{\frac{N}{4}}, \sqrt{\frac{N}{4}}\right)$ (Least Favorable Bayes): minimax

⇒ use posterior mean as point estimate



⇒ risk-functions cross: admissibility
- an admissible estimator is not uniformly dominated by another estimator

- Choose estimators w/ good average risk properties
  - → over possible values of θ ··· but what distribution? ⇒ use prior!

⁂ DIGRESSION

- Complete Class Thm: Any admissible estimator is a Bayes rule (or a limit of Bayes rule) wrt some prior
- Minimax Procedure $\inf_{d \in \mathcal{A}} \sup_{\theta \in \Theta} R(\theta, d)$

(admissible + constant risk ⇒ minimax)

$L(\theta, a) = (\theta - a)^2$, $R(\theta, d) = \mathbb{E}_{\vec{Z}|\theta}[L(\theta, d(\vec{Z})) \mid \theta]$

(θ: target, a: action, $d(\vec{Z})$: decision rule)

$$= \int L(\theta, d(\vec{z})) \cdot f(\vec{z}|\theta) \, d_m \vec{z}$$

($d_m \vec{z}$: for discrete & continuous; multivariate integral ($\vec{Z}$ = vector))

Average risk (Bayes risk) ← avg wrt prior

$$r(\pi, d) = \int R(\theta, d) \cdot \pi(\theta) \, dm(\theta)$$

($r(\pi, d)$: depends on prior; avg over all possible states of the world)

$$= \int\int L(\theta, d(\vec{z})) \cdot f(\vec{z}|\theta) \, dm(\vec{z}) \cdot \pi(\theta) \, dm(\theta)$$

$$= \int \underbrace{\left[\int L(\theta, d(\vec{z})) \cdot f(\vec{z}|\theta) \cdot \pi(\theta) \, dm(\theta) \right]}_{①} dm(\vec{z})$$

(switching order of integration)

Consider we observe $\vec{Z} = \vec{z}$ and choose $d_m(\vec{z})$ to minimize ①

$$d_\pi(\vec{z}) = \arg\min_{a \in \mathcal{A}} \int L(\theta, a) f(\vec{z}|\theta) \pi(\theta) \, dm(\theta)$$

minimizing ① = minimizing posterior expected loss (∵ posterior ∝ likelihood · prior)

$\pi(\theta|\vec{z}) \propto f(\vec{z}|\theta) \cdot \pi(\theta)$

(posterior ∝ likelihood · prior)

Bayes Rule:

$$\pi(\theta|\vec{z}) = \frac{f(\vec{z}|\theta) \cdot \pi(\theta)}{f(\vec{z})}$$

$$f(\vec{z}) = \int f(\vec{z}(t)) \cdot \pi(t) \, dt$$

$$= \arg\min_{a \in \mathcal{A}} \int L(\theta, a) \cdot \pi(\theta|\vec{z}) \, dm(\theta)$$

: choosing $d_\pi(\vec{z})$ to minimize ①

⇒ minimize posterior expected loss

Solving: $d_\pi(\vec{z}) = \arg\min \int L(\theta, a) \cdot \pi(\theta|\vec{z}) \, dm(\theta)$

$$= \mathbb{E}_{\theta|\vec{z}}[(\theta - a)^2 \mid \vec{Z} = \vec{z}]$$

(loss fnc; averaging over θ conditioned on $\vec{z}$)

$$= \mathbb{E}_{\theta|Z}\left[\underbrace{(\theta - \mathbb{E}[\theta|Z=z]}_{a} - \underbrace{(a - \mathbb{E}[\theta|Z=z])}_{b}\right)^2 \Big| Z=z\Big]$$

$$= \underbrace{(a - \mathbb{E}[\theta|Z=z])^2}_{\text{depends on our choice var. } a} + \underbrace{\mathbb{V}(\theta|Z=z)}_{\text{posterior conditional variance}} \Big\} \text{ doesn't depend on } a$$

$$+ 2(a - \mathbb{E}[\theta|Z=z])\underbrace{\mathbb{E}[\theta - \mathbb{E}[\theta|Z=z]|Z=z]}_{0}$$

$\Rightarrow d_\pi(z) = a^* = \mathbb{E}[\theta|Z=z]$

Under squared error loss,
the posterior mean minimizes posterior expected loss.

Plan of Action: To minimize Bayes risk,
(i) we use the likelihood + our prior to compute the posterior,
(ii) then we use the posterior mean as our decision ("point estimate")

- Main Points
  - ① Bayes risk allows us to rank decision rules
  - ② Ranking depends on our prior

$\Rightarrow$ Alternative idea: look for minimax procedure (← no prior needed here)

$$\inf_{d \in \mathcal{A}} \sup_{\theta \in \Theta} R(\theta, d)$$

: LFB

* K-normal means (James-Stein estimator)

- Set-up
  - $F_{X,Y}$ unknown { Y : some outcome of interest (earnings); X : a vector of covariates
  - Training sample : $Y_{(N\times 1)} = (Y_1, \cdots, Y_N)'$, $X_{(N\times K)} = (\underset{(K\times 1)}{X_1}, \cdots, X_N)'$
    $\Rightarrow$ use training sample to construct good predictions of new values of Y
  - Assume that the covariate (feature) values for any future draw of Y will be in X
  - Condition on X $\rightarrow$ treat X as non-stochastic

