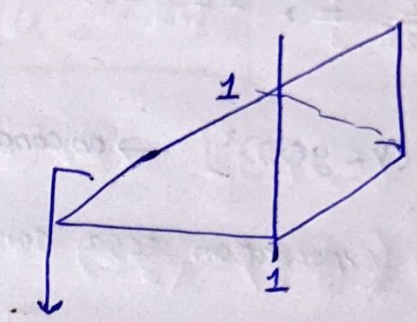
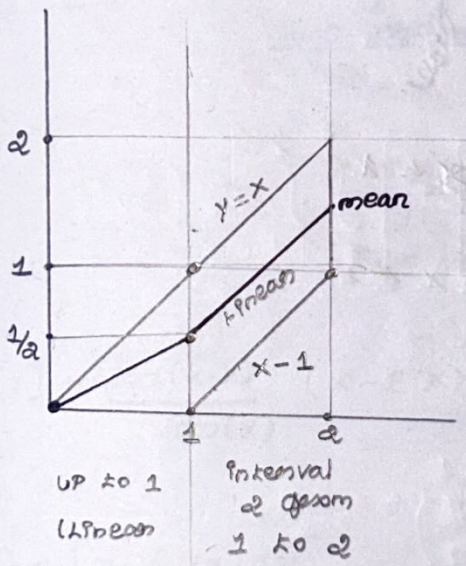
1) find $g(x)$ [LMS estimator]

$$g(x) = E[Y|X]$$

up to $X = 1$ (Linear)





2 linear parts.

* up to 1
* from to 2

) mean curve (Linear)

$$E[Y|X] = \begin{cases} \frac{1}{2}x & 0 \le x \le 1 \\ x - \frac{1}{2} & 1 \le x \le 2 \end{cases}$$

b) $E[(Y - g(x))^2 \mid X = x]$ → Conditional (mean square error in estimate)

Actual, Estimate

$$E\left[(Y - E[Y|X])^2 \mid X = x\right] = \text{Var}(Y|X)$$

$$\text{Var}(Y|X) = \frac{\text{width}}{12} = \frac{(b-a)^2}{12}$$

$$\begin{cases} \left(\frac{x^2}{12}\right), & 0 \le x \le 1 \\ \frac{1}{12}, & 1 \le x \le 2 \end{cases}$$

width $= \frac{(x-(x-1))^2}{12} = \frac{1}{12}$

c) $E\left[(Y-g(X))^2\right]$ → unconditional [Total expectation theorem]

Iterative expectation of a conditional exp is unconditional exp

$$= \int E\left[(Y-g(X))^2 \mid X=x\right] f_X(x)\,dx$$

↓ over all x [JPDF]

Fix x, iterate y → iterate x | continue

$$= \int E\left[(Y-g(X))^2 \mid X=x\right] f_X(x)\,dx$$

$$f_X(x) = \begin{cases} \int_0^x \frac{2}{3}\,dy & 0 \le x \le 1 \\ \int_{x-1}^{x} \frac{2}{3}\,dy & 1 \le x \le 2. \end{cases}$$

$$= \begin{cases} \frac{2}{3}x & 0 \le x \le 1 \\ \frac{2}{3} & 1 \le x \le 2. \end{cases}$$

$$E\left[(Y-g(X))^2\right] = \int_0^1 \frac{x^2}{12} \cdot \frac{2}{3}\,x\,dx + \int_1^2 \frac{1}{12} \cdot \frac{2}{3}\,dx$$

$$= \int_0^1 \frac{x^3}{12} \cdot \frac{2}{3}\,dx + \int_1^2 \frac{1}{12} \cdot \frac{2}{3}\,dx$$

$$= \frac{5}{72}$$

$$= E\left[\text{Var}(Y|X)\right]$$

same

$$= E\left[E\left[(Y-g(X))^2 \mid X\right]\right]$$

$$E[(Y-g(X))^2] = E[\text{Var}(Y|X)]$$

$$\boxed{E\left[E\left[(Y-g(X))^2 \mid X=x\right]\right] = E\left[(Y-g(X))^2\right]}$$

$$E[(Y-g(X))^2] = E\left[(Y-E[Y|X])^2\right]$$

$$= \text{Var}(Y|X)$$

$$= E[\text{Var}(Y|X)] = \text{Var}(Y|X)$$

$$\boxed{E[\text{Var}(Y|X)] = E\left[(Y-g(X))^2\right]}$$

↑ LMS estimator → forces estimator to have the mean m/s.

b)

$h(x) = ?$

Given case

↓

Two linear pieces ($x \sin x$)

$$h(x) = E[Y] + \frac{\text{Cov}(X,Y)}{\text{Var}(X)}(X - E(X))$$

$$E[X] = \int x \cdot f_X(x)\,dx$$

$$= \int_0^1 x \cdot \left(\frac{2}{3}x\right) dx + \int_1^2 x \cdot \left(\frac{2}{3}\right) dx$$

$$= \frac{2}{3}\left(\frac{x^3}{3}\right)_0^1 + \left(\frac{2}{3}\right)\left(\frac{x^2}{2}\right)_1^2$$

$$= \frac{2}{9}(1) + \frac{4-1}{3}$$

$$= \frac{2}{9} + 1$$

$$= \frac{11}{9}$$

$$E[X^2] = \int x^2 f_X(x)\,dx + \int$$

$$= \frac{2}{3}\left(\frac{x^4}{4}\right)_0^1 + \left(\frac{2}{3}\right)\left(\frac{x^3}{3}\right)_1^2$$

$$= 31/18$$

$$\text{Var}(x) = E[x^2] - E[x]^2 = \frac{31}{18} - \left(\frac{11}{9}\right)^2$$

$$= \frac{37}{162}$$

$$E[y] = E[E[y|x]] = \int E[y|x]\, f_x(x)$$

$$= \int_0^1 \left(\frac{1}{2}x\right) \cdot \left(\frac{2}{3}x\right) dx + \int_1^2 \left(x - \frac{1}{2}\right)\left(\frac{2}{3}\right) dx$$

