Lecture - 07 02/18/2020

$P = 1$, $y = \mathbb{R}$ $\quad H = \{\vec{w} \cdot \vec{x} ; \vec{w} \in \mathbb{R}^2\}$ (linear modeling)

x_i = Binary $\Rightarrow h^*(x) = \beta_0 + \beta_1 x$,

$y = h^*(x) + \varepsilon$

$g(x) = b_0 + b_1 X$

Error due to ignoring and misspecfication

A: OLS $\Rightarrow b_0 = \bar{Y} - r \frac{S_y}{S_x} \bar{x}$, $b_1 = r \frac{S_y}{S_x}$

Ordinary least square

$Y = g(x) + e$ ← residual, (all 3 errors)

How well does g predict?

$$SSE = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (Y_i - g(x_i))^2$$

Intepretable? units?

Y metric - squard ; not so intepretable.

$MSE := \frac{1}{n-2} SSE$ units: Y-metric squard not so intepretable

Mean Squard Error ← forget this...

$RMSE := \sqrt{MSE}$ Units: Y, very intepretable

Root mean squard error

Imagine if e was a realization from a normal distribution

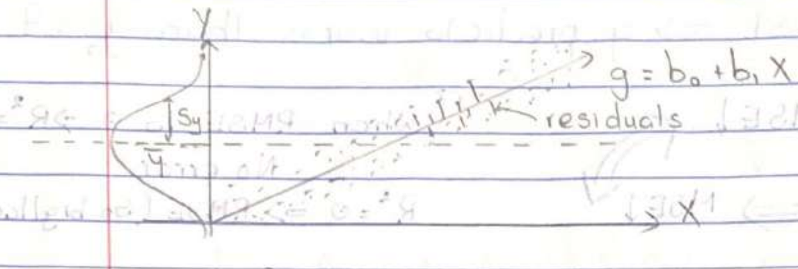
You can show that $[g(x) \pm 2\,RMSE]$ ← $[RMSE, RMSE, RMSE \; g(x)]$

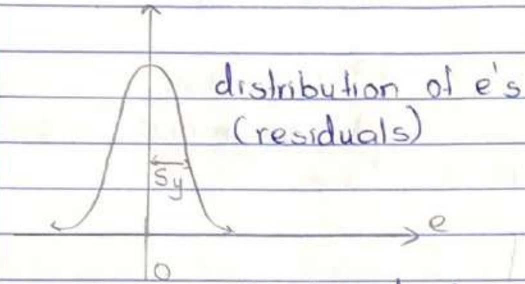
≈ 95% prediction interval

how accurate your model is.

Consider the null model $g_0(x) = \bar{y}$

$$SSE_0 = \sum_{i=1}^{n} (y_i - \bar{y})^2 = SST \quad \text{sum square total}$$
$$= (n-1)\, s^2 y$$





$SSE = \sum e_i^2 = (n-1)\, s^2 e$

$\Delta s^2 = s^2 y - s^2 e$

"Reduction in variance",

"Variance explained"

$$R^2 = \frac{\Delta s^2}{s^2 y} \frac{(\text{explained variance})}{(\text{total variance})} = \frac{s^2 y - s^2 e}{s^2 y} = 1 - \frac{s^2 e}{s^2 y}$$

$$\approx \frac{Var(Y) - Var(\overset{\text{residuals}}{E})}{Var[Y]}$$

$$= \frac{(n-1)\, s^2 y - (n-1)\, s^2 e}{(n-1)\, s^2 y} = \frac{SST - SSE}{SST} = 1 - \frac{SSE}{SST}$$

R^2 is the "proportion of variance explained"

Can $R^2 > 1$?

$1 - \frac{SSE}{SST}$; since $SSE \geq 0$, $SST \geq 0$

$\implies \frac{SSE}{SST} \geq 0$; $1 - \frac{SSE}{SST} \leq 1$

When is $R^2 = 1 \implies SSE = 0$

$\implies e_i = 0 \ \forall i$

$\implies y_i = \hat{y}_i \ \forall i$

Can $R^2 = 0$?

$SSE = SST \Rightarrow g = \bar{y}$

Can $R^2 < 0$?

