Warmup 2:00-2:10

Stat 88 lec 35

German tanks were numbered 1,2,3,...,N, with N unknown, during World War 2 and the Allies needed to estimate N. They captured 5 tanks numbered 20, 31,43, 78 and 92. Can you find an unbiased estimate of N?

Using $\bar{X}$

$E(\bar{X}) = \frac{N+1}{2} \Rightarrow 2E(\bar{X}) - 1 = N$

$T_1 = 2\bar{X} - 1$ is unbiased

since $E(T_1) = 2E(\bar{X}) - 1 = N$ ✓

$T_1 = 2(52.8) - 1 = \boxed{104.6}$

Using $M = \max(X_1, \ldots, X_n)$

gap → ..... 20 ..... 31 ..... 43 ..... 78 ..... 92 ..... N (M)

expected length of a gap $= \frac{N-n}{n+1}$, $n = 5$ sample s.t.e

$E(M) = \frac{N-n}{n+1} \cdot n + n = n\left[\frac{N-n}{n+1} + 1\right]$

$= n\left[\frac{N-n+n+1}{n+1}\right] = \frac{n}{n+1}(N+1)$

$\Rightarrow \frac{n+1}{n} E(M) - 1 = N$

$T_2 = \frac{n+1}{n} M - 1$ is unbiased estimator

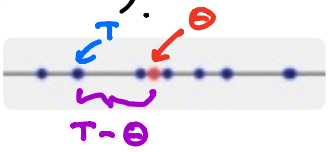
$T_2 = \frac{6}{5}(92) - 1 = \boxed{109.4}$

We need to calculate $\text{Var}(T_1)$ and $\text{Var}(T_2)$ to see which is the better estimator.

Last time

Sec 9.1, 9.2 Bias and variance

We score how good an estimator T at a parameter θ is by $MSE_\theta(T) = E((T-\theta)^2)$.



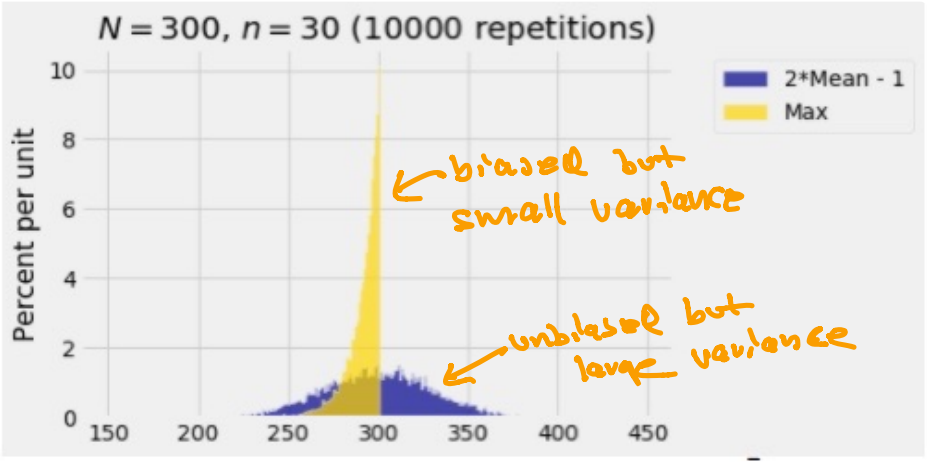
We showed,

$$MSE_\theta(T) = \left(B_\theta(T)\right)^2 + Var_\theta(T)$$

where $B_\theta(T) = E(T) - \theta$ (bias)

$Var_\theta(T) = E\left((T-E(T))^2\right)$ (variance).

The best estimator isn't always unbiased.



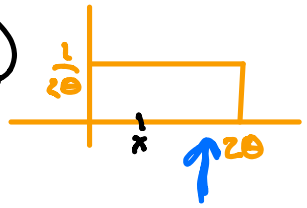
To find an unbiased estimator start with a statistic whose expectation is a linear function of the parameter.

Today ① Sec 11.2 Practice finding an unbiased estimator

② Sec 11.3 least squares regression.

