Chapter 16
Classical Statistical Inference

This chapter and the next …

This chapter and the next are about creating many more estimators.

Keep in mind: estimation is an art, not a science.

- ❖ There is no mathematical proof for the “best” estimator.
- ❖ We pick our favorite estimator subjectively.
- ❖ Lots of room for you to create your own!

Estimating the number of pink jelly beans

Given: Jar has 1000 jelly beans.

Goal: Estimate θ =Number of pink jelly beans (unknown)

Experiment: Randomly sample $n = 20$ jelly beans
with replacement

Why is this a random variable?

$X =$ # pink jelly beans in sample

Why is this a random variable?

$\hat{\theta}(X) =$ estimator of θ

Estimating the number of pink jelly beans

Goal: Estimate θ =Number of pink jelly beans

Experiment: Randomly sample $n = 20$ beans w/ replacement

X = # pink jelly beans in sample

$\hat{\theta}(X)$ = estimator of θ

1000 jelly beans total

Q: What's a reasonable $\hat{\theta}(X)$?

Q: First suppose $X = x$

$$\hat{\theta}(X = x) = \left(\frac{x}{n}\right) \cdot 1000, \quad \forall\, 0 \le x \le n$$

$$\therefore\ \hat{\theta}(X) = \left(\frac{X}{n}\right) \cdot 1000$$

Is our estimator unbiased?

Goal: Estimate θ =Number of pink jelly beans

Experiment: Randomly sample $n = 20$ beans w/ replacement

1000 jelly beans total

X = # pink jelly beans in sample

$$\hat{\theta}(X) = \left(\frac{X}{n}\right) \cdot 1000$$

Q: Is $\hat{\theta}(X)$ unbiased?

<u>Hint</u>: How is X distributed?

$$\mathrm{X} \sim Binomial\left(n, p = \frac{\theta}{1000}\right)$$

$$\rightarrow \mathbf{E}[\mathrm{X}] = n \cdot \frac{\theta}{1000}$$

$$\rightarrow \boxed{\mathbf{E}\left[\hat{\theta}(X)\right]} = \mathbf{E}\left[\frac{X}{n} \cdot 1000\right] = \mathbf{E}[X] \cdot \frac{1000}{n}$$

$$= n \cdot \frac{\theta}{1000} \cdot \frac{1000}{n} = \boxed{\theta}$$

Is our estimator consistent?

$\hat{\theta}(X) = \left(\frac{X}{n}\right) \cdot 1000,$ where X $\sim Binomial\left(n, p = \frac{\theta}{1000}\right)$

Need to show
$\boldsymbol{MSE}\left(\hat{\theta}(X)\right) \to 0$
as $n \to \infty$

Suffices to show
$\boldsymbol{Var}\left(\hat{\theta}(X)\right) \to 0$

Why??

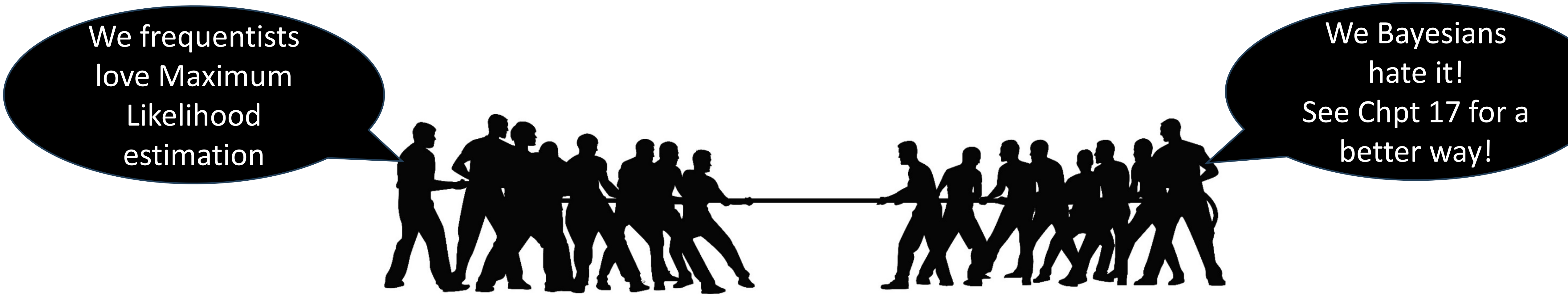
$$\boxed{\mathbf{Var}\left(\hat{\theta}(X)\right)} = \frac{1000^2}{n^2} \cdot np(1-p)$$

$$= \frac{1000^2}{n^2} \cdot n \cdot \frac{\theta}{1000} \cdot \left(1 - \frac{\theta}{1000}\right)$$

$$= \frac{\theta(1000 - \theta)}{n} \xrightarrow[n\to\infty]{} \boxed{0} \checkmark$$

So far we've had no "method" for coming up with estimators.

We now introduce a methodology, called Maximum Likelihood Estimation (MLE), for deriving estimators.



Creating an ML estimator

Goal: Estimate an unknown value θ, given sample data X

This is the "likelihood function"

<u>Step 1</u>: Define

$$\hat{\theta}_{ML}(X = x) = \underset{\theta}{\arg\max}\ \boldsymbol{P}\{X = x \mid \theta\}$$

Our estimator is the value of θ that maximizes the likelihood function

<u>Step 2</u>: Convert $\hat{\theta}_{ML}(X = x)$ to $\hat{\theta}_{ML}(X)$

Example of MLE

Goal: Estimate θ =Number of pink jelly beans

Experiment: Randomly sample $n = 20$ beans w/ replacement

X = # pink jelly beans in sample

Q: Suppose we observe $X = 3$ pinks in our sample.
What is $\boldsymbol{P}\{X = 3 \mid \theta\}$?

$$\boldsymbol{P}\{\,X = 3 \mid \theta\,\} = \binom{20}{3}\left(\frac{\theta}{1000}\right)^{3} \cdot \left(1 - \frac{\theta}{1000}\right)^{17}$$

What value of θ maximizes this?

