Last time.

Sample variance

Sec 11.3 Least square linear regression

Define them before using them

Today.

$\frac{X-\mathbb{E}X}{SD(X)} = \frac{X-\mu_X}{\sigma_X} = \frac{D_X}{\sigma_X}$

Some properties of $r_{X,Y} = \mathbb{E}X^*Y^* = \frac{\mathbb{E}D_X D_Y}{\sigma_X \sigma_Y}$

① Symmetric about X & Y.

$r_{X,Y} = r_{Y,X} = r$

good as long as no ambiguity.

Sec 11.4

② Bounds

by Cauchy-Schwarz inequality.

$\mathbb{E}XY \le \sqrt{\mathbb{E}X^2\,\mathbb{E}Y^2}$

$r = \mathbb{E}X^*Y^* \le \sqrt{\mathbb{E}X^{*2}\,\mathbb{E}Y^{*2}}$

$= 1$

$X^* = \frac{X-\mathbb{E}X}{SD(X)}$ is X in standard units

$\Rightarrow \mathbb{E}X^* = 0,\ SD(X^*) = 1$

$\mathbb{E}X^* = \mathbb{E}\frac{X-\mathbb{E}X}{SD(X)} = \frac{1}{SD(X)}(\mathbb{E}X - \mathbb{E}X) = 0$

$SD(X^*) = SD\left(\frac{X-\mathbb{E}X}{SD(X)}\right) = \frac{1}{SD(X)}SD(X) = 1$

$Var(X^*) = \mathbb{E}X^{*2} - \cancel{(\mathbb{E}X^*)^2} \to 0$

$\Rightarrow \mathbb{E}X^{*2} = Var(X^*) = 1^2 = 1$

Same for $\mathbb{E}Y^{*2} = 1$

lower bound for r $= -$ upper bound for $(-r)$

$-r = \mathbb{E}[(-X^*)\cdot Y^*] \le \sqrt{\mathbb{E}(-X^*)^2\,\mathbb{E}Y^{*2}}$

$= 1 \quad \Rightarrow r \ge -1$

Proof w/o. using C-S:

Consider $\underbrace{\mathbb{E}(X^* \mp Y^*)^2}_{\text{non-neg.}} \ge 0$

$\mathbb{E}X^{*2} \pm 2\mathbb{E}X^*Y^* + \mathbb{E}Y^{*2} \ge 0$

$\mp 2\mathbb{E}X^*Y^* \le \mathbb{E}X^{*2} + \mathbb{E}Y^{*2} = 2$

$r = \mathbb{E}X^*Y^* \le 1$

For lower bound. $-$

$r \ge -1$

bounds for r: -1, $+1$

When $r = 1$? $\quad \mathbb{E}(X^* - Y^*)^2 = 0$

$\Rightarrow X^* - Y^* = 0$

$\Rightarrow \frac{X-\mu_X}{\sigma_X} = \frac{Y-\mu_Y}{\sigma_Y}$

$\Rightarrow X$ is a linear function of Y

$X = aY + b$, $\quad a > 0$

$r = -1 \iff a < 0$

Sec 11.5 The error in regression

Error in regression

$D := Y - \hat{Y}$

$\hat{Y}$: least square linear regression estimator of Y based on X

Residual: the left-over part that cannot be estimated

$Y = \underbrace{\hat{Y}}_{\text{estimator}} + \underbrace{D}_{\text{Residual}}$

$\hat{Y} = \hat{a}X + \hat{b}$ $\qquad \hat{a} = r\cdot\frac{\sigma_Y}{\sigma_X}$, $\hat{b} = \mu_Y - \hat{a}\mu_X$

$= \hat{a}X + \mu_Y - \hat{a}\mu_X$

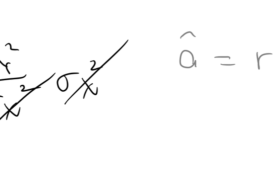
$D = Y - (\hat{a}X + \mu_Y - \hat{a}\mu_X)$

$= (Y - \mu_Y) - \hat{a}(X - \mu_X)$

$= D_Y - \hat{a}D_X$ — no intercept.

