Lecture 3.1 09/?/20

(x) $p=1$, $\mathcal{Y}=\mathbb{R}$, $\mathcal{H}=\{\vec{w}\cdot\vec{x} : \vec{w}\in\mathbb{R}^2\}$ (linear modeling)

$$\Rightarrow h^*(x)=\beta_0+\beta_1 x, \quad y=h^*(x)+\mathcal{E}$$

error due to ignorance and misspecification

$$g(x)=b_0^{\#}+b_1^{\#}x$$

$$\mathcal{A}: OLS \Rightarrow b_0=\bar{y}-r\frac{s_y}{s_x}\bar{x}, \quad b_1=r\frac{s_y}{s_x}$$

$$y=g(x)+e \rightarrow \text{residuals (all 3 errors)}$$

(x) How well does g predict?

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

interpretable? units?

y-metric - squared

Not so important.

$$MSE := \frac{1}{n-2} SSE$$

units: y-metric-squared not so interpretable.

Mean Squared Error

forget this.

$$RMSE := \sqrt{MSE}$$

units: y, very interpretable

root mean squared error

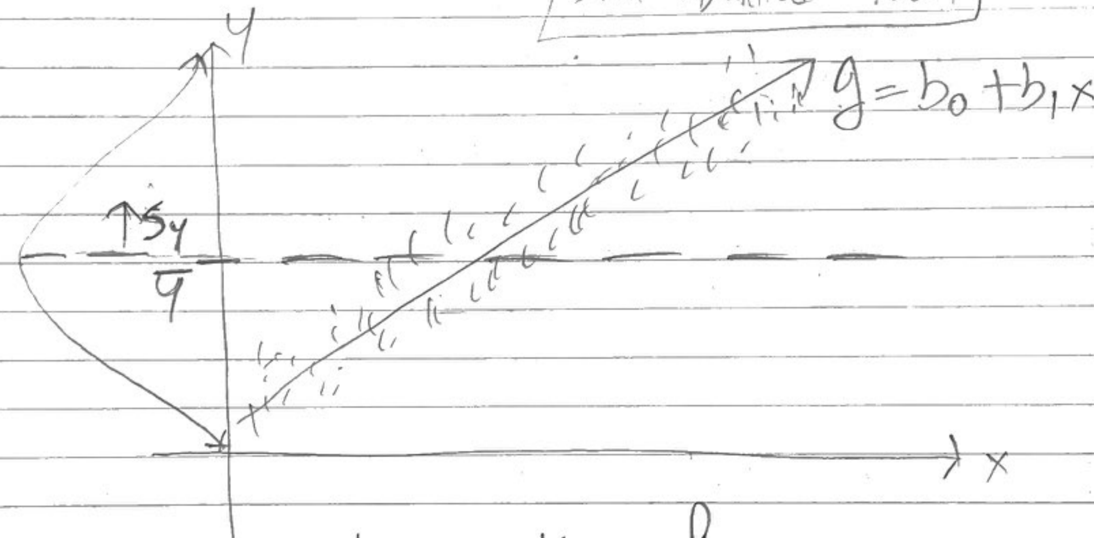
Imagine e with a realization from a normal dist, you can show

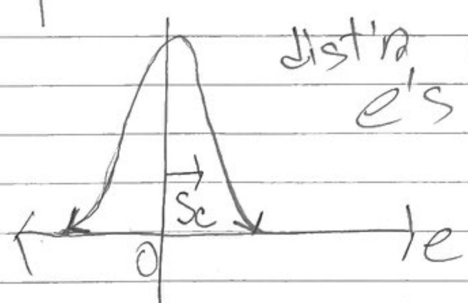
95% prediction interval $[g(x) \pm 2RMSE]$, How accurate your model is.

Consider the null model $\hat{y}_0(x) = \bar{y}$

$$SSE_0 = \sum_{i=1}^{n} (y_i - \bar{y})^2 = SST = (n-1)\, s_y^2$$

$\boxed{\text{Sum squarred Total}}$





dist'n of e's (residials)

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

$\Delta s^2 = s_y^2 - s_e^2$ "reduction in variance", "valiance explained"

$$R^2 = \frac{\Delta s^2}{s_y^2} = \frac{s_y^2 - s_e^2}{s_y^2} = 1 - \frac{s_e^2}{s_y^2} = \frac{(n-1)s_y^2 - (n-1)s_e^2}{(n-1)s_y^2}$$

$$\approx \frac{\mathrm{Var}[Y] - \mathrm{Var}[E]}{\mathrm{Var}[Y]}$$

$E \to \boxed{\text{residials}}$

$\circledast$ R^2 is the "Propotion of Variance explained"

$$= \frac{SST - SSE}{SST} = 1 - \frac{SSE}{SST}$$

$\circledast$ Can $R^2 > 1$?

$$\overset{\shortparallel}{1 - \frac{SSE}{SST}} \leq 1$$

Since $SSE \geq 0$, $SST \geq 0$

$$\Rightarrow \frac{SSE}{SST} \geq 0$$

$$R^2 = 1 \Rightarrow SSE = 0 \Rightarrow e_i = 0 \ \forall i$$

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

$\circledast$ Can $R^2 = 0$?