Example of MLE

Goal: Estimate θ =Number of pink jelly beans

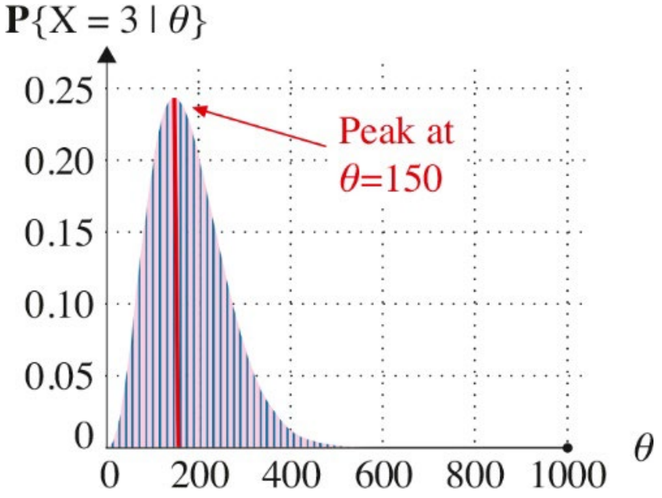
Experiment: Randomly sample $n = 20$ beans w/ replacement

$X = $ # pink jelly beans in sample

Q: Suppose we observe $X = 3$ pinks in our sample. What is $\boldsymbol{P}\{X = 3 \mid \theta\}$?

What value of θ maximizes this?

$\theta = 150$



$$\hat{\theta}_{ML}(X = 3)$$

$$= \underset{\theta}{\arg\max}\ \boldsymbol{P}\{X = x \mid \theta\}$$

$$= 150$$

Example of MLE

Q: What is the likelihood function $\boldsymbol{P}\{X = x \mid \theta\}$?

$$\boldsymbol{P}\{ X = x \mid \theta \} = \binom{n}{x} \left(\frac{\theta}{1000}\right)^{x} \cdot \left(1 - \frac{\theta}{1000}\right)^{n-x}$$

1000 jelly beans total

X = # pink jelly beans in sample

Q: What is $\hat{\theta}_{ML}(X = x) = \underset{\theta}{\arg\max}\ \ \boldsymbol{P}\{X = x \mid \theta\}$?

$$0 = \frac{d}{d\theta} \boldsymbol{P}\{ X = x \mid \theta \} = \frac{d}{d\theta} \binom{n}{x} \left(\frac{\theta}{1000}\right)^{x} \cdot \left(1 - \frac{\theta}{1000}\right)^{n-x}$$

$$= \binom{n}{x} \left(\frac{\theta}{1000}\right)^{x} \cdot (n-x) \cdot \left(1 - \frac{\theta}{1000}\right)^{n-x-1} \cdot \frac{-1}{1000} + \binom{n}{x} \cdot x \left(\frac{\theta}{1000}\right)^{x-1} \cdot \frac{1}{1000} \cdot \left(1 - \frac{\theta}{1000}\right)^{n-x}$$

$$= -\frac{n-x}{1000} \cdot \frac{\theta}{1000} + \frac{x}{1000} \cdot \left(1 - \frac{\theta}{1000}\right)$$

Solving, we get:
$$\theta = \frac{1000x}{n}$$

2^{nd} derivative is negative, so this is a max ✓

Example of MLE

Q: What is the likelihood function $\boldsymbol{P}\{X = x \mid \theta\}$?

$$\boldsymbol{P}\{X = x \mid \theta\} = \binom{n}{x}\left(\frac{\theta}{1000}\right)^{x} \cdot \left(1 - \frac{\theta}{1000}\right)^{n-x}$$

1000 jelly beans total

$X =$ # pink jelly beans in sample

Q: What is $\hat{\theta}_{ML}(X = x) = \underset{\theta}{\arg\max}\ \boldsymbol{P}\{X = x \mid \theta\}$?

$$\hat{\theta}_{ML}(X = x) = \underset{\theta}{\arg\max}\ \boldsymbol{P}\{X = x \mid \theta\} = \frac{1000x}{n}$$

This holds $\forall x$

Hey, this is what we got before!

➔ $\boxed{\hat{\theta}_{ML}(X) = \frac{1000X}{n}}$

PSC Example

#jobs submitted daily to Pittsburgh Supercomputing Center $\sim Poisson(\lambda)$

Goal: Estimate λ

Experiment: Sample #job submissions each day for a month: $X_1, X_2, \ldots, X_{30}$
(assume independent)

What do you think $\hat{\lambda}(X_1, X_2, \ldots, X_{30})$ will be?

$$\hat{\lambda}(X_1, X_2, \ldots, X_{30}) = \frac{X_1 + X_2 + \cdots + X_{30}}{30}$$

PSC Example: MLE method

#jobs submitted daily to Pittsburgh Supercomputing Center $\sim Poisson(\lambda)$

$$\hat{\lambda}_{ML}(X_1 = x_1, X_2 = x_2, \ldots, X_{30} = x_{30})$$

$$= \underset{\lambda}{\arg\max} \; \boldsymbol{P}\{X_1 = x_1, X_2 = x_2, \ldots, X_{30} = x_{30} \mid \lambda\}$$

$$= \underset{\lambda}{\arg\max} \frac{\lambda^{x_1} e^{-\lambda}}{x_1!} \cdot \frac{\lambda^{x_2} e^{-\lambda}}{x_2!} \cdots \frac{\lambda^{x_{30}} e^{-\lambda}}{x_{30}!}$$

$$= \underset{\lambda}{\arg\max} \frac{\lambda^{x_1 + x_2 + \cdots + x_{30}} e^{-30\lambda}}{x_1! \, x_2! \cdots x_{30}!}$$

What should we do
to find the
maximizing λ?

PSC Example: MLE method

$$\hat{\lambda}_{ML}(X_1 = x_1, X_2 = x_2, \ldots, X_{30} = x_{30}) = \arg\max_{\lambda} \underbrace{\frac{\lambda^{x_1+x_2+\cdots+x_{30}} e^{-30\lambda}}{x_1!\,x_2!\cdots x_{30}!}}_{\text{set derivative equal to 0}}$$

$$0 = \frac{d}{d\lambda}\left(\frac{\lambda^{x_1+x_2+\cdots+x_{30}} e^{-30\lambda}}{x_1!\,x_2!\cdots x_{30}!}\right)$$

$$= \frac{(x_1 + x_2 + \cdots + x_{30})\lambda^{x_1+x_2+\cdots+x_{30}-1} e^{-30\lambda} + \lambda^{x_1+x_2+\cdots+x_{30}} \cdot (-30) \cdot e^{-30\lambda}}{x_1!\,x_2!\cdots x_{30}!}$$

$\rightarrow$ $$0 = (x_1 + x_2 + \cdots + x_{30}) + \lambda \cdot (-30)$$

$\rightarrow$ $$\lambda = \frac{x_1 + x_2 + \cdots + x_{30}}{30}$$

PSC Example: MLE method

We have shown:

This holds $\forall x_1, x_2, \ldots, x_{30}$

$$\hat{\lambda}_{ML}(X_1 = x_1, X_2 = x_2, \ldots, X_{30} = x_{30}) = \frac{x_1 + x_2 + \cdots + x_{30}}{30}$$

$$\Rightarrow \hat{\lambda}_{ML}(X_1, X_2, \ldots, X_{30}) = \boxed{\frac{X_1 + X_2 + \cdots + X_{30}}{30}}$$

unsurprising

Log likelihood estimation

Goal: Estimate an unknown value θ, given sample data X

Step 1: Define

This is the likelihood function

This is the log likelihood function

$$\hat{\theta}_{ML}(X = x) = \arg\max_{\theta} \underbrace{\boldsymbol{P}\{X = x \mid \theta\}}_{\text{likelihood function}} = \arg\max_{\theta} \underbrace{\log \boldsymbol{P}\{X = x \mid \theta\}}_{\text{log likelihood function}}$$

equivalent
Why??