① Sec 11.2 Practice finding an unbiased estimator

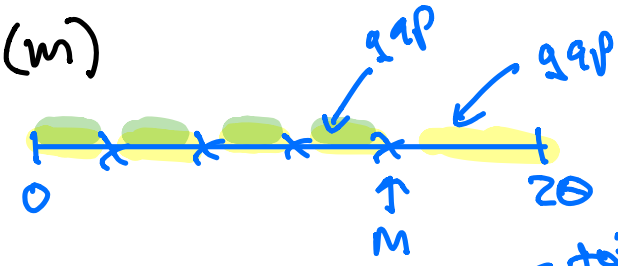
Let $X_1, X_2, \ldots, X_n \overset{iid}{\sim} Unif(0, 2\theta)$



let $M = \max(X_1, \ldots, X_n)$

Is M a biased estimator of 2θ?

Find $E(M)$



average gap length $= \frac{2\theta}{n+1}$ ← total gap length; ← # gaps

$$E(M) = \frac{2\theta}{n+1} \cdot n$$

discrete case.

$E(M) = \frac{N-n}{n+1} \cdot n + n$

$= (n/n+1)(N+1)$

Appendix: See appendix for another way to calculate $E(M)$ using the CDF of M.

Find an unbiased estimator for 2θ.

Solve for 2θ:

$\frac{n+1}{n} E(M) = 2\theta$

$E(M) = \frac{2\theta}{n+1} \cdot n$

$\Rightarrow \boxed{T = \frac{n+1}{n} M}$ is unbiased

② Sec 11.3 least squares regression

Let (x, y) be a random pair of father, son heights from the population

x = father height
y = son height

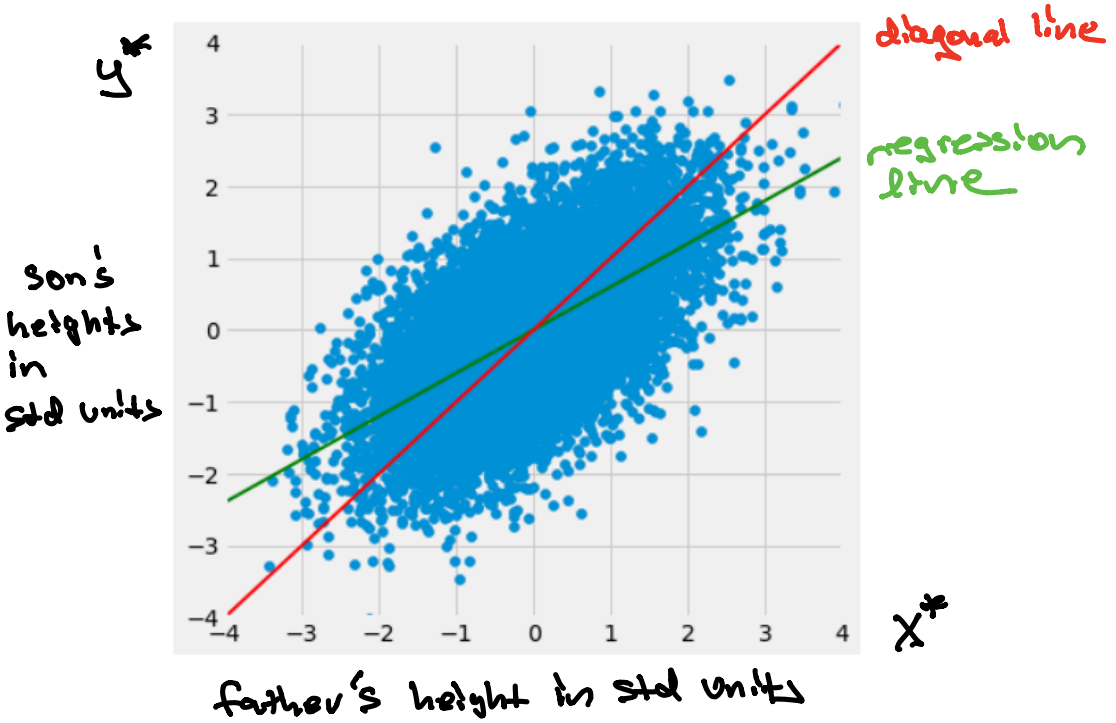
we want to estimate y, call this $\hat{y}$, by the function $\hat{y} = ax + b$ for some a, b,

you plug in x into $\hat{y} = ax + b$ to predict y.