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

$\circledast$ Can $R^2 < 0$?

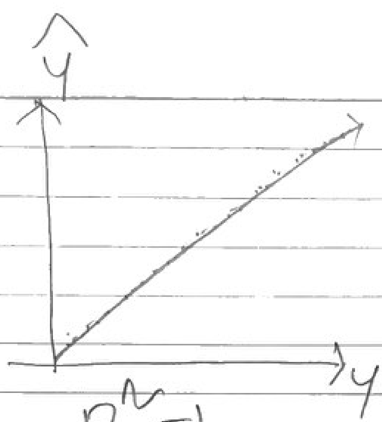
$SSE > SST$

$\Rightarrow \hat{y}$ predicts worse then $\hat{y}_0 = \bar{y}$

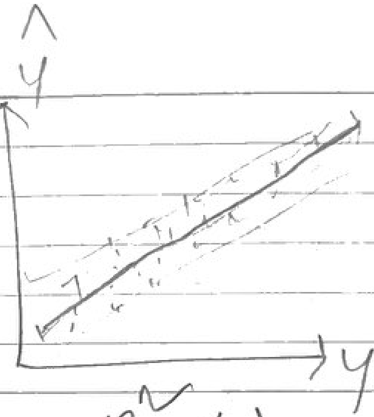
$\circledast$ $R^2 \uparrow \Longleftrightarrow RMSE \downarrow$

$$\Updownarrow$$

$SSE \downarrow \Longleftrightarrow MSE \downarrow$



$R^2 = 1$



$R^2 < 1$

(x) R^2 vs RMSE.

Who is more important?

(x) $x \in \mathcal{Y} = \{\overset{0}{\text{red}}, \overset{1}{\text{green}}\}$

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

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

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

$$g(x) = \begin{cases} \bar{y}_{red} & \text{if } x = 0 \\ \bar{y}_{green} & \text{if } x = 1 \end{cases}$$

Let's prove $\mathcal{A}$: OLS returns this:

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

$$= 0.7 + 0.2 x$$

$$\bar{X} = \frac{\sum x_i}{n} = \frac{n_g}{n} = p \rightarrow \text{prop. of green}$$

$n_g \leftarrow$ # green

$$\bar{Y} = \frac{\sum y_i}{n} = \frac{\sum y_i}{n} = \sum_{\substack{\{i:\, green\} \\ = \{i:\, x_i = 1\}}} y_i + \sum_{\substack{\{i:\, red\} \\ = \{i:\, x_i = 0\}}} y_i$$

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

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

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

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

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

$$= \frac{p\bar{y}_g - p\bar{y}}{p - p^2}$$

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

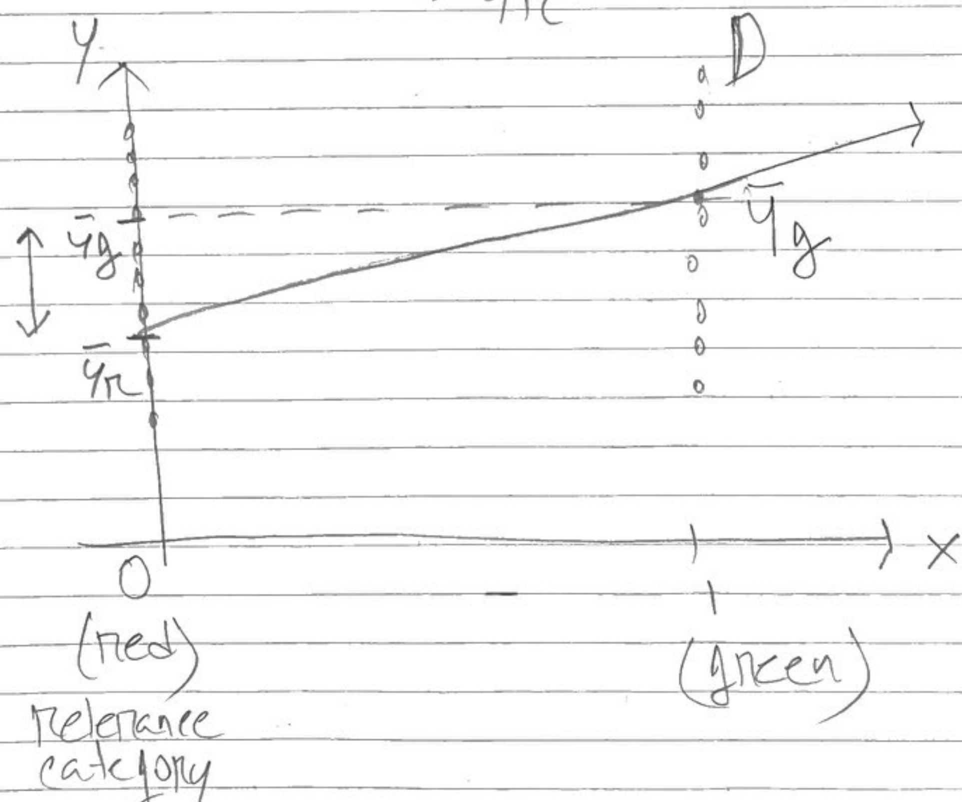
$$= \frac{(1-p)\bar{y}_g - (1-p)\bar{y}_r}{1-p} = \bar{y}_g - \bar{y}_r$$