- **Notation**

$$\underset{(N\times 1)}{m} = \underset{(N\times 1)}{\mathbb{E}[Y]} = \mathbb{E}[Y \mid X]$$

(X: non-stochastic; but depends on specific values of X)

$\hat{m}_i = \hat{m}(X_i)$: our prediction for Y when $X = X_i$

$$\underset{(N\times 1)}{\hat{m}} = \begin{pmatrix} \hat{m}(X_1) \\ \vdots \\ \hat{m}(X_N) \end{pmatrix}$$

→ in general, $\hat{m}$ will depend on $\underset{(N\times 1)}{Y}$ $\left(\hat{m} = d(Y) = d(Y, X)\right)$ ∵ non-stochastic

- **Squared Error Loss** $\left(\|m\| = \sum_{i=1}^{N} m_i^2\right)^{\frac{1}{2}}$

- Risk : $\mathbb{E}[\|(\hat{m} - m)\|^2]$ ⇒ want to choose $\hat{m}$ to be close to $m = \mathbb{E}[Y] = \mathbb{E}[Y \mid X]$

$$= \mathbb{E}\left[\underbrace{\sum_{i=1}^{N} (\hat{m}(X_i) - m(X_i))^2}_{\text{loss}}\right]$$

→ sq. error loss but for N different targets

$$= \begin{bmatrix} m(X_1) \\ \vdots \\ m(X_N) \end{bmatrix} = \begin{bmatrix} \mathbb{E}[Y_1] \\ \vdots \\ \mathbb{E}[Y_N] \end{bmatrix}$$

$$= \sum_{i=1}^{N} \mathbb{E}[(\hat{m}(X_i) - m(X_i))^2]$$

* 11/4 Lecture

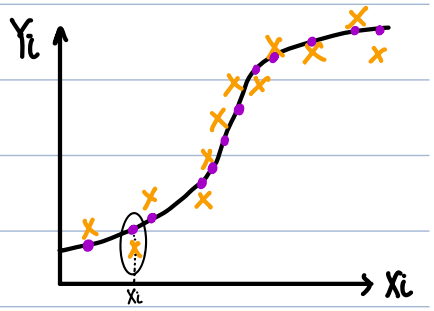
* Set-up to the K-Normal mean

- Training sample : (X_i, Y_i) (Y_i: outcome) $i=1,\cdots,N$, $(X_i, Y_i) \overset{iid}{\sim} F_{X,Y}$
- treat $\underset{(N\times K)}{X} = (X_1, \cdots, X_N)'$ as non-stochastic ("fixed design matrix")
- $m_i = \mathbb{E}[Y_i]$ ∵ note $\mathbb{E}[Y_1] \neq \mathbb{E}[Y_{1n}]$ in general since $X_1 \neq X_{1n}$ in general
- sometimes in an abuse of notation; $m_i = \mathbb{E}[Y|X=X_i] = \mathbb{E}[Y_i]$
- $\underset{(N\times 1)}{Y}$ vector of outcome, $\underset{(N\times 1)}{m} = \mathbb{E}[Y]$

- Prediction Goal : observe $X = X_{19}$, predict Y
- Loss function :

$$\|\hat{m} - m\|^2 = \sum_{i=1}^{N} \underbrace{(\hat{m}_i - m_i)^2}_{d_i(Y)}$$

※ Side-note:
$\begin{cases} Y_i \overset{d}{=} Y_i^* & \because m_i = \mathbb{E}[Y_i] = \mathbb{E}[Y_i^*] \\ Y_i \overset{d}{\neq} Y_j \text{ for } i \neq j & \because X_i \neq X_j \end{cases}$



$\mathbb{E}[\|\hat{m} - m\|^2]$: Risk | MSE

apparent error

→ we want something that works well on avg!

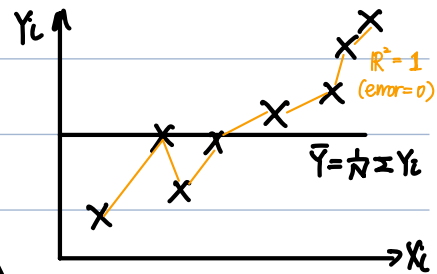
[OLS ≠ BLUE under joint random sampling → EXAM!!]

$$\mathbb{E}[\|Y - \hat{m}\|^2] = \mathbb{E}[\|(Y-m) + (m-\hat{m})\|^2]$$

expected in sample fit

$$= \underbrace{\mathbb{E}[\|Y-m\|^2]}_{①} + \underbrace{\mathbb{E}[\|\hat{m}-m\|^2]}_{②=\text{Risk fnc}} + \underbrace{2\mathbb{E}[(Y-m)'(m-\hat{m})]}_{③\ \text{cross-product fnc}}$$

scale of noise



since $Y_i = m(X_i) + \sigma u_i$, $u_i \sim N(0,1)$, $u_i | X \sim N(0,1)$

some mean + noise

→ $Y_i \sim N(m(X_i), \sigma^2)$: classical regression model

① $\mathbb{E}[\|Y-m\|^2] = \mathbb{E}[\sum_{i=1}^{N}(Y_i - m_i)^2] = \sum_{i=1}^{N} \underbrace{\mathbb{E}[(Y_i-m_i)^2]}_{Var(Y_i)} = N\sigma^2$

