Tasks (3.2)

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + u_i$$

$$X = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 1 \end{pmatrix} \quad y = \begin{pmatrix} 1 \\ 2 \\ 3 \\ 4 \\ 5 \end{pmatrix} \quad \beta = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \end{pmatrix}$$

$$X'X = \begin{pmatrix} 5 & 2 & 1 \\ 2 & 2 & 1 \\ 1 & 1 & 1 \end{pmatrix} \quad (X'X)^{-1} = \begin{pmatrix} \frac{1}{3} & -\frac{1}{3} & 0 \\ -\frac{1}{3} & \frac{4}{3} & -1 \\ 0 & -1 & 2 \end{pmatrix}$$

a) num. of observations $n = 5$

b) num. of regressors $K = 3$

2 dummy-variables and a constant

$x_{ij} \in \{0, 1\}$

c) $\hat{\beta}_{OLS} = (X'X)^{-1} X'y =$

$$= \begin{pmatrix} \frac{1}{3} & -\frac{1}{3} & 0 \\ -\frac{1}{3} & \frac{4}{3} & -1 \\ 0 & -1 & 2 \end{pmatrix} \begin{pmatrix} 1 & 1 & 1 & 1 & 1 \\ 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} 1 \\ 2 \\ 3 \\ 4 \\ 5 \end{pmatrix} =$$

$$= \begin{pmatrix} \frac{1}{3} & -\frac{1}{3} & 0 \\ -\frac{1}{3} & \frac{4}{3} & -1 \\ 0 & -1 & 2 \end{pmatrix} \begin{pmatrix} 15 \\ 9 \\ 5 \end{pmatrix} = \begin{pmatrix} 2 \\ 2 \\ 1 \end{pmatrix}$$

d) $\hat{y} = X\hat{\beta} = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 1 \end{pmatrix} \cdot \begin{pmatrix} 2 \\ 2 \\ 1 \end{pmatrix} = \begin{pmatrix} 2 \\ 2 \\ 2 \\ 4 \\ 5 \end{pmatrix}$

or we can use projection matrix

hat matrix

$P = X(X'X)^{-1}X'$

$Py = \underbrace{X(X'X)^{-1}X'y}_{\hat{\beta}} = X\hat{\beta} = \hat{y}$

$$P = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 1 \end{pmatrix} \begin{pmatrix} \frac{1}{3} & -\frac{1}{3} & 0 \\ -\frac{1}{3} & \frac{4}{3} & -1 \\ 0 & -1 & 2 \end{pmatrix} \begin{pmatrix} 1 & 1 & 1 & 1 & 1 \\ 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix} =$$

$$= \begin{pmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 1 \end{pmatrix} \begin{pmatrix} \frac{1}{3} & \frac{1}{3} & \frac{1}{3} & 0 & 0 \\ -\frac{1}{3} & -\frac{1}{3} & -\frac{1}{3} & 1 & 0 \\ 0 & 0 & 0 & -1 & 1 \end{pmatrix} =$$