$$= \frac{7}{9}$$

$$\text{Cov}(x,y) = E[xy] - E[x]E[y]$$

$$E[xy] = \int_x \int_y xy\, f_{x,y}(x,y)\, dy\, dx$$

$$= \left[\int_0^1 \int_0^x xy \left(\frac{2}{3}\right) dy\, dx\right] + \left[\int_1^2 \int_{x-1}^x xy \left(\frac{2}{3}\right) dy\, dx\right]$$

$$= \frac{41}{36}$$

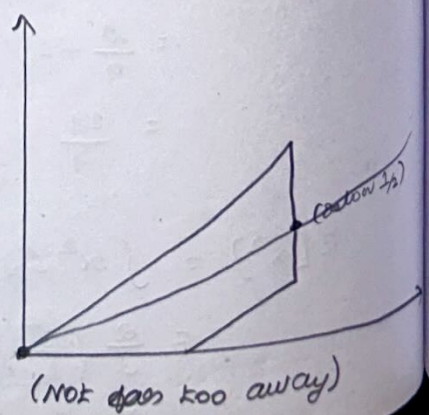
$$\text{Cov}(x,y) = \frac{41}{36} - \left(\frac{11}{9}\right)\left(\frac{41}{36}\right)$$

$$\boxed{\text{Cov}(x,y) = \frac{61}{324}}$$

$$L(x) = E[y] + \frac{\text{cov}(x,y)}{\text{Var}(x)}\,(x - E[x])$$

$$= \frac{7}{9} + \frac{\frac{61}{324}}{\left(\frac{37}{162}\right)}\left(x - \frac{11}{9}\right)$$

$$L(x) = \frac{7}{9} + \frac{61}{74}\left(x - \frac{11}{9}\right) \rightarrow \text{Plot}$$



LMS estimation: $g(x)$

Linear LMS estimation: $L(x)$

which one is better

$E[(Y-L(X))^2]$ $\geq$ $E[(Y-g(X))^2]$

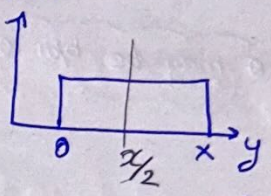
↓ we have made approximation.

↓ designed to minimize the errors.

MAP estimation – Picks the highest value of y in distribution.

Our case: Uniform.

$f_{Y|X}$



"No sensible way of choosing only one"

↓ Cond. expectation is midpoint

"Every single point gives same distribution (conditional))

---

e) $\hat{\theta}$ of conditional mean square errors ← Another problem.

$$E\left[(\hat{\theta}-\theta)^2 \mid X=x\right] = \int_{x}^{1} (\hat{\theta}-\theta)^2 f_{\theta|X}(\theta|x)\, d\theta$$

Recitation 2

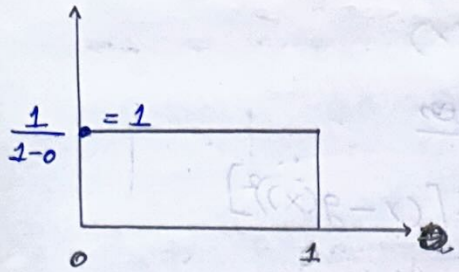
Inferring a parameter of uniform

Romeo & Juliet skeet dating, but Juliet will be late any date by a R.V X, uniformly dist $[0,\theta]$. The parameter θ → unknown → modelled as the value of a R.V θ, uniformly distributed b/w zero & 1 hour.

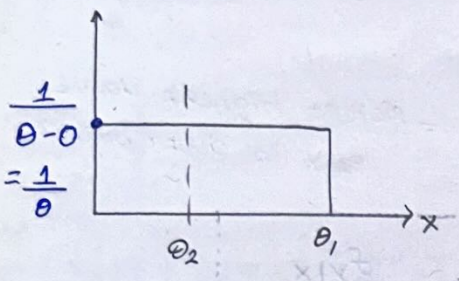
Solu:

a) Assume: Juliet was late by an amount x on their first date, how should Romeo use this info to update the distribution of θ?

Solu:



'Distribution eg θ'



[delay (X) uniformly distributed b/w 0 and θ]

↳ we don't know θ.

∴ θ may be b/w 0 and 1

Players eg uncertainty

$$f_\theta(\theta) = \begin{cases} 1, & 0 \le \theta \le 1 \\ 0, & \text{anywhere else} \end{cases} \longrightarrow \text{Prior}$$

$$f_{x|\theta}(x|\theta) = \begin{cases} \frac{1}{\theta}, & 0 \le x \le \theta \\ 0, & \text{anywhere else.} \end{cases} \longrightarrow \text{Data.}$$

$$f_{\theta|x}(\theta|x) = \frac{f_\theta(\theta)\, f_{x|\theta}(x|\theta)}{f_x(x)} = \begin{cases} \frac{1 \cdot \frac{1}{\theta}}{f_x(x)}, & 0 \le \theta \le 1, \; 0 \le x \le \theta \\ 0 & \text{elsewhere} \end{cases}$$

→ ∵ (x can take any where anywhere)

$$f_x(x) = \int_x^1 f_\theta(\theta)\, f_{x|\theta}(x|\theta)\, d\theta$$

$$= \int_x^1 \left(\frac{1}{\theta}\right) \cdot 1 \cdot d\theta = \left[\log\theta\right]_x^1 = \log 1 - \log x = -\log x = |\log x|$$

$$= \begin{cases} \frac{1}{\theta} \\ \frac{1}{|\log(x)|} \\ 0 \end{cases} \quad 0 \le \theta \le 1 \quad \text{elsewhere.}$$

$\therefore f_X(x) \to +ve$

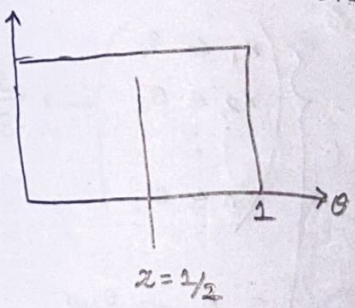
$$= \begin{cases} \frac{1}{\theta |\log x|} & 0 \le \theta \le 1 \\ 0 & \text{elsewhere.} \end{cases}$$

Started with prior $\to \theta \sim [0,1]$ uniform.

update we got

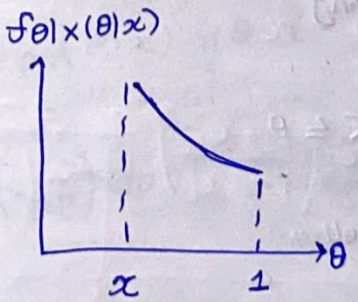
Assume 30 minutes late

when $x = 1/2 \to$ $\therefore$ $0 \le \theta \le 1$ ↓ our case



So our delay must be $x = 1/2$, so θ must start with $1/2$.