③ $\mathbb{E}[(Y-m)'(m-\hat{m})] = \mathbb{E}[(Y-m)'m] - \underbrace{\mathbb{E}[(Y-m)'\hat{m}]}_{④}$

$= \underbrace{\mathbb{E}[Y-m]'}_{(N\times 1)} m - \mathbb{E}[(Y-m)'\hat{m}]$ ($\mathbb{E}[Y-m] = 0$)

$(\mathbb{E}[Y] - m)'m = 0$

④ $\mathbb{E}[(Y-m)'\hat{m}] = \mathbb{E}[\sum(Y_i - m_i)\hat{m}_i] = \sum_{i=1}^{N} \underbrace{\mathbb{E}[(Y_i - m_i)\hat{m}_i]}_{⑤}$

$= \sum_{i=1}^{N} Cov(Y_i, \hat{m}_i)$

$= \begin{pmatrix} Y_1 - m_1 \\ \vdots \\ Y_N - m_N \end{pmatrix}' \begin{pmatrix} \hat{m}_1 \\ \vdots \\ \hat{m}_N \end{pmatrix}$

⑤ $\mathbb{E}[(Y_i - m_i)\hat{m}_i] = \mathbb{E}[Y_i \hat{m}_i] - m_i \mathbb{E}[\hat{m}_i]$ (m_i: not random)

$= \mathbb{E}[Y_i \hat{m}_i] - \mathbb{E}[Y_i]\mathbb{E}[\hat{m}_i]$

$= \mathrm{Cov}(Y_i, \hat{m}_i)$

$\Rightarrow \mathbb{E}[\|Y - \hat{m}\|^2] = \underset{①}{N\sigma^2} + \underset{②}{\mathbb{E}[\|\hat{m} - m\|^2]} - \underset{③}{2\sigma^2 \cdot df(\hat{m})}$

apparent error (aka. expected in sample fit)

where $df(\hat{m}) = \sum \frac{\mathrm{Cov}(Y_i, \hat{m}_i)}{\sigma^2}$

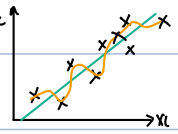
① noise floor ② risk / MSE ③ degrees-of-freedom (model complexity)

$\hookrightarrow$ $\mathbb{E}[\|\hat{m} - m\|^2] = -N\sigma^2 + \underset{\text{apparent error}}{\mathbb{E}[\|Y - \hat{m}\|^2]} + \underset{\text{complexity/df}}{2\sigma^2 \cdot df(\hat{m})}$

Trade-off btw "in sample fit" and "model complexity"

* Connect-the-dots est: $\mathbb{E}[\|Y - \hat{m}\|^2] = 0$ but complexity ⇑

* Sample mean : complexity ⇓ but $\mathbb{E}[\|Y - \hat{m}\|^2]$ huge



* K-normal means

- **Set-up**
  - $Y_i = m(X_i) + \sigma u_i$, $u_i | X \sim N(0,1)$ or $u_i \sim N(0,1)$
    $i = 1, \cdots, N$, $X_i \in \mathcal{X}$, $x \in \mathcal{X}$
  - assumption on m : $m(x) \approx \sum_{k=1}^{K} \alpha_k g_k(x)$ w/ $g_k(x)$ known basis functions, α_k unknown coefficient

smoothing assumption

$N = 1000$, $K = 200$ // (ex) $g_k(x) = x^{k-1}$ where $k = 1, \cdots, K$: series basis

- **Gram-Schmidt Orthogonalization**
  - Euclidean space

$$\langle f, g \rangle = \frac{1}{N} \sum_{i=1}^{N} f(X_i) g(X_i) \quad \text{where } f = \begin{pmatrix} f(X_1) \\ \vdots \\ f(X_N) \end{pmatrix}, \underset{N \times 1}{g} = \begin{pmatrix} g_k(X_1) \\ \vdots \\ g_k(X_N) \end{pmatrix}$$

$$\|f\| = \langle f, f \rangle^{\frac{1}{2}} = \left[\frac{1}{N} \sum_{i=1}^{N} f(x_i)^2\right]^{\frac{1}{2}}$$

$$\mathrm{proj}_f(g) = \pi(g|f) = \frac{\langle f, g \rangle}{\langle f, f \rangle} \cdot f \quad \leftarrow \text{OLS}$$

$u = g - \mathrm{proj}_f(g)$: Projection error

[Pf] $\langle u_1, f \rangle = \langle g - \text{proj}_f(g), f \rangle$

$= \langle g - \frac{\langle f, g \rangle}{\langle f, f \rangle} f, f \rangle$

$= \langle g, f \rangle - \frac{\langle f, g \rangle \langle f, f \rangle}{\langle f, f \rangle}$ (bilinearity)

$= 0$

Recursive Basis function

- $f_1(x) = g_1(x)$: generally a constant
- $f_2(x) = g_2(x) - \frac{\langle g_2, f_1 \rangle}{\langle f_1, f_1 \rangle} f_1(x)$