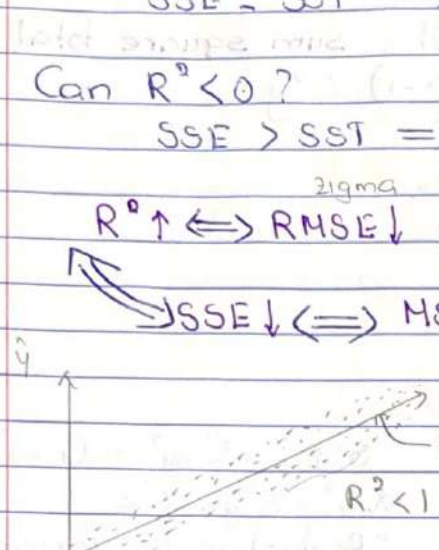
$SSE > SST \Rightarrow g$ predicts worse than $g_0 = \bar{y}$

$$R^2 \uparrow \Leftrightarrow RMSE \downarrow \quad \text{(zigma)}$$

$$SSE \downarrow \Leftrightarrow MSE \downarrow$$

(When RMSE is $0 \Rightarrow R^2 = 1$; No error

$R^2 = 0 \Rightarrow$ RMSE (so big/large #)



R^2 VS. RMSE

Who is more important?

$x \in X = \{\overset{\to 0}{red}, \overset{\to 1}{green}\}$

$x = \mathbb{1}_{x = green} \in \{0, 1\}$

$H = \{w_0 + w_1 x : w_0, w \in \mathbb{R}\}$

$\hat{y} = g(x) = b_0 + b_1 x$

$g(x) = \begin{cases} \bar{y}_{red} & \text{if } x = 0 \\ \bar{y}_{green} & \text{if } x = 1 \end{cases}$ let's prove A: OLS returns this

$= \underset{= b_0}{\bar{y}_{red}} + \underbrace{(\bar{y}_{green} - \bar{y}_{red})}_{= b_1} x$

$\bar{x} = \dfrac{\sum x_i}{n} = \dfrac{n_g}{n} = p \leftarrow$ prob. of green. (n_g = # green)

$\bar{y} = \dfrac{\sum y_i}{n} = \dfrac{\sum_{i: green = \{i : x_i = 1\}} y_i + \sum_{i; red = \{i : x_i = 0\}} y_i}{n}$

$$= \frac{\sum_{\{i: x_i = 1\}} y_i}{n} \cdot \frac{n_g}{n_g} + \frac{\sum_{\{i: x_i = 0\}} y_i}{n} \cdot \frac{(n-n_g)}{(n-n_g)}$$

$$= \frac{\sum_{\{i, x_i = 1\}} y_i}{n_g} \cdot \frac{n_g}{n} + \overbrace{\frac{\sum_{\{i, x_i = 0\}} y_i}{n - n_g}}^{\text{red}} \cdot \frac{n - n_g}{n}$$

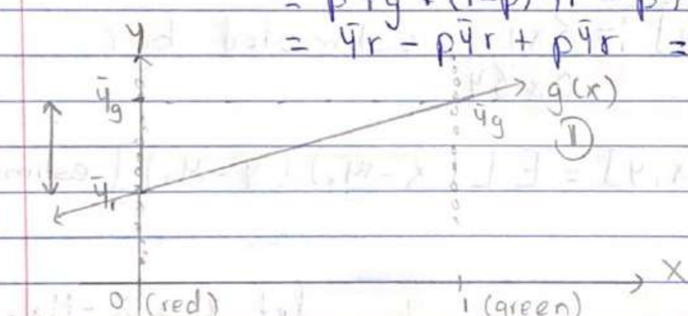
$$= p\bar{y}\,\text{green} + (1-p)\,\bar{y}\,\text{red}$$

$$b_1 = \frac{\sum x_i y_i - n\bar{x}\bar{y}}{\sum x_i^2 - n\bar{x}^2}$$

$$= \frac{n_g \bar{y}_g - np\bar{y}}{n_g - np^2} \cdot \frac{1/n}{1/n} = \frac{p\bar{y}_g - p\bar{y}}{p - p^2} = \frac{\bar{y}_g - \bar{y}}{1-p}$$

$$= \frac{\bar{y}_g - p\bar{y}_g - (1-p)\bar{y}_r}{1-p} = \frac{(1-p)\bar{y}_g - (1-p)\bar{y}_r}{(1-p)} = \bar{y}_g - \bar{y}_r$$

$$b_0 = \bar{y} - b_1\bar{x} = (p\bar{y}_g + (1-p)\bar{y}_r) - (\bar{y}_g - \bar{y}_r)p$$
$$= p\bar{y}_g + (1-p)\bar{y}_r - p\bar{y}_g + p\bar{y}_r$$
$$= \bar{y}_r - p\bar{y}_r + p\bar{y}_r = \bar{y}_r$$



reference category $= \overset{b_0}{\bar{y}_{red}} + \overbrace{(\bar{y}_{green} - \bar{y}_{red})}^{b_1} x$

$= 0.7 + 0.2x$

$$\frac{\bar{y}_g - \bar{y}}{1-p} = \frac{\bar{y}_g - p\bar{y}_g - (1-p)\bar{y}_r}{1-p}$$