To find a and b,
in data 8 you collected n pairs
$(x_1, y_1), \ldots, (x_n, y_n)$ and made a scatter plot.
The regression line is the "best" fitting line $\hat{y} = ax + b$ through your scatter plot.

$$a = \text{slope of the regression line} = r\frac{\text{SD of } y}{\text{SD of } x}$$

$$b = \text{intercept of the regression line}$$
$$= (\text{average of } y) - \text{slope} \times (\text{average of } x)$$

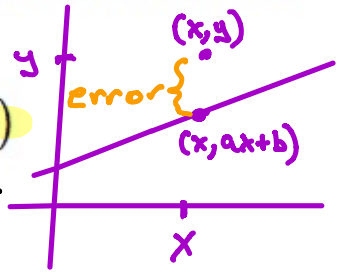


Mean Squared Error

For the random point (X, Y), the mean squared error of a linear predictor of Y based on X depends on the slope a and intercept b of the line used. So let us define $MSE(a, b)$ to be the mean squared error when we use the line $aX + b$ to predict Y. That is,

$$MSE(a, b) = E\big((Y - (aX + b))^2\big)$$

error



we average over all random (x,y) pairs in the population.

Notation

- $E(X) = \mu_X$, $SD(X) = \sigma_X$
- $E(Y) = \mu_Y$, $SD(Y) = \sigma_Y$

we need to take the partial derivative one variable at a time,

Step 1 Best intercept (b) for a fixed slope (a).

Fix slope, a, and solve $\frac{d}{db} mse_a(b) = 0$ for b.

$$mse_a(b) = E\left((y - ax - b)^2\right)$$

$$= E\left(\left((y - ax) - b\right)^2\right) = 0$$

(random variables; constant)

↖ FOIL

$$= E\left((y-ax)^2 - 2(y-ax)b + b^2\right)$$

$E(b^2) = b^2$ since b not random.

$$= E\left((y-ax)^2\right) - 2\left(E(y) - aE(x)\right)b + b^2$$

Solve $\frac{d}{db} mse_a(b) = 0$ for b

$$-2\left(E(y) - aE(x)\right) + 2b = 0$$

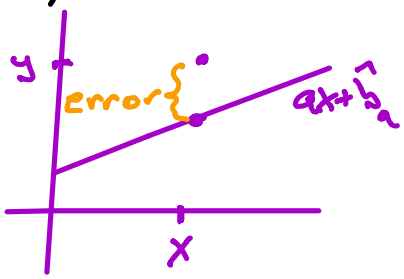
Solve for b

$$\hat{b}_a = E(y) - aE(x)$$

we will write $\boxed{b_a = \mu_y - a\mu_x}$

Step 2 Find the best slope (a),

$$\text{error} = Y - (aX + \underbrace{\hat{b}_a}_{=\mu_Y - a\mu_X})$$



$$= Y - aX - \mu_Y + a\mu_X$$

$$= \underbrace{Y - \mu_Y}_{D_Y} - a\underbrace{(X - \mu_X)}_{D_X}$$

remember $E(D_Y^2) = \sigma_Y^2$

$$MSE(a) = E\left((D_Y - aD_X)^2\right)$$
$$= E(D_Y^2) - 2aE(D_X D_Y) + a^2 E(D_X^2)$$
$$= \sigma_Y^2 - 2aE(D_X D_Y) + a^2\sigma_X^2$$

$$\frac{d}{da} MSE(a) = 0$$

$$\Rightarrow \quad -2E(D_X D_Y) + 2a\sigma_X^2 = 0$$

$$\Rightarrow \quad \hat{a} = \frac{E(D_X D_Y)}{\sigma_X^2}$$

$$= \frac{E\left((X - \mu_X)(Y - \mu_Y)\right)}{\sigma_X^2}$$

we will show equal to $r\frac{\sigma_Y}{\sigma_X}$

So regression line is

$\hat{y} = \hat{a}x + \hat{b}$ where $\hat{a} = \frac{E(D_x D_y)}{\sigma_x^2}$

$\hat{b} = \mu_y - \hat{a}\mu_x$

Correlation

What is $E(D_x^2)$? — Var(X) $D_x = x - \mu_x$

$E(D_x D_y)$ is the covariance of X and Y.