(assume $g_k(x) = x^{k-1}$; $1, x, x^2, \cdots \Rightarrow f_1(x) = \mathbb{1}$

$f_2(x) = x - \bar{x}$ $\left(\bar{x} = \frac{1}{N}\sum x_i\right)$)

- $f_3(x) = g_3(x) - \frac{\langle g_3, f_2 \rangle}{\langle f_2, f_2 \rangle} f_2(x) - \frac{\langle g_3, f_1 \rangle}{\langle f_1, f_1 \rangle} f_1(x)$

  ⋮
- $f_K(x) = g_K(x) - \sum_{k=1}^{K-1} \frac{\langle g_K, f_k \rangle}{\langle f_k, f_k \rangle} f_k(x)$

$\langle f_1, f_2 \rangle = 0$, $\langle f_1, f_3 \rangle = 0$, $\cdots$ $\langle f_k, f_j \rangle = 0$ $k \neq j$, $k, j < K$

$\Rightarrow \underbrace{g_1, \cdots, g_K}_{m(x)} \Rightarrow \underbrace{f_1, \cdots, f_K}_{\text{orthogonal to each other}}$

$\Rightarrow$ orthogonality : $\phi_1(x) = \frac{f_1(x)}{\|f_1(x)\|}$, $\cdots$ $\phi_K(x) = \frac{f_K(x)}{\|f_K(x)\|}$

$\|\phi_k\| = 1$

- **Maximum Likelihood Estimator (MLE)**

$m(x) = \sum_{k=1}^{K} \theta_k \phi_k(x)$

$(Y_i = m(X_i) + \sigma u_i, \; u_i \sim N(0,1), \; \sigma^2 \text{ known})$

$W(X)_{(k \times 1)} = \begin{pmatrix} \phi_1(x) \\ \vdots \\ \phi_k(x) \end{pmatrix}$, $\theta_{(k \times 1)} = \begin{pmatrix} \theta_1 \\ \vdots \\ \theta_k \end{pmatrix}$

log likelihood fnc:

$$\ell(Y \mid X; \theta) = -\frac{N}{2}\ln 2\pi - N\ln\sigma - \frac{1}{2\sigma^2}\sum_{i=1}^{N}\left(Y_i - \underbrace{w(X_i)'\theta}_{W_i}\right)^2$$

$$\hat{\theta}_{ML} = \left[\frac{1}{N}\Sigma W_i W_i'\right]^{-1} \cdot \left[\frac{1}{N}\Sigma W_i Y_i\right]$$

$$\hat{m}(x) = w(x)'\hat{\theta}_{ML}$$

$$\boxed{\frac{1}{N}\sum_{i=1}^{N} W_i W_i' = I_k}$$

↳ Show this: $\frac{1}{N}\sum_{i=1}^{N}$

$$\begin{bmatrix} \phi_1(X_1)^2 & \phi_1(X_1)\phi_2(X_2) & \cdots \\ \phi_2(X_2)\phi_1(X_1) & \phi_2(X_2)^2 & \vdots \\ \vdots & \ddots & \\ \phi_N(X_N)\phi_1(X_1) & \cdots & \phi_N(X_N)^2 \end{bmatrix}$$

$$\hat{\theta}_{ML} = Z_1 = \frac{1}{N}\Sigma W_i Y_i$$

⇒ note that: $\underset{(K\times 1)}{Z_1} \sim N\left(\theta, \frac{\sigma^2}{N} I_k\right)$

K-normal mean

* 11/6 Lecture

- **Set-up**

$Y_i = m(X_i) + \sigma U_i$, $\quad U_i \mid X \sim \mathcal{N}(0,1)$ $\quad i=1,\cdots,N$, $\quad X$ non-stochastic, fixed

σ^2 known

Assume $m(x) = \sum_{k=1}^{K} \alpha_k g_k(x)$

Gran-Schmidt orthonormalization : $m(x) = \sum_{k=1}^{K} \theta_k \phi_k(x)$ (orthonormal basis)

$\hat{\theta}_{ML} = \left[\frac{1}{N}\sum_{i=1}^{N} W_i W_i'\right]^{-1} \cdot \left[\frac{1}{N}\sum_{i=1}^{N} W_i Y_i\right]$ where $\frac{1}{N}\sum_{i=1}^{N} W_i W_i' = I_N$ (by orthonormality (by construction)), $W_i = w(X_i)$

$:= Z_1 = \frac{1}{N}\sum_{i=1}^{N} W_i Y_i$ (Gaussian)

$Z_1 \sim \mathcal{N}\left(\theta, \frac{\sigma^2}{N} I_K\right)$ (Gaussian)

$$\left(\underset{(K\times 1)}{W(X)} = \begin{pmatrix} \phi_1(x) \\ \vdots \\ \phi_K(x) \end{pmatrix}, \quad \theta = \begin{pmatrix} \theta_1 \\ \vdots \\ \theta_K \end{pmatrix}\right)$$

- **Unbiasedness of $\hat{\theta}_{ML}$** (remember X is non-stochastic) (if X is stochastic, mean of θ doesn't always exist)

$E[Z \mid X] = \frac{1}{N}\sum_{i=1}^{N} W_i E[Y_i \mid X]$ $\quad (Y_i = W_i'\theta + \sigma U_i)$ (why?)