$\mathbb{E}D = \underbrace{\mathbb{E}D_Y}_{=0} - \hat{a}\underbrace{\mathbb{E}D_X}_{=0} = 0$

This means, average residual is always 0, no matter what the shape of scatter diagram is } description from Data 8



Or, in the prob. word,
no matter what the joint distr. of (X, Y) is

$Var(D) = \mathbb{E}D^2$ — MSE

Recall $\mathbb{E}D^2 = \mathbb{E}(D_Y - \hat{a}D_X)^2$

$= \mathbb{E}D_Y^2 - 2\hat{a}\,\mathbb{E}D_Y D_X + \hat{a}^2\mathbb{E}D_X^2$ $\qquad r = \frac{\mathbb{E}D_X D_Y}{\sigma_X\sigma_Y}$

$= \sigma_Y^2 - 2\hat{a}\cdot r\sigma_X\sigma_Y + \hat{a}^2\sigma_X^2$ $\qquad \hat{a} = r\frac{\sigma_Y}{\sigma_X}$

$= \sigma_Y^2 - 2r\frac{\sigma_Y}{\cancel{\sigma_X}}\cdot r\cancel{\sigma_X}\sigma_Y + r^2\frac{\sigma_Y^2}{\cancel{\sigma_X^2}}\cancel{\sigma_X^2}$

$= \sigma_Y^2 - 2r^2\sigma_Y^2 + r^2\sigma_Y^2$

$= (1 - r^2)\sigma_Y^2$, $\quad r = r_{X,Y}$

$SD(D) = \sqrt{1-r^2}\,\sigma_Y$ — Same formula as in Data 8

Since $\mathbb{E}D = 0$, when $SD(D)$ is small, by Chebyshev's inequality D is with high prob close to 0

? r is close to $+1$ or -1

i.e. $\hat{Y} \approx Y$ with high prob. $\Rightarrow$ good estimation.

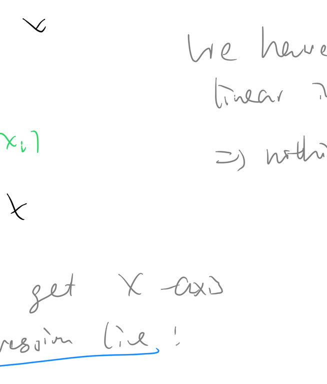
Another extreme case. $r = 0$ $\quad SD(D) = \sigma_Y$

$\hat{a} = r\frac{\sigma_Y}{\sigma_X} = 0 \quad \Rightarrow$ the best linear estimator of Y using X
= the best constant estimator

$\Rightarrow$ Do not need X / X provides no extra info.
when we estimate Y linearly

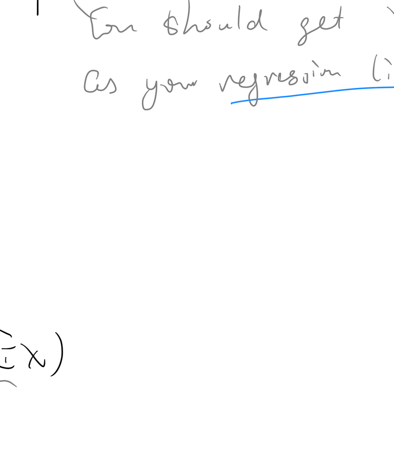
To summarize. $|r|$ quantifies the amount of linear association between X and Y

Note: even when $r = 0$, it is totally possible that X and Y have a non-linear association



The residual D, and X

Data 8: residual plot — flat



Imagine that you are doing another linear regression of D based on X

We have extracted all linear info in $\hat{Y}$
$\Rightarrow$ nothing left in D.