$x \le \theta \le 1$



θ less than $x \to$ no way she arrives late by x time.

"More prob of being x" → High prob curve.

b) More data

n samples, $X = (X_1 + \dots + X_n)$ (more days)

"Conditionally independent $X_1, \dots, X_n$" → Assumption.

"prior hasn't changed"

$$f_\theta(\theta) = \begin{cases} 1, & 0 \le \theta \le 1 \\ 0, & \text{anywhere else} \end{cases}$$

$$f_{X|\theta}(x|\theta) = f_{X_1,\dots,X_n|\theta}(x_1, x_2 \dots x_n|\theta) \quad \xrightarrow{\text{conditionally}} \text{Independent}$$

$$= f_{X_1|\theta}(x_1|\theta) \cdots f_{X_n|\theta}(x_n|\theta)$$

∴ Each one is uniformly dist, b/w 0 and θ

$$\boxed{f_{X|\theta}(x|\theta) = \frac{1}{\theta^n}} \rightarrow 0 \le x \le \theta$$

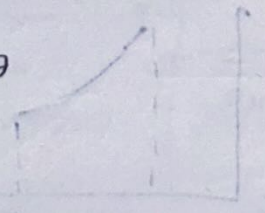
$$\boxed{x \le \theta \le 1}$$

$$\begin{cases} \left.\begin{array}{l} x_1 \le \theta_1 \\ x_2 \le \theta_2 \\ \vdots \\ x_n \le \theta_n \end{array}\right\} \rightarrow \frac{1}{\theta^n} \\ 0 \rightarrow \text{o.w.} \end{cases}$$

Let find:

$$\overline{X} = \max\{x_1, \dots x_n\}$$

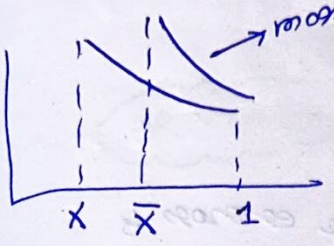
$$f_{X|\theta} = \begin{cases} \frac{1}{\theta^n}, & \overline{X} \le \theta \\ 0 & \text{other} \end{cases}$$



$$f_{\theta|X}(\theta|x) = \frac{f_\theta(\theta)\, f_{X|\theta}(x|\theta)}{f_X(x)} = \frac{1 \cdot \frac{1}{\theta^n}}{f_X(x)}, \quad \overline{X} \le \theta$$

$$f_{\bar{X}}(x) = \int_{\bar{x}}^{1} \frac{1}{\theta^n}\, d\theta$$

$$= \begin{cases} \dfrac{\frac{1}{\theta^n}}{\int_{\bar{x}}^{1} \frac{1}{\theta^n}\, d\theta} & , \bar{x} \le \theta \\ 0 & , \text{others} \end{cases}$$



one single point

Posterior distribution: MAP

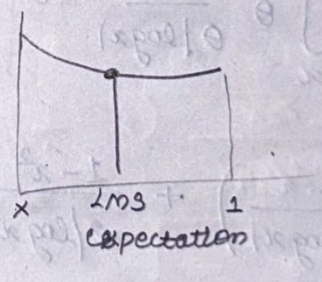
a) $\hat{\theta}_{map} = x$ (x has highest value)

↓

30 minutes late

[worst case]

a) take expectation



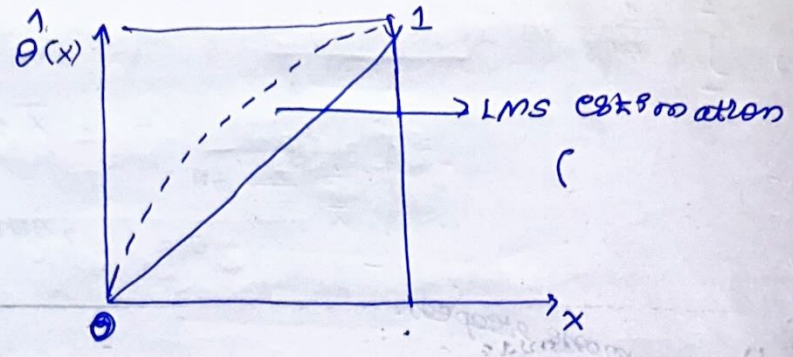
a) LMS = $E[\theta | x = x]$

$$= \int_{x}^{1} \theta\, f_{\theta|x}(\theta|x)\, d\theta$$

$$= \int_{x}^{1} \theta \cdot \frac{1}{\theta^n}\, d\theta \qquad \theta \Big/ \frac{1}{\log x}\cdot\theta \; d\theta$$

$$= \int_{x}^{1} \frac{1}{|\log x|}\, d\theta$$

$$= \frac{1-x}{|\log x|} \longrightarrow \text{LMS estimation}$$



Part-2

a) $\hat{\theta}$ → Initial conditional mean square error is

$$E\left[(\hat{\theta}-\theta)^2 \mid x=x\right] = \int_x^1 (\hat{\theta}-\theta)^2 f_{\theta|x}(\theta|x)\, d\theta$$

$$= \int_x^1 (\hat{\theta}^2 - 2\hat{\theta}\theta + \theta^2) \frac{1}{\theta|\log x|}\, d\theta$$

$$= \hat{\theta}^2 (1) - 2\hat{\theta} \int_x^1 \theta \frac{1}{\theta|\log x|}\, d\theta + \int_x^1 \theta^2 \frac{1}{\theta|\log x|}\, d\theta$$

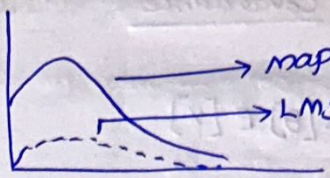
$$= \hat{\theta}^2 - 2\hat{\theta}\cdot\left(\frac{1-x}{|\log x|}\right) + \frac{1-x^2}{2|\log x|}$$

map: $\hat{\theta} = x$

$$= x^2 - 2x\left(\frac{1-x}{|\log x|}\right) + \frac{1-x^2}{2|\log x|}$$

LMS: $\hat{\theta} = \dfrac{1-x}{|\log x|}$

$$\Rightarrow \frac{1-x^2}{2|\log x|} - \left(\frac{1-x}{\log x}\right)^2$$



LMS is better.