If X = father's ht. (ft)
Y = son's ht. (ft)

$E(D_x D_y)$ has units ft^2

dividing it by $\sigma_x \sigma_y$,

$r = \frac{E(D_x D_y)}{\sigma_x \sigma_y}$ is unitless and called the correlation coefficient of X, Y.

Covariance $E(D_x D_y) = r\sigma_x \sigma_y$

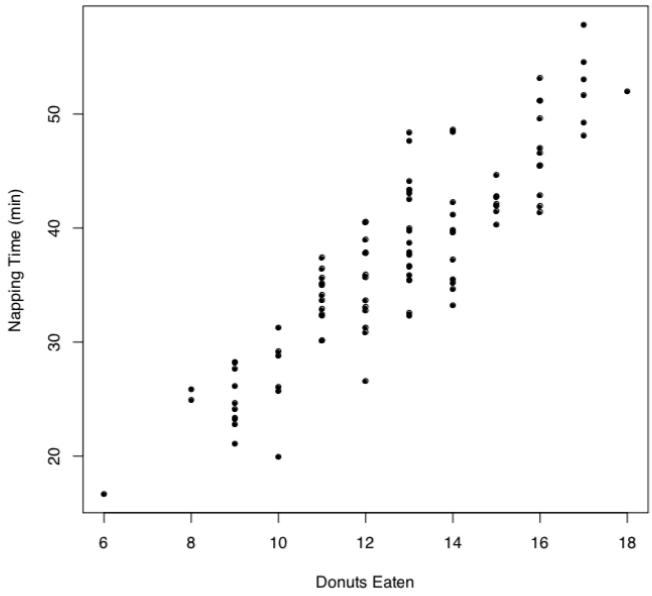
so $\hat{a} = \frac{E(D_x D_y)}{\sigma_x^2} = \frac{r\sigma_x \sigma_y}{\sigma_x^2} = \boxed{\frac{r\sigma_y}{\sigma_x}}$

ex

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

$\hat{a} = r\frac{\sigma_y}{\sigma_x}$

$\hat{b} = \mu_y - \hat{a}\mu_x$



Suppose

$\bar{x} = 11.5$ $SD(x) = 2.5$

$\bar{y} = 30$ $SD(y) = 8$

$r = .5$

Suppose I tell you that Homer ate 14 donuts today. What is your best prediction for the time he spent napping?

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

$\hat{a} = .5\left(\frac{8}{2.5}\right) = 1.6$

$\hat{b} = 30 - 1.6(11.5) = 11.6$

so $\hat{y} = 1.6(14) + 11.6 = 34$ min

Appendix

Let $X_1, X_2, \ldots, X_n \overset{iid}{\sim} \text{Unif}(0, 2\theta)$

let $M = \max(X_1, \ldots, X_n)$

Calculate the density of M by first calculating the cdf of M.

$$F(m) = P(M \le m) = P(X_1 \le m, \ldots, X_n \le m)$$
$$= P(X_1 \le m)^n$$
$$= \left(m\left(\frac{1}{2\theta}\right)\right)^n$$
$$f(m) = \frac{d}{dm} F(m) = n\left(m\left(\frac{1}{2\theta}\right)\right)^{n-1} \cdot \frac{1}{2\theta}$$
$$= \boxed{n m^{n-1} \cdot \frac{1}{(2\theta)^n}}$$

Calculate $E(M)$

$$E(M) = \int_0^{2\theta} m f(m)\, dm = \frac{n}{(2\theta)^n} \int_0^{2\theta} m^n\, dm$$
$$= \frac{n}{(2\theta)^n} \frac{m^{n+1}}{n+1} \Big|_0^{2\theta}$$
$$= \boxed{(2\theta)\frac{n}{n+1}}$$