Step 2: Convert $\hat{\theta}_{ML}(X = x)$ to $\hat{\theta}_{ML}(X)$

PSC example revisited with log likelihood

#jobs submitted daily to Pittsburgh Supercomputing Center $\sim Poisson(\lambda)$

the unknown

$$\hat{\lambda}_{ML}(X_1 = x_1, X_2 = x_2, \ldots, X_{30} = x_{30})$$

$$= \arg\max_{\lambda} \ \ln(\boldsymbol{P}\{X_1 = x_1, X_2 = x_2, \ldots, X_{30} = x_{30} | \ \lambda\})$$

$$= \arg\max_{\lambda} \ \ln(\boldsymbol{P}\{X_1 = x_1 \mid \lambda\} \cdot \boldsymbol{P}\{X_2 = x_2 \mid \lambda\} \cdots \boldsymbol{P}\{X_{30} = x_{30} \mid \lambda\})$$

$$= \arg\max_{\lambda} \sum_{i=1}^{30} \ln(\boldsymbol{P}\{X_i = x_i \mid \lambda\})$$

$$= \arg\max_{\lambda} \sum_{i=1}^{30} \ln\left(\frac{e^{-\lambda}\lambda^{x_i}}{x_i!}\right)$$

PSC example revisited with log likelihood

$$\hat{\lambda}_{ML}(X_1 = x_1, X_2 = x_2, \ldots, X_{30} = x_{30})$$

$$= \arg\max_{\lambda} \sum_{i=1}^{30} \ln\left(\frac{e^{-\lambda}\lambda^{x_i}}{x_i!}\right)$$

$$= \arg\max_{\lambda} \left(-30\lambda + \sum_{i=1}^{30} x_i \ln(\lambda) - \sum_{i=1}^{30} \ln(x_i!)\right)$$

Why can we drop this?

$$= \arg\max_{\lambda} \left(-30\lambda + \sum_{i=1}^{30} x_i \ln(\lambda)\right)$$

PSC example revisited with log likelihood

$$\hat{\lambda}_{ML}(X_1 = x_1, X_2 = x_2, \ldots, X_{30} = x_{30}) = \arg\max_{\lambda} \underbrace{\left(-30\lambda + \sum_{i=1}^{30} x_i \ln(\lambda)\right)}_{\text{set derivative equal to 0}}$$

What should we do to find the maximizing λ?

$$0 = \frac{d}{d\lambda}\left(-30\lambda + \sum_{i=1}^{30} x_i \ln(\lambda)\right) = -30 + \left(\sum_{i=1}^{30} x_i\right) \cdot \frac{1}{\lambda}$$

➔ $\boxed{\lambda = \frac{x_1 + x_2 + \cdots + x_{30}}{30}}$

➔ $\boxed{\hat{\lambda}_{ML}(X_1 = x_1 X_2 = x_2, \ldots, X_{30} = x_{30}) = \frac{x_1 + x_2 + \cdots + x_{30}}{30}}$

MLE with data from continuous distribution

MLE single variable:
We wish to estimate an unknown value θ.

- If the sample data is represented by discrete r.v. X, then we define:

$$\hat{\theta}_{ML}(X = x) = \underset{\theta}{\arg\max} \ \boldsymbol{P}\{X = x \mid \theta\}$$

- If the sample data is represented by continuous r.v. X, then we define:

$$\hat{\theta}_{ML}(X = x) = \underset{\theta}{\arg\max} \ f_{X|\theta}(x)$$

MLE with data from continuous distribution

MLE multiple variables:
We wish to estimate an unknown value θ.

- If the sample data is represented by discrete r.v.s $X_1, X_2, \ldots, X_n$, then we define:

$$\hat{\theta}_{ML}(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n) = \underset{\theta}{\arg\max} \; \boldsymbol{P}\{X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n \mid \theta\}$$

- If the sample data is represented by continuous r.v.s $X_1, X_2, \ldots, X_n$, then we define:

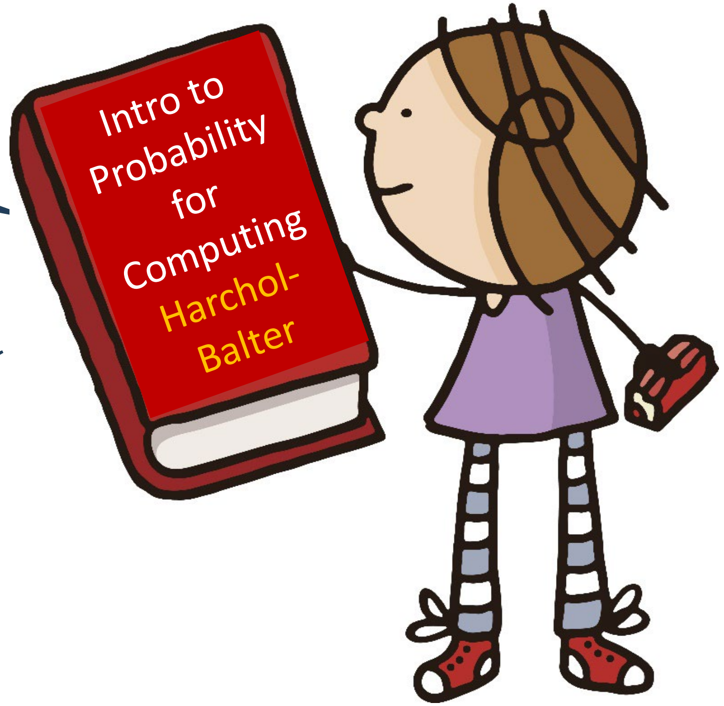
$$\hat{\theta}_{ML}(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n) = \underset{\theta}{\arg\max} \; f_{X_1, X_2, \ldots, X_n \mid \theta}(x_1, x_2, \ldots, x_n)$$

MLE Example: time spent on homework

How long can I expect to work on homework?