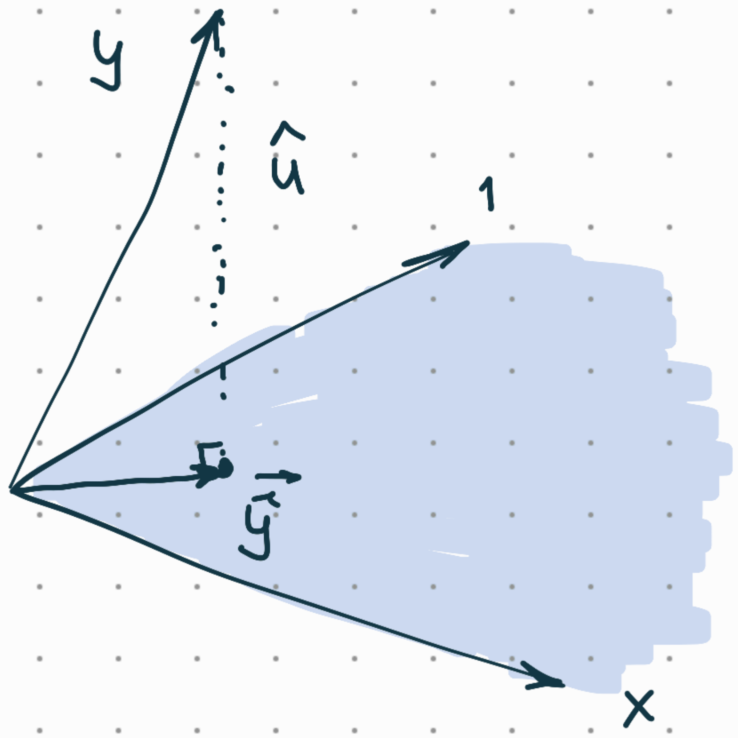
$$= \begin{pmatrix} \frac{1}{3} & \frac{1}{3} & \frac{1}{3} & 0 & 0 \\ \frac{1}{3} & \frac{1}{3} & \frac{1}{3} & 0 & 0 \\ \frac{1}{3} & \frac{1}{3} & \frac{1}{3} & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix} \cdot \times \begin{pmatrix} 1 \\ 2 \\ 3 \\ 4 \\ 5 \end{pmatrix}$$

$$= \begin{pmatrix} 2 \\ 2 \\ \vdots \end{pmatrix}$$

$$Py = \begin{pmatrix} 2 \\ 2 \\ 4 \\ 5 \end{pmatrix}$$

$M = I - P$ annihilator matrix

$$My = y - Py = y - \hat{y} = \hat{u}$$



$$\hat{y} = f(1, x)$$

$$P = X\underbrace{(X'X)^{-1}X'}_{A} = XA$$

$$\hat{y}'\hat{u} = (Py)'\hat{u} = y'P'(I-P)y =$$

$$= y'(P - \underbrace{PP}_{})y = y'(P-P)y = 0$$

idempotent matrix $PP = P$

$\Rightarrow \hat{y} \perp \hat{u}$

e) Analysis of Variance

$$\{ y = Py + My = \hat{y} + \hat{u}$$

$$\begin{cases} \ldots \\ \hat{y} \perp \hat{u} \end{cases}$$

if X contains a constant

$$\underbrace{\sum (y_i - \bar{y})^2}_{} = \underbrace{\sum (\hat{y}_i - \bar{y})^2}_{SSE} + \underbrace{\sum \hat{u}_i^2}_{SSR}$$

SST sum of sq. total

SSE sum of sq. explained

SSR sum of sq. residuals

$$y = \begin{pmatrix} 1 \\ 2 \\ 3 \\ 4 \\ 5 \end{pmatrix} \quad \hat{y} = \begin{pmatrix} 2 \\ 2 \\ 2 \\ 4 \\ 5 \end{pmatrix} \quad \hat{u} = \begin{pmatrix} -1 \\ 0 \\ 1 \\ 0 \\ 0 \end{pmatrix}$$