Slope $= \hat{a}_{D,X} = r_{D,X}\cdot\frac{\sigma_D}{\sigma_X}$

You should get X-axis as your regression line!

WTS slope $= 0$

$\Uparrow$

$r_{D,X} = \frac{Cov(D,X)}{\sigma_D\sigma_X} = 0$

$\Uparrow$

$Cov(D,X) = \mathbb{E}(D - \underbrace{\mathbb{E}D}_{=0})(\underbrace{X - \mathbb{E}X}_{D_X})$

$= \mathbb{E}\,D D_X$

$= \mathbb{E}[D_X(D_Y - \hat{a}D_X)]$

$= \mathbb{E}D_X D_Y - \hat{a}\,\mathbb{E}D_X^2$

$= r\sigma_X\sigma_Y - r\cdot\frac{\sigma_Y}{\cancel{\sigma_X}}\cdot\sigma_X^{\cancel{2}} = 0$

Intercept $= \mu_D - \hat{a}_{D,X}\mu_X$

$= \mathbb{E}D - 0 = 0$

Sec 12.1 Simple linear regression model

Single input $\longrightarrow$ machine / black box $\longrightarrow$ signal (linear function of input) $\longrightarrow$ error in observation $\longrightarrow$ output

What we know: input & output.

Wish to learn about the machine.
i.e. linear params,

Formally, n (input, output)
(x_i, Y_i) $\quad i = 1, 2, \dots, n$

$Y_i = \boxed{\beta_0 + \beta_1 x_i} + \boxed{\varepsilon_i}$ $\quad i = 1, 2, \dots, n.$

β_0, β_1 are unknown params that we wish to estimate

x_i's are consider known. fixed value

$\varepsilon_i \overset{i.i.d}{\sim} N(0, \sigma^2)$ $\quad$ (same var. for all i)

$\Rightarrow Y_i \sim N(\beta_0 + \beta_1 x_i, \sigma^2)$, indpt.

We wish to give an estimator of Y as

$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$

$Error_i = Y_i - \hat{Y}_i$

Target: find $(\hat{\beta}_0, \hat{\beta}_1)$ to minimizing MSE $= \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$

$Y_i \sim N(\underbrace{\beta_0 + \beta_1 x_i}_{\text{unknown constant}}, \sigma^2)$ $\quad$ indpt.

"The randomness of Y_i comes only from the noise ε_i

Therefore, the average output is

$\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i \sim N(\beta_0 + \beta_1\bar{x}, \frac{\sigma^2}{n})$

as linear combination of indpt normals

$\mathbb{E}\bar{Y} = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}Y_i = \frac{1}{n}\sum_{i=1}^{n}(\beta_0 + \beta_1 x_i)$

$= \frac{1}{n}\left[n\beta_0 + \beta_1\sum_{i=1}^{n}x_i\right]$

$= \beta_0 + \beta_1\frac{1}{n}\sum_{i=1}^{n}x_i$

$= \beta_0 + \beta_1\bar{x}$

"the same linear transform of the average input"

$Var(\bar{Y}) = Var(\text{Constant} + \frac{1}{n}\sum_{i=1}^{n}\varepsilon_i)$

$\bar{Y} = \frac{1}{n}\sum_{i=1}^{n}Y_i$, $\quad Y_i = \underbrace{\beta_0 + \beta_1 x_i}_{\text{const.}} + \underbrace{\varepsilon_i}_{\text{random}}$

$= Var(\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i)$

$= \frac{1}{n^2}Var(\underbrace{\sum_{i=1}^{n}\varepsilon_i}_{\text{indpt}})$

$= \frac{1}{n^2}\sum_{i=1}^{n}Var(\varepsilon_i)$

$= \frac{1}{n^2}\sum_{i=1}^{n}\sigma^2$

$= \frac{1}{n^2}\cdot n\sigma^2 = \frac{\sigma^2}{n}$