The distribution is $Uniform(0, b)$

Great, but what is b?



Sample some students and estimate it!

MLE Example: time spent on homework

Given: Time students spend on homework each week $\sim Uniform(0, b)$

Goal: Estimate the unknown b

Experiment: Sample 3 students independently at random: X_1, X_2, X_3

Estimator: $\hat{b}(X_1, X_2, X_3)$

Start by deriving
$\hat{b}(X_1 = x_1, X_2 = x_2, X_3 = x_3)$

MLE Example: time spent on homework

Given: Time students spend on homework each week $\sim Uniform(0, b)$

$$\hat{b}_{ML}(X_1 = x_1, X_2 = x_2, X_3 = x_3) = \underset{b}{\arg\max}\ f_{X_1,X_2,X_3|b}(x_1, x_2, x_3)$$

$$f_{X_1,X_2,X_3|b}(x_1, x_2, x_3) = \begin{cases} \frac{1}{b^3} & \text{if } 0 < x_1, x_2, x_3 \le b \\ 0 & \text{o.w.} \end{cases}$$

Is (x_1, x_2, x_3) possible here?

What value of b maximizes this?

$$= \begin{cases} \boxed{\frac{1}{b^3}} & \text{if } b \ge \max(x_1, x_2, x_3) \\ 0 & \text{o.w.} \end{cases}$$

$b = \max(x_1, x_2, x_3)$

MLE Example: time spent on homework

Given: Time students spend on homework each week $\sim Uniform(0, b)$

$$\hat{b}_{ML}(X_1 = x_1, X_2 = x_2, X_3 = x_3) = \max(x_1, x_2, x_3)$$

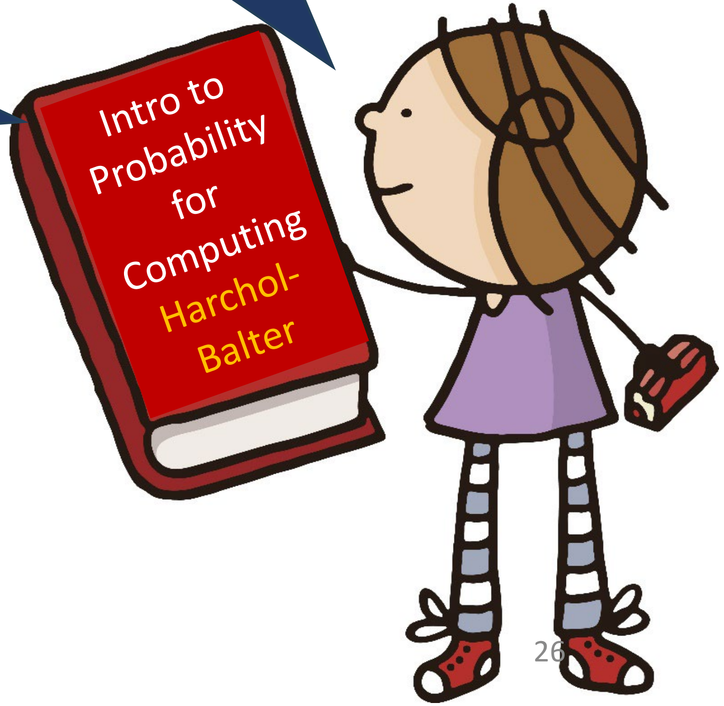
$$\Rightarrow \hat{b}_{ML}(X_1, X_2, X_3) = \max(X_1, X_2, X_3)$$

Huh??
I asked 3 random students:
$X_1 = 10$, $X_2 = 5$, $X_3 = 15$
Should $b = 15$?

What if I only have 1 sample?!

The MLE is just one possible estimator, and not always the best one.

In Exercise 16.5, you'll show that $\hat{b}_{ML}$ is a biased estimator, but can be made into an unbiased estimator pretty easily.



MLE Example: estimating std of temperature

Given: The high temp in Pittsburgh in June $\sim Normal(\mu = 79, \sigma^2)$

Goal: Estimate the unknown σ via MLE

Experiment: Observe temperature on n randomly sampled independent June days: $X_1, X_2, \dots, X_n$

$$\hat{\sigma}_{ML}(X_1 = x_1, X_2 = x_2, \dots, X_n = x_n) = \underset{\sigma}{\arg\max}\ f_{X_1,X_2,\dots,X_n|\sigma}(x_1, x_2, \dots, x_n)$$

$$= \underset{\sigma}{\arg\max}\ \ln\left(f_{X_1,X_2,\dots,X_n|\sigma}(x_1, x_2, \dots, x_n)\right)$$

$$= \underset{\sigma}{\arg\max}\ \ln\left(\prod_{i=1}^{n} f_{X_i|\sigma}(x_i)\right)$$

MLE Example: estimating std of temperature

Given: The high temp June $\sim Normal(\mu = 79, \sigma^2)$, where σ is unknown

$$\hat{\sigma}_{ML}(X_1 = x_1, X_2 = x_2, \dots, X_n = x_n) = \underset{\sigma}{\arg\max}\ \ln\left(\prod_{i=1}^{n} f_{X_i|\sigma}(x_i)\right)$$

$$= \underset{\sigma}{\arg\max} \sum_{i=1}^{n} \ln\left(f_{X_i|\sigma}(x_i)\right)$$

$$= \underset{\sigma}{\arg\max} \sum_{i=1}^{n} \ln\left(\frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}}\right)$$

$$= \underset{\sigma}{\arg\max} \sum_{i=1}^{n} \left(-\frac{(x_i-\mu)^2}{2\sigma^2} - \ln\sigma - \ln\sqrt{2\pi}\right)$$

MLE Example: estimating std of temperature

Given: The high temp June ~ $Normal(\mu = 79, \sigma^2)$, where σ is unknown

$$\hat{\sigma}_{ML}(X_1 = x_1, X_2 = x_2, \dots, X_n = x_n) = \arg\max_{\sigma} \sum_{i=1}^{n} \left(-\frac{(x_i - \mu)^2}{2\sigma^2} - \ln \sigma - \ln \sqrt{2\pi} \right)$$

$$= \arg\max_{\sigma} \underbrace{-\frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \mu)^2 - n \ln \sigma - n \ln \sqrt{2\pi}}_{\boxed{\text{Set the derivative to 0}}}$$

$$0 = \frac{d}{d\sigma}\left(-\frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \mu)^2 - n \ln \sigma - n \ln \sqrt{2\pi} \right) = \frac{1}{\sigma^3} \sum_{i=1}^{n} (x_i - \mu)^2 - \frac{n}{\sigma}$$