(can get rid of X since X non-stochastic, but to emphasize conditioning on X)

$= \underbrace{\frac{1}{N}\sum_{i=1}^{N} W_i W_i'}_{I_N}\theta + \frac{\sigma}{N}\sum_{i=1}^{N} W_i \underbrace{E[U_i \mid X]}_{0}$

$= \theta$: OLS is unbiased when X is non-stochastic

Covariance matrix

$V[Z_k \mid X] = V\left[\frac{1}{N}\sum_{i=1}^{N} \underbrace{\phi_k(X_i)}_{W_{ki}} Y_i \mid X\right]$ (diagonal elements)

$\because Y_i$'s independent : $V(\Sigma) = \Sigma V$

$= \frac{1}{N^2}\sum_{i=1}^{N} V(\phi_k(X_i) Y_i \mid X)$ $\quad V(aX+bY) = a^2 V(X) + b^2 V(Y) + 2ab\,C(X,Y)$

$= \frac{\sigma^2}{N} \underbrace{\frac{1}{N}\sum_{i=1}^{N} \phi_k(X_k)^2}_{\|\phi_k\|^2 = 1}$

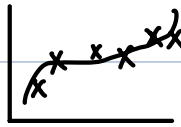
$= \frac{\sigma^2}{N}$ diagonal elements $(k=1,\cdots,K)$

$C(Z_j, Z_k \mid X) = C\left[\frac{1}{N}\sum_{i=1}^{N} \phi_j(X_i) Y_i, \frac{1}{N}\sum_{i=1}^{N} \phi_k(X_i) Y_i \mid X\right]$ (off diagonal elements = 0)

$= \frac{\sigma^2}{N} \cdot \underbrace{\frac{1}{N}\sum_{i=1}^{N} \phi_j(X_i)\cdot\phi_k(X_i)}_{\langle \phi_j, \phi_k \rangle = 0}, \quad j \neq k$

j-th k k-th elements of the orthogonal basis : orthogonal by construction → inner product = 0

$\Rightarrow \underset{(K\times 1)}{\hat{\theta}_{ML} = Z_1} \sim \mathcal{N}\left(\theta, \frac{\sigma^2}{N} I_K\right)$



- Getting back to "Risk" (motivation: forecasting problem)

$$\underbrace{\|\hat{m} - m\|^2}_{\text{squared error loss}} = \frac{1}{N}\sum_{i=1}^{N}(\underbrace{w_i'\theta_{ML}}_{m(x_i)} - w_i'\theta)$$

$\hat{\theta}_{ML} \equiv Z$ (to emphasize $\hat{\theta}_{ML}$ is r.v.)

$$w_i = \begin{pmatrix} \phi_1(x_i) \\ \vdots \\ \phi_K(x_i) \end{pmatrix}$$

$$= \frac{1}{N}\sum_{i=1}^{N} w_i'(Z - \theta)$$

$$= \frac{1}{N}\sum_{i=1}^{N}\left(\sum_{k=1}^{K}\underbrace{\phi_k(x_i)}_{\text{orthonormal to one another}}\cdot(Z_k - \theta_k)\right)^2$$

$$= \sum_{k=1}^{K}(Z_k - \theta_k)^2$$

Loss function: $L(Z,\theta) = \|Z-\theta\|^2 = \sum_{k=1}^{K}(Z_k - \theta_k)^2$

Risk function: $R(Z,\theta) = \mathbb{E}[L(Z,\theta)]$

$$\mathbb{E}\left[\sum_{k=1}^{K}\underbrace{(Z_k - \theta_k)^2}_{\text{multivariate normal, identity cov. matrix}}\right] = \sum_{k=1}^{K}\mathbb{E}[(Z_k-\theta_k)^2] = \frac{K\sigma^2}{N}$$

The Risk of MLE varies with:
- (i) model complexity, K ↑ (not a choice!)
- (ii) amount of data, N ↓
- (iii) noise floor, σ^2 ↑

* James-Stein Type Estimators

- Loss & Risk function

Loss: $\mathcal{L} = \{\underset{K\times K}{C}Z : C = \text{diag}\{c_1, \cdots, c_K\},\ c_k \in [0,1],\ k=1,\cdots,K\}$

Risk: $\mathbb{E}\left[\sum_{k=1}^{K}(c_k Z_k - \theta_k)^2\right] = \mathbb{E}\left[\sum_{k=1}^{K}(c_k(Z_k-\theta_k) - (1-c_k)\theta_k)^2\right]$

$$= \underbrace{\frac{\sigma^2}{N}\sum_{k=1}^{K}c_k^2}_{V(C,K)} + \underbrace{\sum_{k=1}^{K}(1-c_k)^2\theta_k}_{\text{Bias}^2}$$

- "Oracle"

(Thought experiment: think of a new θ est. → which C would minimize risk?)

$$c_k^* = \frac{\theta_k^2}{\frac{\sigma^2}{N} + \theta_k^2},\quad k=1,\cdots,K \quad \Rightarrow \quad C^*\cdot Z = \hat{\theta}_0$$

↳ imagine it's like a constant

if θ big → deviation = lots of bias → scale a lot
small → not much bias, more variance → don't scale a lot

(plug c_k^* back into the risk fnc)

$$\Rightarrow \inf_{\hat{\theta} \in \mathcal{I}} R(\hat{\theta}_0, \theta) = \frac{\sigma^2}{N} \sum_{k=1}^{K} \left(\frac{\theta_k^2}{\frac{\sigma^2}{N} + \theta_k^2} \right) < \frac{K\sigma^2}{N} \left(= R_{ML} \right)$$: "bound"

- **Stein - Unbiased - Risk - Estimate (SURE)** { ✓ model selection, ✓ cross validation, ✓ cross fitting.