$\bar{y} = 3$

$SST = 2^2 + 1^2 + 0 + 1^2 + 2^2 = 10$

$SSR = 1 + 1 = 2$

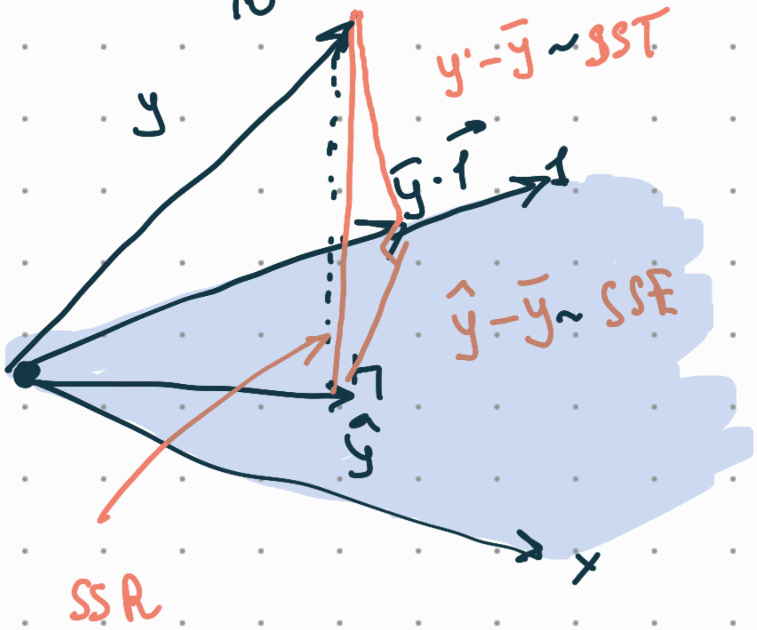
$SSE = 8$

f) coefficient of determination

$$R^2 = \frac{SSR}{SST} \overset{\text{constant}}{=} 1 - \frac{SSE}{SST}$$

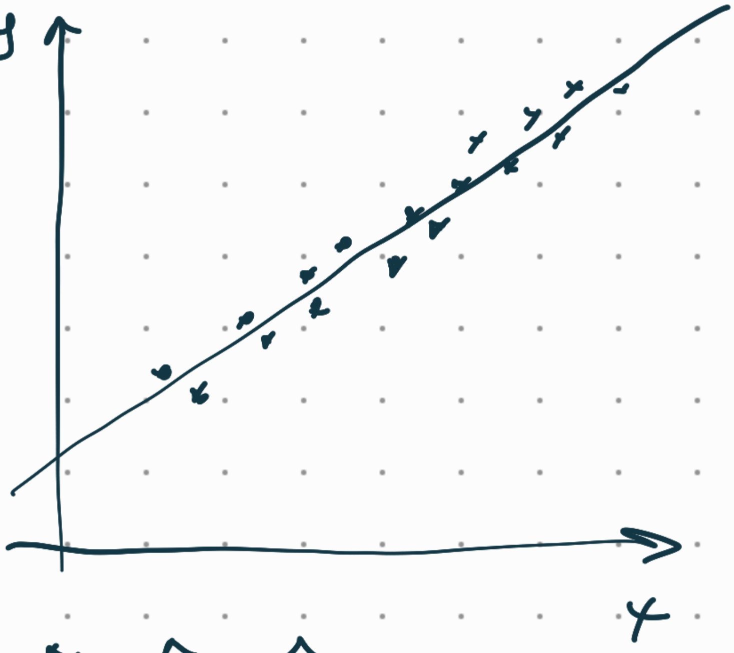
If there is a constant in regression, $R^2 \in [0; 1]$

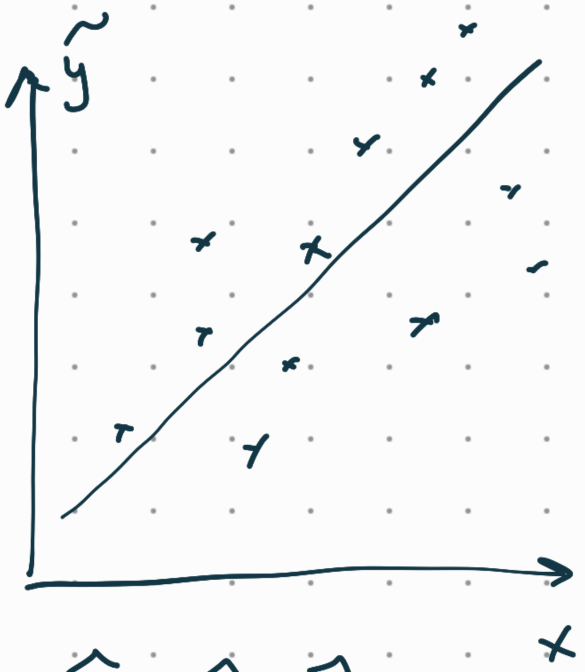
$R^2 = \frac{2}{10} = 0{,}8$ ← good or bad?