MLE Example: estimating std of temperature

Given: The high temp June ~ $Normal(\mu = 79, \sigma^2)$, where σ is unknown

$$0 = \frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2 - n$$

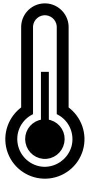
(We multiplied both sides by n.)

$$\rightarrow \sigma = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \mu)^2}{n}}$$

$$\rightarrow \hat{\sigma}_{ML}(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n) = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \mu)^2}{n}}$$

$$\rightarrow \boxed{\hat{\sigma}_{ML}(X_1, X_2, \ldots, X_n) = \sqrt{\frac{\sum_{i=1}^{n}(X_i - \mu)^2}{n}}}$$

MLE estimation of > 1 parameter



Given: The high temp in Pittsburgh in June $\sim Normal(\mu, \sigma^2)$

Goal: Estimate **both** unknowns μ and σ via MLE

Experiment: Observe temperature on n randomly sampled independent June days: $X_1, X_2, \ldots, X_n$

$$\begin{pmatrix} \hat{\mu}_{ML}\,(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n) \\ \hat{\sigma}_{ML}(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n) \end{pmatrix} = \underset{\mu,\sigma}{\arg\max}\ \ln\Big(\underbrace{f_{X_1,X_2,\ldots,X_n|\mu,\sigma}(x_1, x_2, \ldots, x_n)}_{\boxed{\text{Likelihood function } g(\mu,\sigma)}} \Big)$$

$$\boxed{\frac{\partial \ln g(\mu,\sigma)}{\partial \mu} = 0} \quad \text{AND} \quad \boxed{\frac{\partial \ln g(\mu,\sigma)}{\partial \sigma} = 0}$$

MLE estimation of > 1 parameter

$$\begin{pmatrix} \hat{\mu}_{ML}\,(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n) \\ \hat{\sigma}_{ML}\,(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n) \end{pmatrix} = \underset{\mu,\sigma}{\operatorname{arg\,max}} \ln\Big(\underbrace{f_{X_1,X_2,\ldots,X_n|\mu,\sigma}(x_1, x_2, \ldots, x_n)}_{\boxed{\text{Likelihood function } g(\mu,\sigma)}} \Big)$$

$$\ln(g(\mu,\sigma)) = -\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2 - n\ln\sigma - n\ln\sqrt{2\pi}$$

$$0 = \frac{\partial \ln g(\mu,\sigma)}{\partial \mu} = \frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i - \mu)$$

AND

$$0 = \frac{\partial \ln g(\mu,\sigma)}{\partial \sigma} = \frac{1}{\sigma^3}\sum_{i=1}^{n}(x_i - \mu)^2 - \frac{n}{\sigma}$$

$$\Rightarrow \boxed{\mu = \frac{x_1 + x_2 + \cdots + x_n}{n}}$$

$$\Rightarrow \boxed{\sigma = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2}}$$

MLE estimation of > 1 parameter

$$\begin{pmatrix} \hat{\mu}_{ML}\,(X_1 = x_1, X_2 = x_2, \dots, X_n = x_n) \\ \hat{\sigma}_{ML}(X_1 = x_1, X_2 = x_2, \dots, X_n = x_n) \end{pmatrix} = \underset{\mu,\sigma}{\arg\max}\ \ln\left(f_{X_1,X_2,\dots,X_n|\mu,\sigma}(x_1, x_2, \dots, x_n) \right)$$

$$\mu = \frac{x_1 + x_2 + \cdots + x_n}{n}$$

$$\sigma = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2}$$

need to
substitute in μ

$$\sigma = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(x_i - \frac{x_1 + x_2 + \cdots + x_n}{n}\right)^2}$$

MLE estimation of > 1 parameter

$$\hat{\mu}_{ML}\,(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n) = \frac{x_1 + x_2 + \cdots + x_n}{n}$$

$$\hat{\sigma}_{ML}(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n) = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(x_i - \frac{x_1 + x_2 + \cdots + x_n}{n}\right)^2}$$

$$\rightarrow \boxed{\hat{\mu}_{ML}\,(X_1, X_2, \ldots, X_n\,) = \frac{X_1 + X_2 + \cdots + X_n}{n}}$$

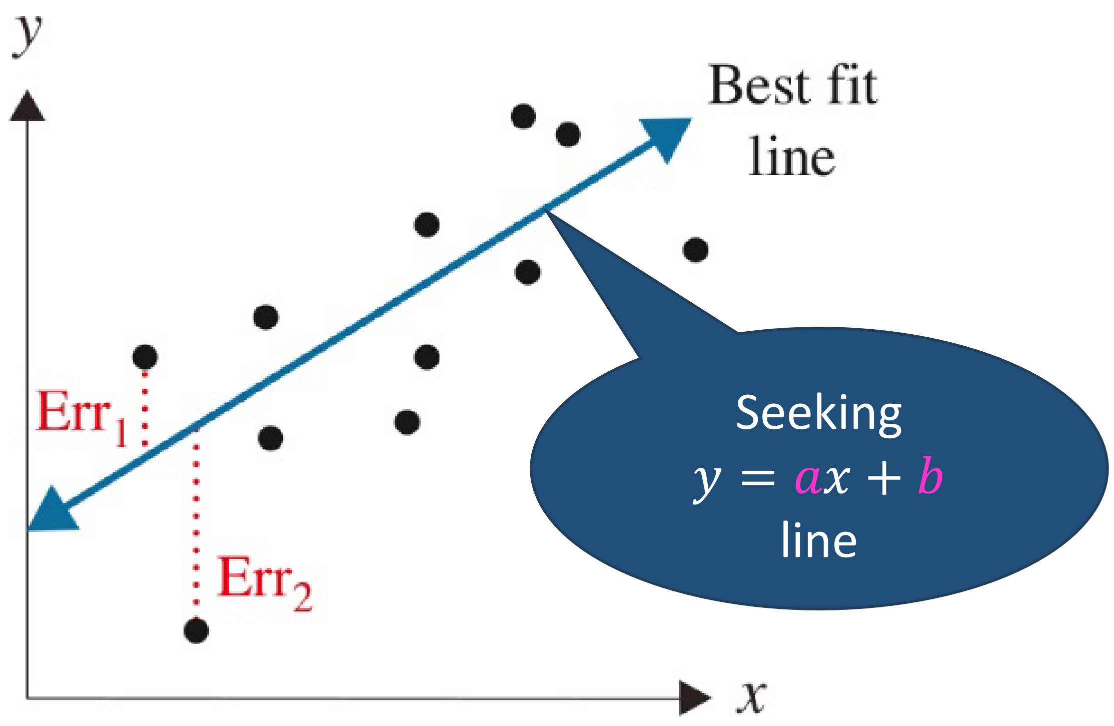
$$\rightarrow \boxed{\hat{\sigma}_{ML}(X_1, X_2, \ldots, X_n\,) = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(X_i - \frac{X_1 + X_2 + \cdots + X_n}{n}\right)^2}}$$

Linear Regression: a new kind of optimization

Experiment: Obtain n data points generated through some experiment, represented by random variables: $(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)$

Goal: Given specific instance, $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$, find the line that best this.