$Z \sim N(\theta, \sigma^2 I_K)$, $\hat{\theta} = \hat{\theta}(Z)$, $g(Z) = \hat{\theta}(Z) - Z$

$$\hat{R}_{SURE}(Z) = -K\sigma^2 + 2\sigma^2 \cdot \sum_{k=1}^{K} \frac{\partial g_k(Z)}{\partial Z_k} + \sum_{k=1}^{K} (\hat{\theta}_k - Z_k)^2$$: unbiased risk estimator

$$\mathbb{E}[\hat{R}_{SURE}(Z)] = \mathbb{E}[\|\hat{\theta} - \theta\|^2]$$: UNBIASED

- **James-Stein Type Estimators**

$$\hat{\theta}_{JS}(Z) = \left(1 - \frac{(K-2)}{Z'Z} \frac{\sigma^2}{N}\right) Z$$

$$g_{JS}(Z) = -\left(\frac{(K-2)\sigma^2/N}{Z'Z}\right) Z \quad \Rightarrow \quad \sum_{k=1}^{K} \frac{\partial g_k(Z)}{\partial Z_k} = -\frac{(K-2)^2 \cdot \frac{\sigma^2}{N}}{Z'Z}$$

$$\underbrace{\hat{R}_{SURE}(Z)}_{\hat{\theta}_{JS}} = \frac{K}{N}\sigma^2 - \frac{(K-2)^2}{Z'Z} \cdot \frac{\sigma^4}{N^2}$$

$$R(\hat{\theta}_{JS}, \theta) = \mathbb{E}[\hat{R}_{SURE} \mid Z] = \underbrace{\frac{K}{N}\sigma^2}_{R(\hat{\theta}_{ML}, \theta)} - (K-2)^2 \cdot \frac{\sigma^4}{N^2} \cdot \underbrace{\mathbb{E}\left[\frac{1}{Z'Z}\right]}_{\sim \chi^2} \quad \text{①} \qquad (\text{for } K \geq 3) \quad < R(\hat{\theta}_{ML}, \theta)$$

$Z_k^2 = \frac{\sigma^2}{N}\left(\frac{\theta_k}{\sigma/\sqrt{N}} + u\right)^2$, $u \sim N(0,1)$, $Z_k \sim N\left(\theta_k, \frac{\sigma^2}{N}\right)$

$Z'Z \sim \frac{\sigma^2}{N} V$ where V is a non-central χ^2 r.v. w/ K degrees of freedom and non-centrality parameter $p = \sum_{k=1}^{K} N(\theta_k^2/\sigma^2)$

$V \sim \chi^2_{K+2W}$ where $W \sim \text{Poisson}\left(\frac{p}{2}\right)$, $\mathbb{E}[W] = \frac{p}{2}$: fact

$$\mathbb{E}\left[\frac{1}{Z'Z}\right] = \frac{N}{\sigma^2}\mathbb{E}\left[\frac{1}{V}\right]$$

$$(\text{L.I.E.}) = \frac{N}{\sigma^2}\mathbb{E}\left[\underbrace{\mathbb{E}\left[\frac{1}{V} \mid W\right]}_{\text{conditioned on } W,\ \text{becomes central } \chi^2 \text{ r.v.}}\right]$$

∵ $K>2$, $S \sim \chi^2_K$, $\mathbb{E}\left[\frac{1}{S}\right] = \frac{1}{K-2}$

$$= \frac{N}{\sigma^2}\mathbb{E}\left[\frac{1}{K-2+2W}\right]$$

$$(\text{Jensen's Inequality}) \geq \frac{N}{\sigma^2} \frac{1}{K-2+N\sum_{k=1}^{K}(\theta_k^2/\sigma^2)} \quad \left(= \frac{N}{\sigma^2} \cdot \frac{1}{K-2+2\mathbb{E}[W]}\right)$$

$$= \frac{1}{(K-2)\cdot\frac{\sigma^2}{N} + \|\theta\|^2}$$

$$R(\hat{\theta}_{JS}, \theta) \leq \frac{K}{N}\sigma^2 - \frac{(K-2)^2 \frac{\sigma^4}{N^2}}{(K-2)^2 \frac{\sigma^2}{N} + \|\theta\|^2} \leq R(\hat{\theta}_{ML}, \theta)$$

$$= \frac{2}{N}\sigma^2 + \frac{(K-2)\sigma^2/N \, \|\hat{\theta}\|^2}{(K-2)\frac{1}{N} + \|\theta\|}$$