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

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

$$= \bar{y}_r - \cancel{p\bar{y}_r} + \cancel{p\bar{y}_r}$$

$$= \bar{y}_r$$



$\circledast$ $L = 3$ e.g.

$x \in \{red, green, blue\}$

$x_1 = \mathbb{1}_{x = green} \in \{0,1\}$

$x_2 = \mathbb{1}_{x = blue} \in \{0,1\}$

$$\hat{y} = \overset{0.8}{\underset{\bar{y}_r}{b_0}} + \overset{0.3}{\underset{(\bar{y}_g - \bar{y}_r)}{b_1}} x_1 + \overset{-0.4}{\underset{(\bar{y}_b - \bar{y}_r)}{b_2}} x_2$$

$\circledast$ $x \in \{Low, medium, high\}$ ordinal category

$\Downarrow$

$x_1 = \mathbb{1}_{x = M}$, $x_2 = \mathbb{1}_{x = H}$

$$\hat{y} = \begin{cases} \bar{y}_L & \text{if } x = Low \\ \bar{y}_m & \text{if } x = medium \\ \bar{y}_H & \text{if } x = High \end{cases}$$

$\circledast$ What if you wish to constraint

$\hat{y}(Low) < \hat{y}(Med) < \hat{y}(high)$?

Does A: OLS give you this? $\boxed{NO}$

(*) Consider the r.v's X, Y

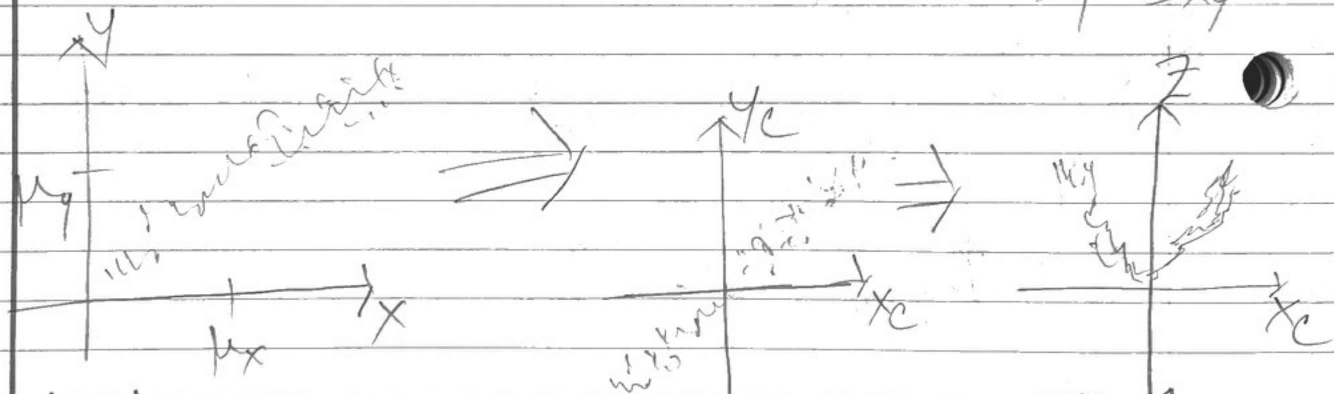
They are dependent if... ("associated" if...)

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

$\rho := \mathrm{Corr}[X, Y] = \dfrac{\sigma_{xy}}{\sigma_x \sigma_y}$ estimated by r.