Linear LMS

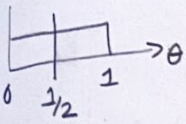
$1-x^2 \rightarrow$ Not linear

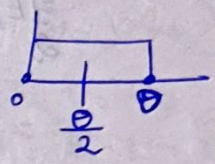
$\log x \rightarrow$ Not linear.

Linear: $ax+b$

$$\hat{\theta}_{LLMS} = E[\theta] + \frac{Cov(x,\theta)}{var(x)}(x - E[x])$$

$E[\theta] = \frac{1}{2}$





$$E[x] = E[E[X|\theta]]$$

$$= E\left[\frac{\theta}{2}\right]$$

$$= \frac{1}{2}E[\theta]$$

$$= \frac{1}{2}\left(\frac{1}{2}\right) = \frac{1}{4}.$$

$$Var(X) = E[Var(X|\theta)] + Var(E[X|\theta])$$

Variance $= \frac{(b-a)^2}{12}$

$$= E\left[\frac{\theta^2}{12}\right] + Var\left(\frac{\theta}{2}\right)$$

$$= \frac{1}{12}E[\theta^2] + \frac{1}{4}Var(\theta)$$

$$= \frac{1}{12}\left[Var(\theta) + (E[\theta])^2\right] + \frac{1}{4}Var\theta$$

$$= \frac{4}{12}Var\theta + \frac{1}{12}(E[\theta])^2$$

$$= \frac{1}{3}Var\theta + \frac{1}{12}(E[\theta])^2$$

$$= \frac{1}{3}\left(\frac{1^2}{12}\right) + \frac{1}{12}\left(\frac{1}{2}\right)^2 = \frac{1}{3}\left(\frac{1}{12}\right) + \frac{1}{12}\cdot\frac{1}{4}$$

$$= 7/144$$

Covariance

$$Cov(\theta, X) = E[\theta X] - E[\theta]\,E[X]$$

$$\boxed{E[\theta X] = E\left[E[\theta X|\theta]\right]}$$

$$Cov(\theta, X) = E\left[E[\theta X|\theta]\right] - \frac{1}{4}\cdot\frac{1}{12}$$

$$= E\left[\theta E[X|\theta]\right] - \frac{1}{48} \qquad \because \theta \text{ is known (Constant)}$$

$$= E\left[\theta \cdot \frac{\theta}{2}\right] - \frac{1}{48}$$

$$= \frac{1}{2}E[\theta^2] - \frac{1}{48} \qquad E[\theta^2] = \frac{1}{12} + \frac{1}{4} = \frac{1}{3}$$

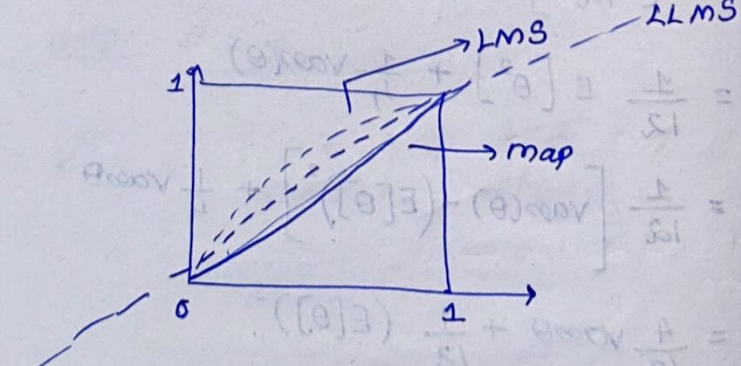
$$= \frac{1}{2}\left(\frac{1}{3}\right) - \frac{1}{48}$$

$$= 1/24$$

LLMS

$$\hat{\theta}_{LLMS} = \frac{1}{2} + \frac{1/24}{7/44}\left(x - \frac{1}{4}\right)$$

$$\boxed{\hat{\theta}_{LLMS} = \frac{6}{7}x + \frac{2}{7}}$$

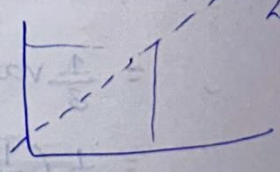


$X = 1$ (hours late)

LLMS → 8/7 (greater than 1)

θ bounded by 1.

↳ side effect of LLMS.



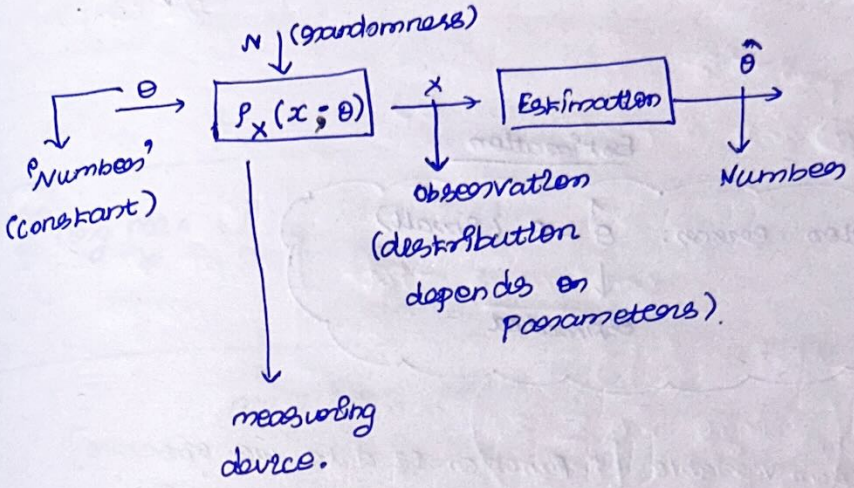
Advantage = simple

Classical Inference : 2 — Lecture : 23

↓

* Don't want to assume prior distribution.
* Confidence interval.