⇒ When $k \geq 3$, $R(\hat{\theta}_{JS}, \theta) \leq R(\hat{\theta}_{ML}, \theta)$ for $\forall \theta \in \Theta$

∴ MLE is inadmissible? ($\exists \hat{\theta}_{JS}$ that dominates $\hat{\theta}_{ML}$)

※ Ensemble Risk $\sum_{k=1}^{K} (\hat{\theta}_{Jk} - \theta_k)^2$ vs Component Risk $(\hat{\theta}_{Jk} - \theta_k)^2$

→ shrink

The James–Stein estimator made this point dramatically in 1961, and made it in the context of just a few unknown parameters, not hundreds or thousands. It begins the story of *shrinkage estimation*, in which deliberate biases are introduced to improve overall performance, at a possible danger to individual estimates. Chapters XXX and XXX will carry on the story in its modern implementations.

* 11/13 Lecture

※ Finishing up K-Normal means (& JS est)

- **Set-up**

$Z \sim N(\theta, \frac{\sigma^2}{N} I_K)$, $Y = m(X) + \sigma u$, $u \sim N(0, \sigma^2)$

$m(X) = \sum_{k=1}^{K} \varphi_k(x)\, \theta_k$ (large enough s.t. we don't have to worry about whether K is suff. large) ($K=200$, $N=1000$)

$\| \hat{m} - m \|^2 = \sum_{k=1}^{K} (\hat{\theta}_k - \theta_k)^2$: sq. error loss

$\hat{\theta}_{ML} = Z = \underset{(N\times K)(N\times 1)}{W'Y}$ where $(W'W) = I_K$

→ MLE is inadmissible

→ showed this by demonstrating that the risk function of $\hat{\theta}_{JS}$ lies below that of $\hat{\theta}_{ML}$ for $\forall \theta \in \Theta$ $(K \geq 3)$

- **JS Estimator**

$\hat{\theta}_{JS}(Z) = \left(1 - \frac{(K-2)}{Z'Z}\frac{\sigma^2}{N}\right) Z$ ("shrinkage estimator")

$\mathcal{L} = \{ \underset{K\times K}{C} Z : C = \text{diag}\{c_1, \cdots, c_K\},\ c_k \in [0,1],\ k=1,\cdots,K \}$

multiply each coefficient by 0~1 { x1 : MLE (unbiased) ; x0 : zero } ↑ bias ↑ retain all variance

Recall: we consider "oracle" choice for C.

Use SURE to think about how to choose C in practice.

$\hat{R}_{SURE}(Z, c) = \frac{K}{N}\sigma^2 - \frac{2}{N}\sigma^2 \sum_{k=1}^{K}(1-c_k) + \sum_{k=1}^{K}(\hat{\theta}_k - Z_k)^2$ $(\hat{\theta} = CZ)$ (*)

$\left(R_{SURE}(Z) = K\sigma^2 + 2\sigma^2 \sum_{k=1}^{K} \frac{\partial g_k(Z)}{\partial Z_k} + \sum_{k=1}^{K}(\theta_k - Z_k)^2\right)$

$E[\hat{R}_{SURE}(Z, c)]$: verify our earlier calculation.

Choose $\hat{C}$ to minimize an estimate of risk

$(*) = \underbrace{\frac{\sigma^2}{N}\sum_{k=1}^{K} c_k^2}_{\text{variance reduction}} + \underbrace{\sum_{k=1}^{K}\left(Z_k^2 - \frac{\sigma^2}{N}\right)(1-c_k)^2}_{\text{estimate of bias sq.}}$ (heuristic)

$\left(E[Z_k^2] = \underbrace{V(Z_k)}_{\sigma^2/N} + \underbrace{E[Z_k]^2}_{\theta_k^2}\right)$

$\left(\theta_{k,0} = c_k Z_k,\ V(\hat{\theta}_{k,0}) = c_k^2 \cdot V(Z_k) = c_k^2 \cdot \frac{\sigma^2}{N}\right)$

(FOC) $\boxed{\hat{C}_k = 1 - \frac{N^{-1}\sigma^2}{Z_k^2}}$ $k=1,\cdots,K$; $Z_k^2 = \frac{\sigma^2}{N} + \theta_k^2 + \text{noise}$

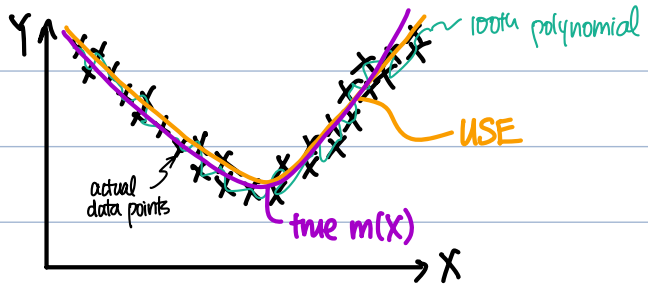
⇒ USE : Universal Series Estimator

- How to choose optimal c_k?
  - How does it connect w/ oracle?
    - ⇒ oracle shrinks a lot when θ_k is small relative to "noise", σ^2
  - Oracle: choose c^*_k to minimize Risk (evaluating risk requires omniscience)
  - USE: choose $\hat{c}_k$ to minimize SURE (b/c we don't know the actual risk!) (= Empirical Risk Minimization)