Q := $L = 3$ e.g

$X \in$ [red, green, blue]

$X_1 = \mathbb{1}_{x = \text{green}} \in [0,1]$

$X_2 = \mathbb{1}_{x = \text{blue}} \in [0,1]$

$$\hat{y} = \overset{0.8}{b_0} + \overset{0.3}{b_1} X_1 + \overset{-0.4}{b_2} X_2$$

$$b_0 = \bar{y}_r, \quad b_1 = \bar{y}_g - \bar{y}_r, \quad b_2 = \bar{y}_b - \bar{y}_r$$

$X \in$ [low, medium, high] ordinal categorical.

$X_1 := \mathbb{1}_{x = m}$, $Y_1 = \mathbb{1}_{x = H}$

$$\hat{y} = \begin{cases} \hat{y}_L & \text{if } x = \text{low} \\ \hat{y}_M & \text{if } x = \text{medium} \\ \hat{y}_H & \text{if } x = \text{High} \end{cases}$$

However, what if you with to contain $\hat{y}$ (low) < $\hat{y}$ (med) < $\hat{y}$ (high) ?

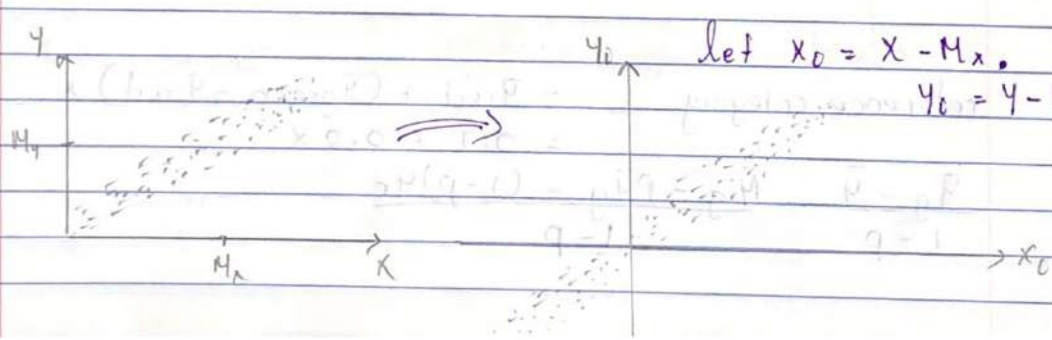
Does. A: OLS give you this? No!

They are depends if "associated" if

$\exists x_1, x_2$ s.t $P(Y \mid X = x_1) \neq P(Y \mid X = x_2)$

$\rho := \text{Cov}[x,y] := \dfrac{\sigma_{xy}}{\sigma_x \sigma_y}$ estimated by r.

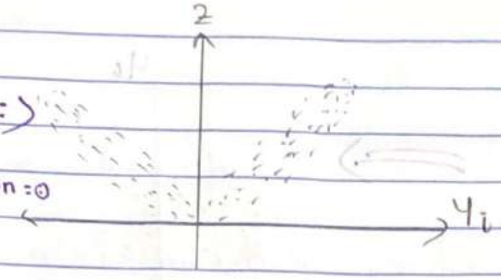
$\sigma_{xy} := \text{Cov}[x,y] = E[(X - \mu_x)(Y - \mu_y)]$ estimated by S_{xy}



let $X_0 = X - M_x$.

$Y_0 = Y - M_y$

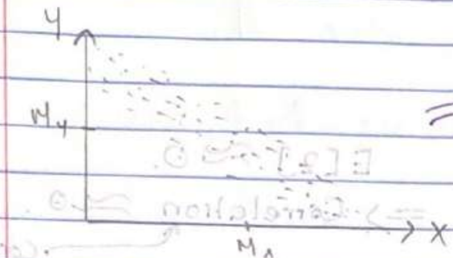
$Z = X_c Y_c$ $\Rightarrow$
centered
mean expectation $= 0$



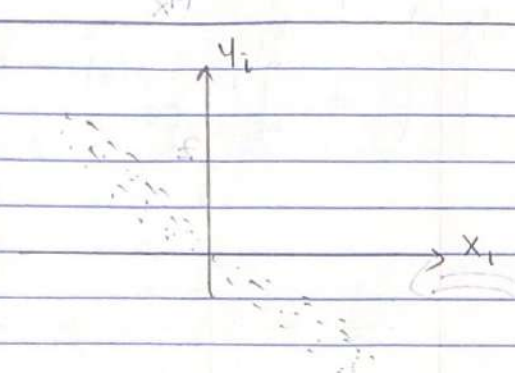
$E[Z] > 0$.