Classical statistics



Condition : when we have two R.V (Here only 1 R.V & a constant)

$p_X(x;\theta)$

↳ x has a certain distribution
(has a certain unknown parameter)

Estimator

Takes X & result $\hat{\theta}$ (unknown parameter).

$p_X(x;\theta)$ ⟶ 1-dimensional quantity.

Vectors :

↳ unknown parameters in background.

$p_{X_1 \ldots X_n}(x_1, \ldots x_n ; \theta_1, \ldots \theta_n)$

* Not conditional probabilities. (θ is not random).

eg :

A normal with this mean or with that mean

(Alternative candidates of a prob model).

our job: choose which one is the correct model.

eg: Coin — Small no. of models

Bias: 1/2 or 3/4

* Decide: which is the correct model? → lots of alternative models.

Hypothesis → choose one by estimation → Biased → (0.6, 0.7, ...)
→ unbiased → $(\frac{1}{2})$

Estimation

estimation error: $\hat{\theta} - \theta$ (small)
↓
estimated

$\hat{\theta}$ → Random variable [∵ Function of data we observe]
even though constant θ.

distribution of $\hat{\theta}$ → depends on x → distribution of x → Based on θ

Problem

Pick a θ, under which the data observed (x's most likely to occur).

$$\hat{\theta}_{ML} = \arg\max_{\theta} P_x(x;\theta)$$

∴ choose that θ → which causes our observed data more likely.

[Under a particular θ → x has this prob of occurring].

In Bayesian MAP estimation

maximum aposteriori distribution

$$\hat{\theta}_{MAP} = \arg\max_{\theta} P_{\theta|x}(\theta|x)$$

$$\hat{\theta}_{MAP} = \arg\max_{\theta} \frac{P_{x|\theta}(x|\theta)\, P_{\theta}(\theta)}{P_x(x)}$$

using ~~posterior~~ distribution of θ, we will find the distribution (posterior) of θ using the data of x.

↓

In our case $\hat{\theta}_{ML} = \arg\max_{\theta} P_X(x;\theta)$ ← x from a part. θ ← opposite

$\hat{\theta}_{MAP} = \arg\max_{\theta} P_{\theta|X}(\theta|x)$

$[\theta \rightarrow R.V]$

$$\hat{\theta}_{MAP} = \arg\max_{\theta} \frac{P_{x|\theta}(x|\theta)\, P_\theta(\theta)}{P_X(x)}$$

→ Bayes law.

→ Doesn't depend on θ.

↓ most likely value of θ

$\therefore \hat{\theta}_{map}$ → maximize

$$\hat{\theta}_{map} = \arg\max_{\theta} \frac{P_{x|\theta}(x|\theta)\, P_\theta(\theta)}{\text{denominator}}$$

↳ denominator don't have θ

If $P_\theta(\theta)$ (prior) is constant →

$$\hat{\theta}_{map} = \arg\max_{\theta} P_{x|\theta}(x|\theta)$$

same as

$$\hat{\theta}_{ML} = \arg\max_{\theta} P_X(x;\theta)$$

Same form when prior is constant. [Similar]

In Bayes law: we assumed a prior

Here : No priors

$x_1, x_2, \ldots, x_n$, exponential (θ)

$\theta \rightarrow$ 1 dimensional (Several data)

$$\max_{\theta} \prod_{i=1}^{n} \theta e^{-\theta x_i}$$

↓ pmf (discrete)

for a particular $\hat{\theta}_{ML} = \arg\max_{\theta} P_X(x;\theta)$

↓ Not conditional

↓ density.

(value of θ for a particular θ)

continuous — [what value of θ makes x (observe) more likely]

$$\max_{\theta} \left(n \log\theta - \theta \sum_{i=1}^{n} x_i \right)$$

maximize (By taking log then derivative) w.r.t θ

$$\hat{\theta}_{ML} = \frac{n}{x_1 + \cdots + x_n}$$

Estimators.

[Estimator]

$$\hat{\theta}_n = \frac{n}{x_1 + x_2 + \cdots + x_n}$$ → Sample mean

↓ (R.V) ← R.V (Function of R.V)

Inverse of

In exponential distribution : $\theta = \frac{1}{\text{mean}}$, [mean $= \frac{1}{\lambda}$] $\lambda = \theta$

Good or bad estimate ?

$\hat{\theta}_n \rightarrow$ close to θ [bias error]

Prop:

★ Takes: x_i 's & produces θ [random – Not exact]

★ Ideally: we don't want errors [on average: correct value]

Unbiased: $E[\hat{\theta}_n] = 0$

prob dist of $\hat{\theta}$ → affected by x_i → In turn by θ.

$E_\theta[\hat{\theta}_n] = \int f_{\hat{\theta}_n}(\hat{\theta};\theta) \cdot \hat{\theta}\, d\hat{\theta}$

$\hat{\theta}$ → function of x.

density of x is affected by θ

∴ x_i affected by θ

under a particular θ ∴ $\hat{\theta}$ → from a particular value of θ.

$E[\hat{\theta}_n] = \theta$ → Irrespective of θ (took)

No matter of θ → we don't want to be biased.

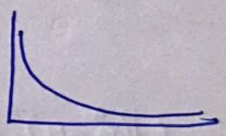
Irrespective of θ (took), $E[\hat{\theta}_n] = \theta$ [Real value]

Our estimate — Biased

Sample size : 1

$\theta = \frac{1}{x_1}$ (x_1 has a fair amount of density in 0)

$\approx \frac{1}{0} = \infty \neq \theta$



∴ (Biased upwards a lot)

"Under some conditions - we can make them asymptotically unbiased" - By having more data

"more data" - more correct data.