$\sigma_{xy} = \mathrm{Cov}[X, Y] = E[(X - \mu_x)(Y - \mu_y)]$

estimated by s_{xy}



Let $X_c = X - \mu_x$

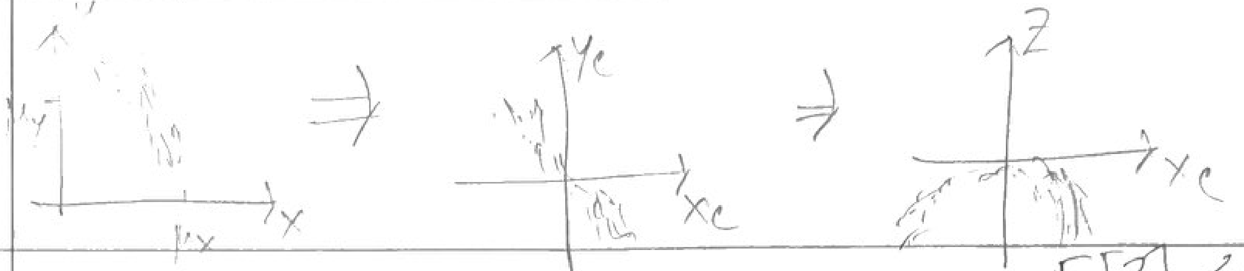
$Y_c = Y - \mu_y$

$E[Z] > 0$

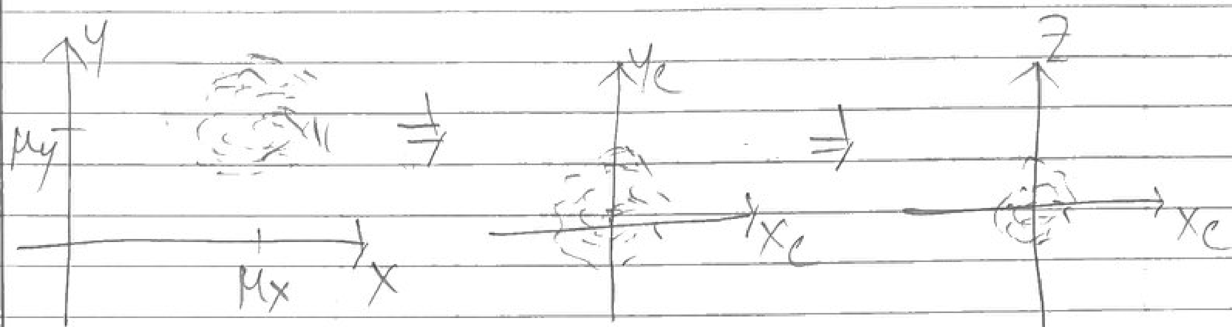
$\sigma_{xy} = E[(X - \mu_x)(Y - \mu_y)]$

$= E[X_c Y_c] = E[Z]$

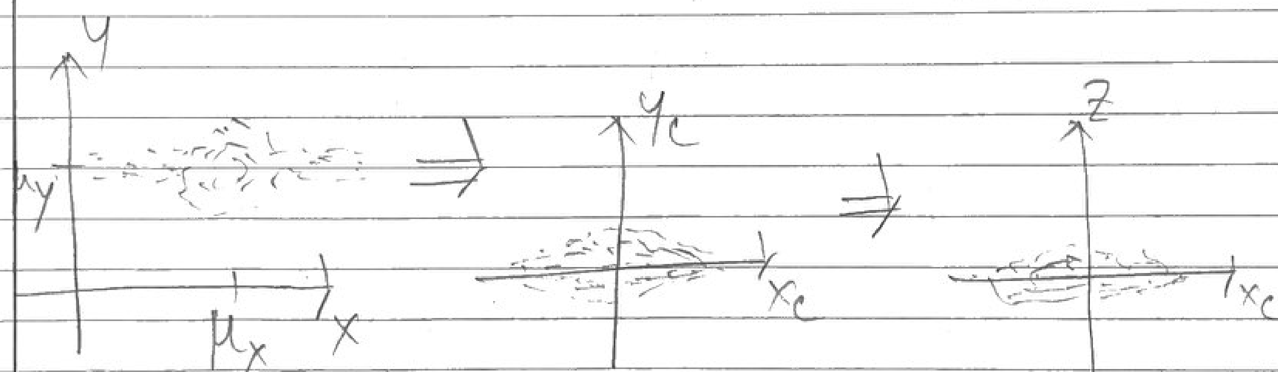
$Z = X_c Y_c$



$E[Z] < 0$



$E[Z] \approx 0$



$E[Z] \approx 0$

Correlation $\subset$ {association} $\Rightarrow$ Correlation ≈ 0 but X, Y are dependent OR "association"

||

"linear association"

$P(Y|X=x_1)$ $\neq$ $P(Y|X=x_2)$