Task 2 Is R^2 a good measure of anything?

a) Is large R^2 a good measure of fit in any case? No





$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 X \longleftrightarrow \hat{\tilde{y}} = \hat{\beta}_0 + \hat{\beta}_1 X$

the same $\hat{\beta}_0, \hat{\beta}_1$

the same SSE!

the same fitting!

But $SST_y < SST_{\tilde{y}}$, because
in case of $\tilde{y}$ the variance is bigger $\Rightarrow$

$$\Rightarrow R^2_y = \frac{SSE_y}{SST_y} > \frac{SSE_{\tilde{y}}}{SST_{\tilde{y}}} \quad :($$

$\boxed{\text{we should compare models only with fixed } \{y_1, y_2 \dots y_n\} \text{ to have SST fixed}}$

---

b) We had $\{(x_i, y_i)\}_{i=1}^{n}$, SSE_1,
SST_1,
SSR_1

now we have additional (x_{n+1}, y_{n+1}), new observ.
SSE_2, SST_2, SSR_2

What is the reaction of R^2?

- $R^2 = \dfrac{SSE}{SST}$, $\quad SST_1 = \sum_{i=1}^{n} (y_i - \bar{y}_{old})^2$

- $SST_2 = \sum^{n+1} (y_i - \bar{y}_{new})^2 = \sum^{n} (y_i - \bar{y}_{new})^2 +$

$SST_2 = \sum_{i=1} (y_i - \bar{y}_{new})^2 = \sum_{i=1} (\dots$

$+ (y_{n+1} - \bar{y}_{new})^2 = \sum_{i=1} (y_i - \bar{y}_{old} + \bar{y}_{old} - \bar{y}_{new})^2 +$

$+ (y_{n+1} - \bar{y}_{new})^2 = \left[(a+b)^2 = a^2 + b^2 + 2ab \right] =$

$= \sum_{i=1}^{n} (y_i - \bar{y}_{old})^2 + \sum_{i=1}^{n} (\bar{y}_{old} - \bar{y}_{new})^2 +$

$+ \underbrace{2 \sum_{i=1}^{n} (y_i - \bar{y}_{old})(y_i - \bar{y}_{new})}_{=0} + (y_{n+1} - \bar{y}_{new})^2 \ominus$

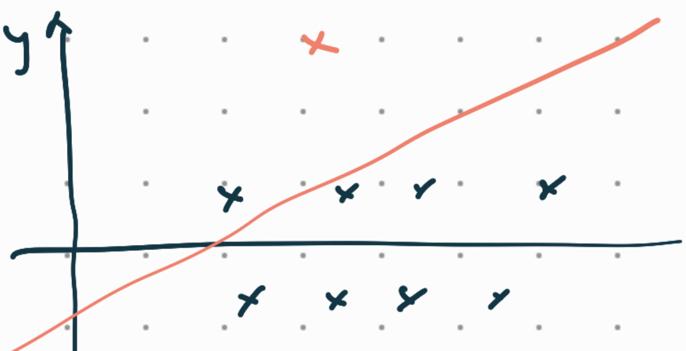
$0, \quad \sum_{i=1}^{n} (y_i - \bar{y}) = \sum_{i=1}^{n} y_i - n\bar{y} =$

$= n\bar{y} - n\bar{y} = 0$

$\ominus \underbrace{\sum_{i=1}^{n} (y_i - \bar{y}_{old})^2}_{= SST_1} + \underbrace{n(\bar{y}_{old} - \bar{y}_{new})^2}_{\geq 0} + \underbrace{(y_{n+1} - \bar{y}_{new})^2}_{\geq 0} \Rightarrow$

$\Rightarrow SST_2 \geq SST_1$!