This will **NOT** be Maximum Likelihood Estimation!



Linear Regression

<u>Defn</u>: **(Linear Regression)** Let $(X_1, Y_1), (X_2, Y_2), \dots, (X_n, Y_n)$ be a set of data sample points.

Define $\hat{a}$ and $\hat{b}$ to be estimators for line parameters.

Define:

$$\widehat{Y_i} \equiv \hat{a}X_i + \hat{b}$$

$\widehat{Y_i}$ is an estimated dependent r.v.

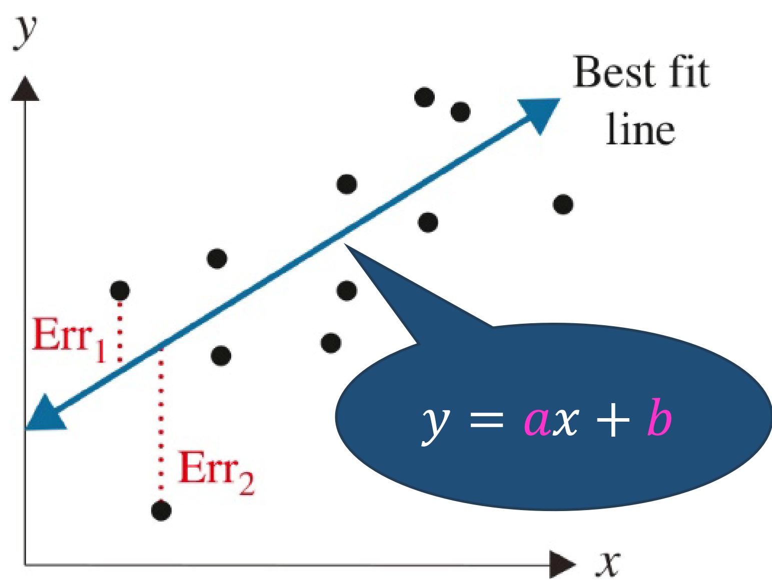
The **point-wise error** is defined as:

$$\mathbf{Err}_{\boldsymbol{i}} = Y_i - \widehat{Y_i}$$

The **sample average squared error (SASE)** is then:

$$\mathbf{SASE}(\widehat{Y_1}, \widehat{Y_2}, \dots, \widehat{Y_n}) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{Err}_{\boldsymbol{i}}{}^2 = \frac{1}{n}\sum_{i=1}^{n} \left(Y_i - \widehat{Y_i}\right)^2$$

The goal of **linear regression** is to find estimates $\hat{a}$ and $\hat{b}$ that minimize $\mathbf{SASE}(\widehat{Y_1}, \widehat{Y_2}, \dots, \widehat{Y_n})$.



Linear Regression

Goal: Derive estimators

$$\hat{a}((X_1, Y_1), \ldots, (X_n, Y_n)) \quad \text{and} \quad \hat{b}((X_1, Y_1), \ldots, (X_n, Y_n))$$

which minimize $\mathbf{SASE}\left(\widehat{Y_1}, \widehat{Y_2}, \ldots, \widehat{Y_n}\right)$, where $\widehat{Y_i} \equiv \hat{a}X_i + \hat{b}$.

Why can't we define $\hat{a}$ and $\hat{b}$ to be ML estimators, where we find a and b which maximize

$$\boldsymbol{P}\{(X_1 = x_1, Y_1 = y_1), \ldots, (X_n = x_n, Y_n = y_n) | \, a, b\} \text{ ?}$$

There's no probability in this problem. Once you specify the $X_i's$ and a and b, then the $Y_i's$ are also fully specified.

We're not doing MLE, just trying to minimize the SASE. But the methodology will feel like MLE.

Linear Regression

Goal: Derive estimators

$$\hat{a}((X_1, Y_1), \dots, (X_n, Y_n)) \quad \text{and} \quad \hat{b}((X_1, Y_1), \dots, (X_n, Y_n))$$

which minimize $\mathbf{SASE}\left(\widehat{Y_1}, \widehat{Y_2}, \dots, \widehat{Y_n}\right)$, where $\widehat{Y_i} \equiv \hat{a}X_i + \hat{b}$.

Given: set of specific points: $(x_1, y_1), \dots, (x_n, y_n)$
Given: specific choice of a and b
Define:

$$g(a, b) = \mathbf{SASE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - (ax_i + b))^2$$

$$\begin{pmatrix} \hat{a}((X_1 = x_1, Y_1 = y_1), \dots, (X_n = x_n, Y_n = y_n)) \\ \hat{b}((X_1 = x_1, Y_1 = y_1), \dots, (X_n = x_n, Y_n = y_n)) \end{pmatrix} = \underset{a,b}{\arg\min}\ g(a, b)$$

Linear Regression

$$\begin{pmatrix} \hat{a}((X_1 = x_1, Y_1 = y_1), \ldots, (X_n = x_n, Y_n = y_n)) \\ \hat{b}((X_1 = x_1, Y_1 = y_1), \ldots, (X_n = x_n, Y_n = y_n)) \end{pmatrix} = \underset{a,b}{\arg\min}\ g(a,b)$$

$$= \underset{a,b}{\arg\min} \frac{1}{n} \sum_{i=1}^{n} (y_i - (ax_i + b))^2$$

$$= \underset{a,b}{\arg\min} \underbrace{\sum_{i=1}^{n} (y_i - (ax_i + b))^2}$$

Set **both** partial derivatives to 0

$$\boxed{\frac{\partial g(a,b)}{\partial a} = 0}$$

AND

$$\boxed{\frac{\partial g(a,b)}{\partial b} = 0}$$

Linear Regression

$$g(a,b) = \sum_{i=1}^{n} (y_i - (ax_i + b))^2$$

$$\boxed{\frac{\partial g(a,b)}{\partial b} = 0}$$

$$0 = \sum_{i=1}^{n} \frac{\partial}{\partial b} (y_i - (ax_i + b))^2$$

$$0 = -2 \sum_{i=1}^{n} (y_i - (ax_i + b))$$

$$0 = \sum_{i=1}^{n} (y_i - (ax_i + b)) = \sum_{i=1}^{n} y_i - a \sum_{i=1}^{n} x_i - nb$$