∴ Oracle: $c_k^* = \frac{\theta_k^2}{\sigma^2/N + \theta_k^2}$

if we knew the truth, this is the optimal shrinkage coefficient (let $c \approx 0$ if true $\theta \approx 0$)

- **Summary**
  - Setting: $\{X_i, Y_i\}_{i=1}^N$, $Y_i = m(X_i) + \sigma u_i$,
    $m(X) = \sum_{k=1}^K P_k(X_i)\,\theta_k$
  - Steps:
    1. use OLS to compute $\hat{\theta}_{MLE} = \mathbb{Z}$ → inadmissible (that's why we start looking into shrinking the coefficients)
    2. Compute $\hat{c}_k = 1 - \frac{\sigma^2/N}{\mathbb{Z}_k^2}$ $(k = 1, \cdots, K)$
    3. $\hat{m}^{USE}(X_i) = \sum_{k=1}^K P_k(X_i)\,\hat{\theta}_k^{USE}$
       where $\hat{\theta}_k^{USE} = \hat{c}_k \cdot \hat{\theta}_{MLE} = \hat{c}_k \cdot \mathbb{Z}$
       $\hat{m}^{OLS}(X_i) = \sum_{k=1}^K P_k(X_i)\,\hat{\theta}_k^{OLS}$



* Bayesian Bootstrapping

- **Intro**
  - measuring uncertainty (want to start thinking about uncertainty wrt "sample in hand")
    - Bayesian
    - Frequentist (large sample)
  - Why Bayesian bootstrapping:
    - (i) a way to quantify uncertainty from Bayesian perspective
    - (ii) introduction to simulation (w/ numerical integration)
    - (iii) Bayesian b.s. is also a weighted frequentist b.s.

• Set-up

$\underset{(K\times 1)\times 1}{Z} = (\underset{(K\times 1)}{X}{}', \underset{(1\times 1)}{Y})' \in \mathbb{Z} = \{z_1, \cdots, z_J\}$ (J is massive)

(r.v. vector) — support is a set

(ex) Y : earnings, $Y_i \in \{0, 1, \cdots, 88{,}000\}$

$K=3$:
- X_1 : gender, $X_{1i} \in \{0, 1\}$
- X_2 : yrs of schooling, $X_{2i} \in \{0, \cdots, 20\}$
- X_3 : AFQT score, $X_{3i} \in \{1, \cdots, 100\}$

$J = 2 \times 21 \times 100 \times 88{,}001 = 369{,}604{,}200$

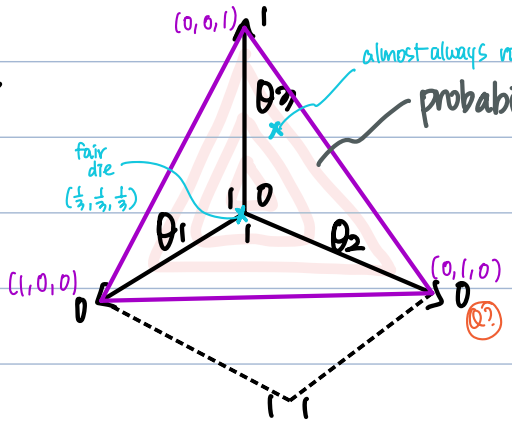
$Z \sim F_Z$

$\Pr[Z = z_j \mid \theta] = \theta_j$ where $\theta = (\theta_1, \cdots, \theta_J)'$ (population shares)

$\theta \in \Theta = \mathbb{S}^{J-1} = \{\theta = (\theta_1, \cdots, \theta_J)' \in \mathbb{R}^J,\ \theta_j \geq 0,\ \sum_{j=1}^{J} \theta_j = 1\}$

probability (unit) simplex — all non-negative — sum to 1

imagine a 6-sided die but only w/ 3 numbers (1,1,2,2,3,3)



⇒ Each point of the simplex is a possible population. (state of the world)

• Likelihood (multinomial) (∵ Bayesian!)

Set-up: $\theta \in \Theta = \mathbb{S}^{J-1}$, $\Pr(Z = z_j \mid \theta) = \theta_j$ (θ : "population")

$\underset{(N\times 1)}{\mathbb{Z}} = (z_1, \cdots, z_N) = z = (1, 1, 1, 3, 2, 1, 3, 2, \cdots)$

Likelihood : $f(z \mid \theta) = \underbrace{\frac{N!}{N_1! \times \cdots \times N_J!}}_{\text{capturing different ordering of draws}} \cdot \prod_{j=1}^{J} \theta_j^{N_j}$

Prior (Dirichlet) : $\underbrace{\pi(\theta_1, \cdots, \theta_J)}_{\text{describes a "population of populations"}} = \frac{\Gamma(\sum_{j=1}^{J} \alpha_j)}{\prod_{j=1}^{J} \Gamma(\alpha_j)} \cdot \prod_{j=1}^{J} \theta_j^{\alpha_j - 1}$ [gamma fnc]

("population of populations": from a bag of dice → pick a die ("population") → roll a die ("sample"))

assigns probability mass to simplex