In poll problem

* More sample size - more accurate answer

"Desirable property of estimators"

more data: more accurate θ (unknown parameters)

↓

maximum likelihood estimation

(good approach in this case: more data).

→ $\theta_n \to 0$

$$\hat{\theta}_n = \frac{n}{x_1 + \cdots + x_2} \to \frac{1}{E[X]}$$

From weak law of convergence (large numbers)

Sample mean converges to the expectation.

$$\frac{x_1 + \cdots + x_n}{n} \Rightarrow E[X] = \frac{1}{\theta}$$

$\left[\frac{1}{\lambda}$, over case $X = \theta\right]$

$$\frac{1}{E[X]} = \theta$$

$\hat{\theta}_n = \theta$ → certainly correct one. [Irrespective of θ [over data]]

Small – Squared error

$$E\left[(\hat{\theta} - \theta)^2\right] = \text{var}(\hat{\theta} - \theta) + \left(E[\hat{\theta} - \theta]\right)^2$$

↓ Fix a little θ

$$= \text{var}(\hat{\theta}) + (\text{bias})_\theta^2$$

↳ depends on θ

$\text{var}(X) = E[X^2] - (E[X])^2$

∴ Average we fall above or below (more or less than actual value)

↓ Bias

Means square error → Bias, → $\text{var}(\hat{\theta})$) Contributions

↳ estimation

Design: Both of the bias & $\text{var}(\hat{\theta})$ small [one's own]

Assume

$X \sim N(\theta, 1)$

$\hat{\theta} = 100$

$E[(\hat{\theta} - \theta)^2] = 0 + (\text{Bias})^2 = (100 - \theta)^2$

If $\theta = 0 \rightarrow$ Large errors.

moral: ways, we can make variance small in cost of large bias (need to take care of both)

Other method: High school method (Say)

Estimate mean:

* $X_1, \ldots, X_n$ (iid) mean θ, variance σ^2
* $X_i = \theta + W_i$ (random) $\rightarrow$ measurement
* $W_i \rightarrow$ iid, mean $= 0$, variance $= \sigma^2$

$\hat{\theta}_n = M_n = \text{Sample mean} = \frac{X_1 + \ldots + X_n}{n}$

"Reasonable way"

No need of variance

[No need of distribution of X]

Properties: good?

* $E[\hat{\theta}_n] = \theta \quad = E\left[\frac{X_1 + X_2 + \ldots X_n}{n}\right] = \frac{1}{n} \, n E[X]$

$= \theta \qquad [E[X] = \theta]$

*weak large numbers law: converges to true parameter (consistency)

mean square errors

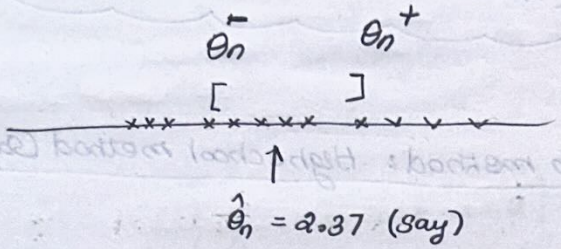
$$MSE = E\left[(\hat{\theta}_n - \theta)^2\right] \approx \frac{Var(X_i)}{n}$$

$$\approx \frac{\sigma^2}{n}$$

$$\left[Var(\hat{\theta}_n) + \left(E\left[\hat{\theta}_n - \theta\right]\right)^2 \right]$$

As $n \to \infty$, $MSE = 0$.

Sample mean: simple, easy, nice



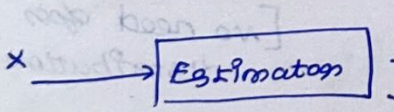
So we need to give some interval: how likely. → Report end points

Report $[\theta_n^-, \theta_n^+]$

↑ depend on data (after seeing over data)

R.V

Confidential Interval



$[\hat{\theta}^-, \hat{\theta}^+]$

Producing an interval too.

we want it to highly likely to have over output result.

$$P(\hat{\theta}_n^- \le \theta \le \hat{\theta}_n^+) \ge 1 - \alpha \quad (\alpha \to 0.95)$$

$\therefore \alpha \approx 0.05$ (large enough)

↓

Prob 0.95 this will happen!

with a 95% prob: b/w 1.96 & 2.56

number [not random]

$\therefore \theta \rightarrow$ not random.

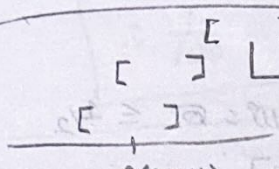
"where is the randomness"

↓

Interval → with 95% prob it has real value.

$\therefore$ Probability: 95% confidence of being the real value

---



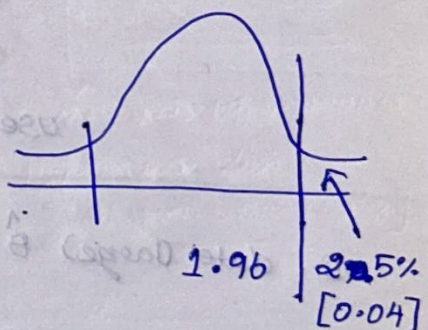
95% of days – we will be lucky.

Confidential Interval

$\hat{\theta}_n = (X_1 + \dots + X_n)/n$ [Sample mean]

$\phi(1.96) = 1 - \frac{0.05}{2}$

$$P\left(\frac{|\hat{\theta}_n - \theta|}{\sigma/\sqrt{n}} \leq 1.96\right) \approx 0.95 \text{ (CLT)}$$



$$P\left(\hat{\theta}_n - \frac{1.96\sigma}{\sqrt{n}} \leq \theta \leq \theta_n + \frac{1.96\sigma}{\sqrt{n}}\right) \approx 0.95.$$

lower end — upper end.

more generally: let z be s.t $\phi(z) = 1 - \alpha/2$

$$P\left(\hat{\theta}_n - \frac{z\sigma}{\sqrt{n}} \leq \theta \; \hat{\theta}_n + \frac{z\sigma}{\sqrt{n}}\right) \approx 1 - \sigma$$

* when n is large — more certain - small interval [very little noise]
* vice versa.

$$\boxed{\phi(z) = 1 - \alpha/2}$$

↓

Used central limit theorem (approximate) $\approx 95\%$ may not be accurate.