$\sigma_{xy} = E[(X - \mu_x)(Y - \mu_y)]$
$= E[X_i Y_i] = E[Z]$

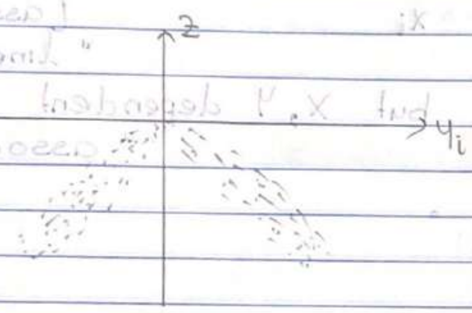
opposite



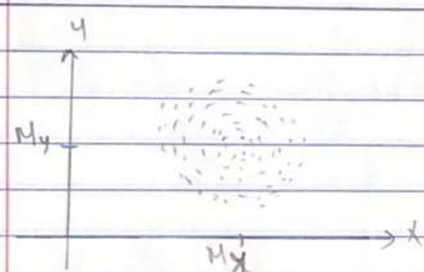
$\Rightarrow$



$\Rightarrow$

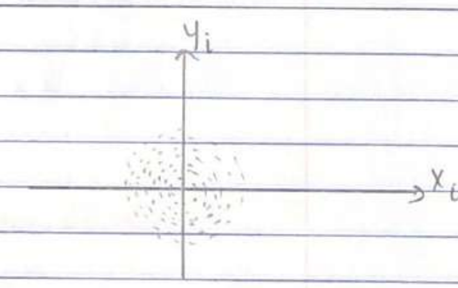


$E[Z] < 0$.

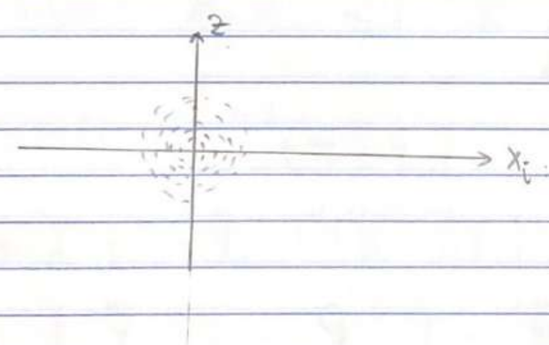


$\Rightarrow$

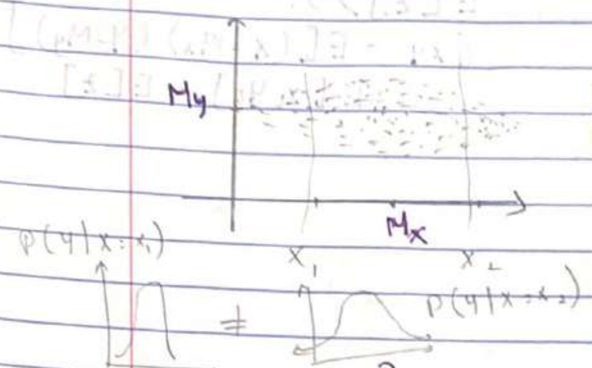
no covariance
no correlation



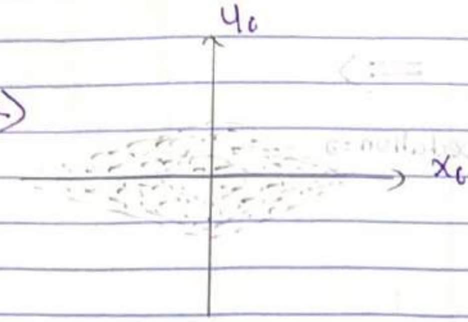
$\Rightarrow$



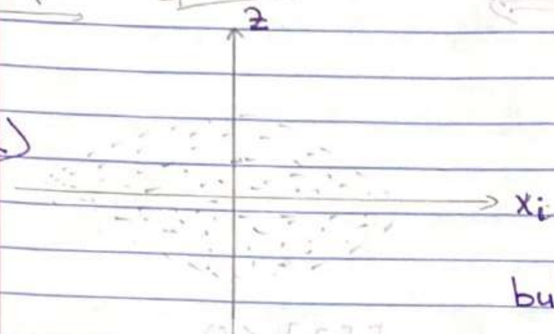
$E[Z] \approx 0$



$\Longrightarrow$



$\Longrightarrow$



$E[z] \approx 0.$

$\Rightarrow$ Correlation ≈ 0.

Correlation $\subseteq$ [associations]

"linear association"

but x, y dependent or "associated"

In HW sample correlation $r^2 = R^2$ for $p=1$