$$b = \frac{\sum_{i=1}^{n} y_i}{n} - a \frac{\sum_{i=1}^{n} x_i}{n}$$

Linear Regression

$$\boxed{\frac{\partial g(a,b)}{\partial b} = 0}$$

$$g(a,b) = \sum_{i=1}^{n} (y_i - (ax_i + b))^2$$

$$b = \frac{\sum_{i=1}^{n} y_i}{n} - a\frac{\sum_{i=1}^{n} x_i}{n}$$

$$\boxed{b = \bar{y} - a\bar{x}}$$

where:

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n} \qquad \bar{y} = \frac{y_1 + y_2 + \cdots + y_n}{n}$$

Linear Regression

$$\boxed{\frac{\partial g(a,b)}{\partial a} = 0}$$

$$g(a,b) = \sum_{i=1}^{n} (y_i - (ax_i + b))^2$$

$$0 = \sum_{i=1}^{n} \frac{\partial}{\partial a} (y_i - (ax_i + b))^2$$

$$0 = -2 \sum_{i=1}^{n} \big(y_i - (ax_i + b)\big) \cdot x_i$$

$$0 = \sum_{i=1}^{n} \big(y_i - (ax_i + b)\big) \cdot x_i$$

$$= \sum_{i=1}^{n} y_i x_i - b \sum_{i=1}^{n} x_i - a \sum_{i=1}^{n} x_i^2$$

To solve for a ,
let's first
substitute in:
$b = \bar{y} - a\bar{x}$

Linear Regression

$$0 = \sum_{i=1}^{n} y_i x_i - b \sum_{i=1}^{n} x_i - a \sum_{i=1}^{n} x_i^2$$

$$\boxed{b = \bar{y} - a\bar{x}}$$

$$0 = \sum_{i=1}^{n} y_i x_i - (\bar{y} - a\bar{x}) \sum_{i=1}^{n} x_i - a \sum_{i=1}^{n} x_i^2$$

$$0 = \sum_{i=1}^{n} x_i (y_i - \bar{y}) + \sum_{i=1}^{n} x_i a\bar{x} - a \sum_{i=1}^{n} x_i^2$$

$$\sum_{i=1}^{n} x_i (y_i - \bar{y}) = a\left(\sum_{i=1}^{n} x_i^2 - \sum_{i=1}^{n} x_i \bar{x}\right) \Rightarrow \boxed{a = \frac{\sum_{i=1}^{n} x_i (y_i - \bar{y})}{\sum_{i=1}^{n} x_i (x_i - \bar{x})}}$$

Linear Regression

$$a = \frac{\sum_{i=1}^{n} x_i\,(y_i - \bar{y})}{\sum_{i=1}^{n} x_i(x_i - \bar{x})}$$

$$b = \bar{y} - a\bar{x}$$

$$= \frac{\sum_{i=1}^{n} x_i\,(y_i - \bar{y}) - \sum_{i=1}^{n} \bar{x}\,(y_i - \bar{y})}{\sum_{i=1}^{n} x_i(x_i - \bar{x}) - \sum_{i=1}^{n} \bar{x}\,(x_i - \bar{x})}$$

because

$$\sum_{i=1}^{n}(y_i - \bar{y}) = 0 = \sum_{i=1}^{n}(x_i - \bar{x})$$

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

substitute in $\hat{a}$

$$\hat{b}\big((X_1 = x_1, Y_1 = y_1), \ldots, (X_n = x_n, Y_n = y_n)\big) = \bar{y} - \hat{a}\bar{x}$$

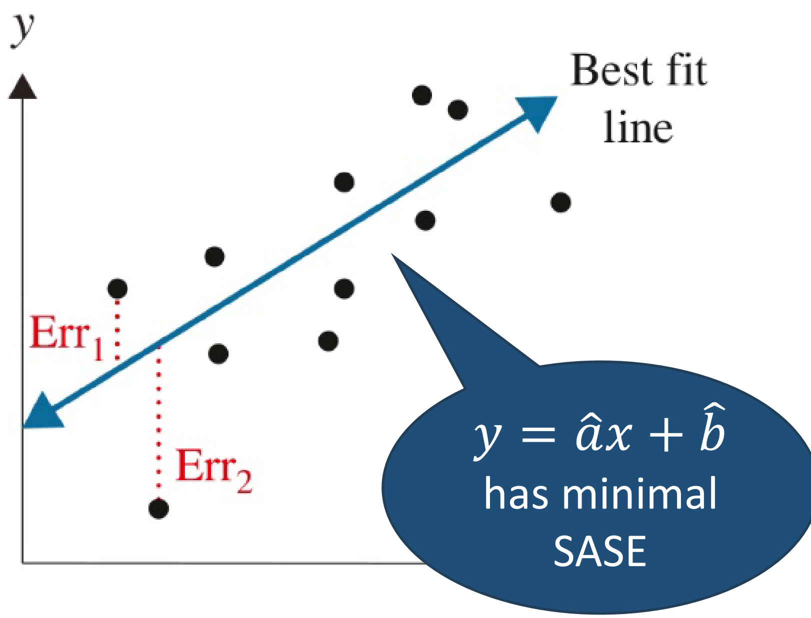
$$\hat{a}\big((X_1 = x_1, Y_1 = y_1), \ldots, (X_n = x_n, Y_n = y_n)\big) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

These hold $\forall (x_i, y_i)$

Linear Regression: Interpretation

$$\hat{b}\big((X_1,Y_1),\dots,(X_n,Y_n)\big) = \bar{Y} - \hat{a}\bar{X}$$

$$\hat{a}\big((X_1,Y_1),\dots,(X_n,Y_n)\big) = \frac{\sum_{i=1}^{n}(X_i-\bar{X})(Y_i-\bar{Y})}{\sum_{i=1}^{n}(X_i-\bar{X})^2}$$



- Makes perfect sense: $\bar{Y} = \hat{a}\bar{X} + \hat{b}$
- Also has interpretation:

$$\hat{a}\big((X_1,Y_1),\dots,(X_n,Y_n)\big) = \frac{\frac{1}{n-1}\sum_{i=1}^{n}(X_i-\bar{X})(Y_i-\bar{Y})}{\frac{1}{n-1}\sum_{i=1}^{n}(X_i-\bar{X})^2} = \frac{\mathbf{Cov}(X,Y)}{\mathbf{Var}(X)}$$

Positive slope if X and Y are positively correlated

Unbiased sample variance of $X_i's$

R^2 Goodness of Fit

To evaluate our linear regression, we use the R^2 measure.