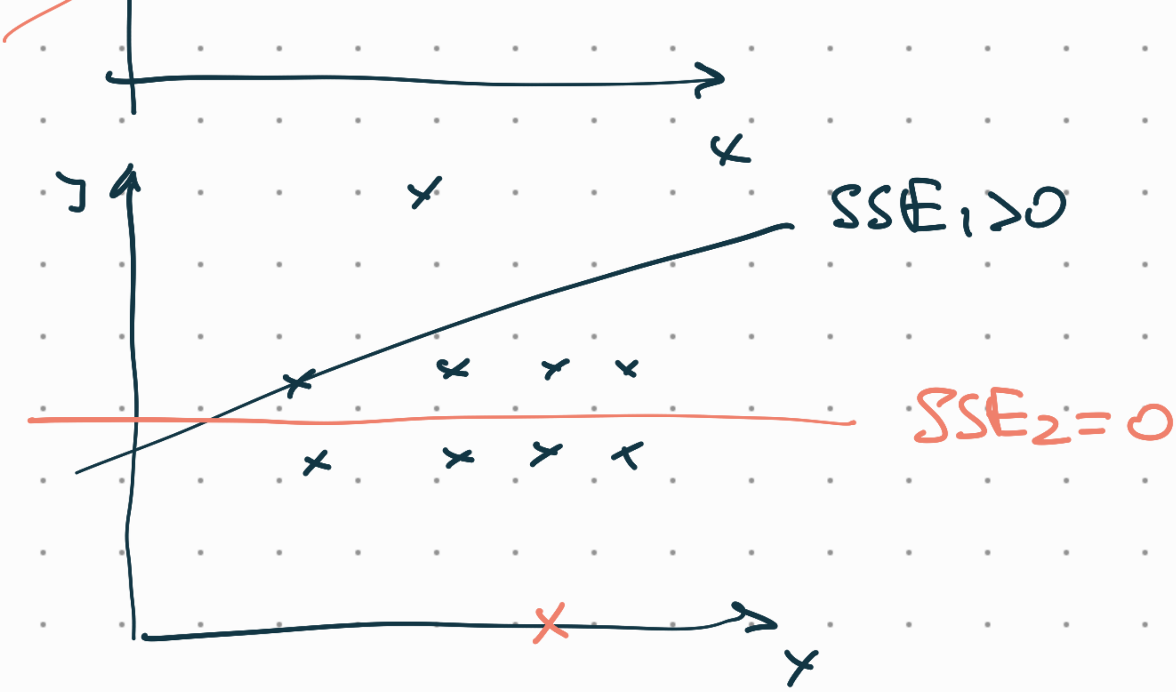
- what about SSE_2 vs SSE_1?



$SSE_2 > 0$

$SSE_1 = 0$

SSE↑



SSE $\downarrow$

- $R^2 = \frac{SSE}{SST}$ $\quad \left.\begin{matrix} SST \uparrow \\ SSE\ ? \end{matrix}\right\} R^2\ ?$

we don't know how R^2 will change!

---

c) <u>New regressors</u> $\qquad R_2^2$ vs R_1^2 ~ ?

1. $y_i = \beta_0 + \beta_1 x_i + u_i$ | 2. $y_i = \beta_0 + \beta_1 x_i + \beta_2 z_i + u_i$

y – salary
x – KPI
z – retrograde mercury   or some unimportant regressor for salary prediction.

- $R_1^2 = \frac{SSE_1}{SST_1}$ $\qquad R_2^2 = \frac{SSE_2}{SST_2}$

$$SST_1 = SST_2 = \sum_{i=1}^{n} (y_i - \bar{y})^2$$ → no changes

$SSE_1 \leq SSE_2$ because our predictions coudn't get worse if we have more information (even it is irrelevant)

⇒ $R_2^2 \geq R_1^2$ and R^2 always grows with the increasing number of any regressors

⇒ It is bette to use adjusted R^2:

$$R^2 = 1 - \frac{SSR}{SST} \qquad R^2_{adj} = 1 - \frac{SSR / n-k}{SST / n-1}$$

R^2_{adj} with penalty k – number of regressors (include constant)

$n-k$ → degrees of freedom for random variable SSR

$n-1 \rightarrow$ degrees of freedom for random variable SST.