Difficulty:

I don't know σ.

Case: unknown σ

1) use upperbound on σ: If X_i Bernoulli: $\sigma \leq 1/2$

[most biggest confidence level]

2) Estimate from data.

* $\sigma = \sqrt{\theta(1-\theta)}$ → Bernoulli

we don't know θ [But we have estimate]

$= \sqrt{\hat{\theta}(1-\hat{\theta})}$

· use S.D in the construction of confidential interval.

As data (large) $\hat{\theta} \approx \theta$ (good estimate)

$\sigma = \sqrt{\hat{\theta}(1-\hat{\theta})}$ also good.

3) Don't have any formulas like Case (2)

Generic method: estimate variance → mean of $[(X_i - \theta)^2]$

↓

obtain lots of measurement then take average

$\sigma_n^2 = \frac{1}{n} \sum_{i=1}^{n} (X_i - \theta)^2 \longrightarrow \sigma^2$

(By weak law of large numbers)

we don't know θ

$[\theta \rightarrow \text{mean}]$

$$S_n^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \hat{\theta}_n) \longrightarrow \sigma^2$$

(Unbiased $E[S_n^2] = \sigma^2$)

As loorge data.

$n-1$ [$\because$ For being unbiased $\rightarrow$ factor]

$\therefore \frac{1}{100} \approx \frac{1}{99}$ $\rightarrow$ Not a big deal.

---

overall

1) Estimate sample mean
2) S.D variance
3) use $\rightarrow$ to find confidential interval

major $\rightarrow$ use $\hat{\theta}$ to find S.D.

Pretend
Sample mean
Normal distribution
(central limit theorem)

classical inference – II

Linear regression: → Today

* Likelihood fn x based on θ
* Similar to Bayesian without priors.

Picking more likely θ

$\max_\theta f_X(x;\theta)$ → choose more likely θ.

pick θ (even though syntactically but appears) where x is (observed) more likely.

Sample mean estimate:

$$\hat{\theta_n} = \frac{x_1 + \cdots + x_n}{n}$$ → Common mean [Take average]

$1-\alpha$ confidential interval: $P(\hat{\theta_n}^- \le \theta \le \hat{\theta_n}^+) \ge 1-\alpha$, $\forall\theta$ → Particular

For Sample mean – CI

z be s.t $\phi(z) = 1 - \alpha/2$

$$P\left(\hat{\theta}_n - \frac{z\sigma}{\sqrt{n}} \le \theta \le \hat{\theta_n} + \frac{z\sigma}{\sqrt{n}}\right) \approx 1-\sigma$$

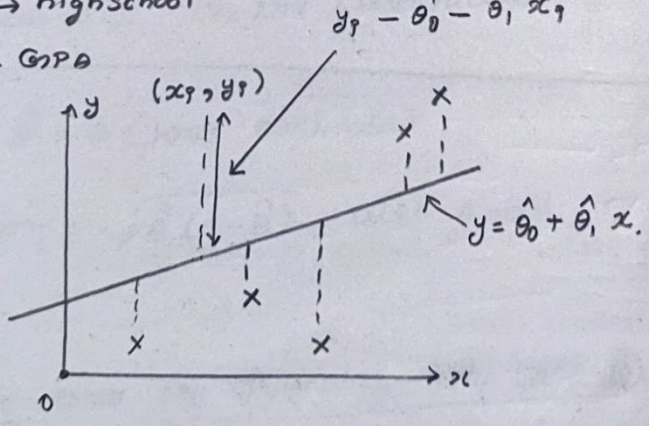
The interval has prob of 95% of having the real value in it.

∴ θ is not random [constant]

Pretend: Sample mean ≈ Normal distribution (good when n is large)

$x \to$ SAT → highschool

$Y \to$ MIT – GPA



Data: $(x_1, y_1), \ldots (x_n, y_n)$

Model: $y \approx \theta_0 + \theta_1 x$ → Linear relationship

$$\min_{\theta_0, \theta_1} \sum_{i=1}^{n} (y_i - \theta_0 - \theta_1 x_i)^2$$ → Squared errors

"we can't expect the line to pass through all points"

↓

we expect them to be close.
(randomness)

For a particula θ_0, θ_1: calculate error

choose θ_0, θ_1 with minimum error.

↓

Square error

one interpretation:

* $Y_i = \theta_0 + \theta_1 x_i + w_i$, $w_i \sim N(0, \sigma^2)$, IID

↓ random noise.

GPA — SAT + randomness.

Random noise: Independent $\sim N(0, \sigma^2)$ p.d.f

likelihood:

likelihood: $f_{X,Y|\theta}(x, y : \theta)$: ?

For a particular w: $ce^{-w^2/2\sigma^2}$

for $Y_i = \theta_0 + \theta_1 x_i + w_i$

likelihood is

$$c \cdot \exp\left\{ -\frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - \theta_0 - \theta_1 x_i)^2 \right\}$$

* Take logs, take derivatives

* Least Square ⟷ pretend w_i p.d.f normal.

maximize this

$\max \theta_0, \theta_1$

After log:

max : (since − ve sign → smallest as possible)

Linear regression

Conclusion: choose to do linear regression on maximum likelihood

where $y_i = \theta_0 + \theta_1 x_i + w_i$ $(w_i \sim N(0, \sigma^2))$

Assumes this

carry out

$$\min_{\theta_1, \theta_0} \sum_{i=1}^{n} (y_i - \theta_0 - \theta_1 x_i)^2$$

↓

then only C. exp $\left\{ -\frac{1}{2\sigma^2} \right\}$ will be maximum

derivative → 0

$\frac{d}{d\theta}$ (quadratic function) → linear system

set derivatives to zero:

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n} \quad \bar{y} = \frac{y_1 + \cdots + y_n}{n}$$

$$\hat{\theta}_1 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

→ Solving quadratic

$$\hat{\theta}_0 = \bar{y} - \hat{\theta}_1 \bar{x}$$

Interpretation

x, y → random variables

w − independent of x

$E[y] = \theta_0 + \theta_1 E[x] + 0$ [w has zero mean]

$$\theta_0 = E[y] - \theta_1 E[x]$$

$$\boxed{\theta_0 = \bar{y} - \theta_1 \bar{x}}$$

"Estimate $\hat{\theta}_1$"

$$\boxed{\hat{\theta}_0 = \bar{y} - \hat{\theta}_1 \bar{x}}$$

Estimate $\hat{\theta}_1$

Assume for simplicity $E[x] = E[w] = 0$

$$y = \theta_0 + \theta_1 x + w$$

mult by x

$$xy = \theta_0 x + \theta_1 x^2 + wx$$

$$cov(xy) = \theta_0 E[x] + \theta_1 E[x^2] + cov(wx)$$

$$cov(xy) = 0 + \theta_1 E[x^2] + 0$$

$$cov(xy) = \theta_1 E[(x-0)^2]$$

$$cov(xy) = \theta_1 var(x)$$

$$\boxed{\theta_1 = \frac{cov(x,y)}{var(x)}}$$

when mean is not zero $E[x] \neq 0$

$$\theta_1 = \frac{cov(x,y)}{var(x)} = \frac{E[(x - E[x])(y - E[y])]}{E[(x - E[x])^2]}$$

Linear regression

model $\longrightarrow$ $y \approx \theta_0 + \theta_1 x$

$$\min_{\theta_0, \theta_1} \sum_{i=1}^{n} (y_i - \theta_0 - \theta_1 x_i)^2$$

"GPA may affected by several factors"

Multiple linear regression

data: (x_i, x_i', x_i'', y_i), $i = 1, \dots, n$

model: $y \approx \theta_0 + \theta x + \theta' x' + \theta'' x''$

formulation:

$$\min_{\theta, \theta', \theta''} \sum_{i=1}^{n} (y_i - \theta_0 - \theta x_i - \theta' x_i' - \theta'' x_i'')^2$$

↓

Take derivation (Set to zero)

↓

system of linear equ.

choosing right variables: → (quadratic may fit) or something,

model $y \approx \theta_0 + \theta_1 h(x)$

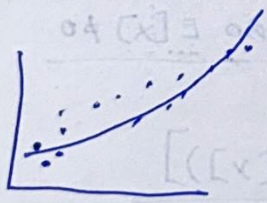
eg: $y \approx \theta_0 + \theta_1 x^2$

"work with data points $(y_i, h(x))$

formulation:

$$\min_{\theta} \sum_{i=1}^{n} (y_i - \theta_0 - \theta_1 h_1(x_i))^2$$

"some: GPA has something to do with SAT
(or) courses eg SAT"



Linear model: poor fit.

"A quadratic model may be good"

Choose: x^2 as planetary variable.

How to deal with $h(x)$

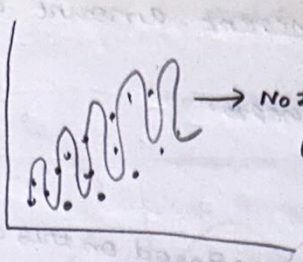
"multiple linear regression"

↓

may be 10th degree polynomial

(prediction error: extremely small)

linear — Any others (quadratic?)



* Avoid: too many parameters when little data is available'

Confidential Intervals

* Estimate σ [how much randomness?]
* R^2 → measure of explanatory power.

$$y = \theta_0 + \theta_1 x + w$$

↳ Estimate σ^2 [Std. errors]

* How good we can predict

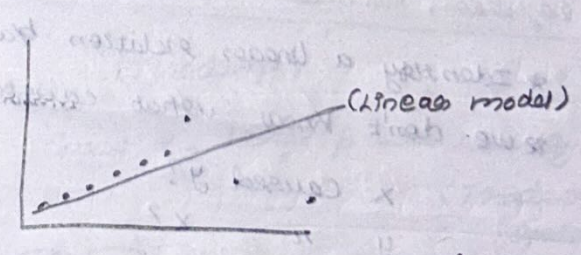
$\frac{Var(Y|X)}{Var(Y)}$ → Comparing: randomness of Y knowing X & randomness of Y.

---

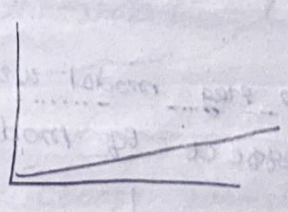
* How much randomness of Y is attached with X.

Concerns

1) has zero Skedacisity



'w has small variance'

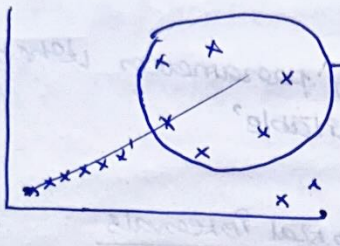


'w has large variances'

* 'W is not identically distributed — variance of w depends on x'

"with different x spaces - different amount of noise"

Aim: minimize sum of squared errors.



→ Based on this data (noisy one) → (we need to build our model)

↓

Not the right thing. (deal in many ways)

multiple explanatory variable

$$Y_{GPA} = \theta_0 + \theta_1 SAT_{1st\ attempt} + \theta_2 SAT_{2nd\ attempt}$$

very likely to be close.

"Ignore one score - go with others"

"sensitive: data - θ_1 big, θ_2 small"

↓

Our answer drastically changes

"Careful" - make special test.

Run linear regression, Run linear model:

* Identify a linear relation b/w x & y
* we don't know what caused which?

x caused y?

y " x?

x, y caused by any?

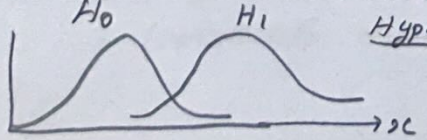
" Building a linear reg model with small errors doesn't tell causal effect of model"

↓

Tells

close association b/w the variables.

Hypothesis testing



which distribution is correct?

H_0 or H_1