<u>Defn</u>: **(R^2 goodness of fit)** Given a set of data sample points $(X_1 = x_1, Y_1 = y_1), \ldots, (X_n = x_n, Y_n = y_n)$ with estimated linear fit:

$$y = \hat{a}x + \hat{b}$$

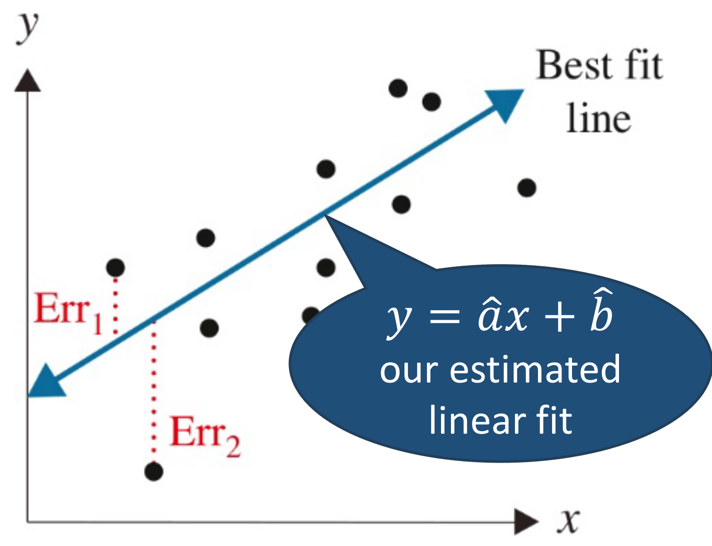
Define

$$\hat{y_i} = \hat{a}x_i + \hat{b}$$

to be the **estimated dependent value** for the i^{th} point.

Then we define the goodness of fit of the line by

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y_i})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

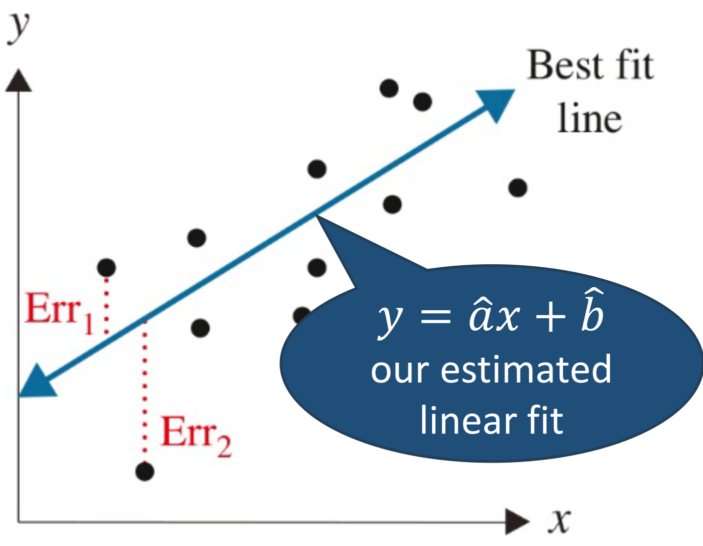


$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n}$$

$$\bar{y} = \frac{y_1 + y_2 + \cdots + y_n}{n}$$

R^2 Goodness of Fit

To evaluate our linear regression, we use the R^2 measure.



<u>Defn</u>: **(R^2 goodness of fit)** Given a set of data sample points $(X_1 = x_1, Y_1 = y_1), \ldots, (X_n = x_n, Y_n = y_n)$ with estimated linear fit:

$$y = \hat{a}x + \hat{b}$$

Define

$$\hat{y_i} = \hat{a}x_i + \hat{b}$$

to be the **estimated dependent value** for the i^{th} point.

$$R^2 = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y_i})^2}{\sum_{i=1}^n (y_i - \bar{y})^2}$$

$$= 1 - \frac{\frac{1}{n}\sum_{i=1}^n (y_i - \hat{y_i})^2}{\frac{1}{n}\sum_{i=1}^n (y_i - \bar{y})^2} = 1 - \frac{\text{sample avg. sqrd error}}{\text{sample variance}}$$

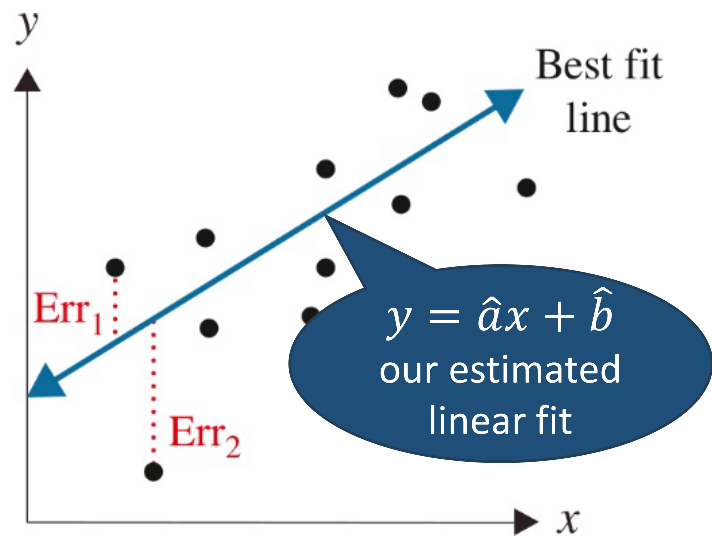
Goal is that this should be near 1.

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n}$$

$$\bar{y} = \frac{y_1 + y_2 + \cdots + y_n}{n}$$

How do we know $0 \leq R^2 \leq 1$?

$$R^2 = 1 - \frac{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2}$$



Suffices to show that:

This is SASE where the estimator is $\hat{y}_i$

This is SASE where the estimator is $\bar{y}$

$$\frac{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2} < 1$$

$$\hat{y}_i = \hat{a}x_i + \hat{b}$$

$$\bar{y} = \frac{y_1 + y_2 + \cdots + y_n}{n}$$

But, by definition, $\hat{y}_i$ is the estimator that minimizes SASE! So the numerator is smaller than the denominator!