Lecture 3 - 24 Oct 2023 (part 2)

PARTITION OF THE SUM OF SQUARES AND COEFFICIENT OF DETERMINATION R^2

We observe two variables x and y on n units.
A first descriptive statistic we can compute is the correlation coefficient

$$r_{xy} = \frac{s_{xy}}{s_x s_y} \in [-1,1]$$

where
$$s_{xy} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})$$
$$s_x^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2$$
$$s_y^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i-\bar{y})^2$$

which is a measure of the strenght of a linear relationship between the two variables

In the context of linear regression we can derive another (related) quantity to assess the strenght of the linear relationship between the variables involved: the coefficient R^2.

• PARTITION of the SUM OF SQUARES

The variability of $y_1,\dots,y_n$ is usually summarized using $s_y^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i-\bar{y})^2$ ← here we are not looking at the model

The numerator is also called the TOTAL sum of squares (SST): "deviance"

it is just "how difficult it is to study y"

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

Consider now the simple linear model.
We want to partition the SST into two parts:

"what the model can explain"

1. the variation that is accounted for by the model → REGRESSION sum of squares : SSR
2. the variation that is left unexplained by the model → RESIDUAL (ERROR) sum of squares : SSE

"what the model can not explain"

We use the following quantities:
- observed values y_i $i=1,\dots,n$
- predicted values $\hat{y}_i$ $i=1,\dots,n$
- residuals $e_i = y_i - \hat{y}_i$ $i=1,\dots,n$

$\to \hat{\beta}_1 = \bar{y} - \hat{\beta}_2\bar{x}$

$$\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n}(y_i-\hat{y}_i)^2 = \sum_{i=1}^{n}(y_i-\hat{\beta}_1-\hat{\beta}_2 x_i)^2 = \sum_{i=1}^{n}(y_i-\bar{y}+\hat{\beta}_2\bar{x}-\hat{\beta}_2 x_i)^2 =$$
$$= \sum_{i=1}^{n}\left[(y_i-\bar{y})^2 + \hat{\beta}_2^2(x_i-\bar{x})^2 - 2\hat{\beta}_2(y_i-\bar{y})(x_i-\bar{x})\right] =$$
$$= \sum_{i=1}^{n}(y_i-\bar{y})^2 + \hat{\beta}_2^2\sum_{i=1}^{n}(x_i-\bar{x})^2 - 2\boxed{\hat{\beta}_2\sum_{i=1}^{n}(y_i-\bar{y})(x_i-\bar{x})}$$
$$= \sum_{i=1}^{n}(y_i-\bar{y})^2 + \hat{\beta}_2^2\sum_{i=1}^{n}(x_i-\bar{x})^2 - 2\hat{\beta}_2^2\sum_{i=1}^{n}(x_i-\bar{x})^2$$
$$= \sum_{i=1}^{n}(y_i-\bar{y})^2 - \hat{\beta}_2^2\sum_{i=1}^{n}(x_i-\bar{x})^2$$

recall that $\hat{\beta}_2 = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sum_{i=1}^{n}(x_i-\bar{x})^2} \Rightarrow$

$$\hat{\beta}_2\sum_{i=1}^{n}(x_i-\bar{x})^2 = \sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})$$
$$\Rightarrow \hat{\beta}_2^2\,\Sigma(x_i-\bar{x})^2 = \boxed{\hat{\beta}_2\,\Sigma(x_i-\bar{x})(y_i-\bar{y})}$$

Now, we notice that
$$\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2 = \sum_{i=1}^{n}(\hat{\beta}_1+\hat{\beta}_2 x_i-\bar{y})^2 = \sum_{i=1}^{n}(\cancel{\bar{y}}-\hat{\beta}_2\bar{x}+\hat{\beta}_2 x_i-\cancel{\bar{y}})^2 = \sum_{i=1}^{n}\left[\hat{\beta}_2(x_i-\bar{x})\right]^2 = \hat{\beta}_2^2\sum_{i=1}^{n}(x_i-\bar{x})^2$$

Hence we obtain $\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n}(y_i-\bar{y})^2 - \sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2$ or,

$$\underbrace{\sum_{i=1}^{n}(y_i-\bar{y})^2}_{\substack{SST \\ \text{total ss.}}} = \underbrace{\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2}_{\substack{SSR \\ \text{regression ss.}}} + \underbrace{\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}_{\substack{SSE \\ \text{error/residual ss.}}}$$

How do we interpret the three quantities?
If I only have $(y_1,\dots,y_n)$ and no additional information, the best "model" (guess) I can do to predict y is its mean $\bar{y}$. Indeed in the model $Y_i = \beta_1 + \varepsilon_i$ I obtain $\tilde{\beta}_1 = \bar{y} \Rightarrow \tilde{y}_i = \bar{y}$ for all $i=1,\dots,n$.
Hence SST is the amount of variability in the data that is left unexplained in the absence of any additional information (covariates).
After I observe x, my prediction for y becomes $\hat{y}_i$. $(\hat{y}_i-\bar{y})$ is the discrepancy between what I would have predicted in the absence of covariates and what I actually predict when I have them.
Hence SSR is the additional amount of variability explained by the model compared to modeling the data only with their mean $\bar{y}$.
Finally, SSE is what is left unexplained

RECAP:

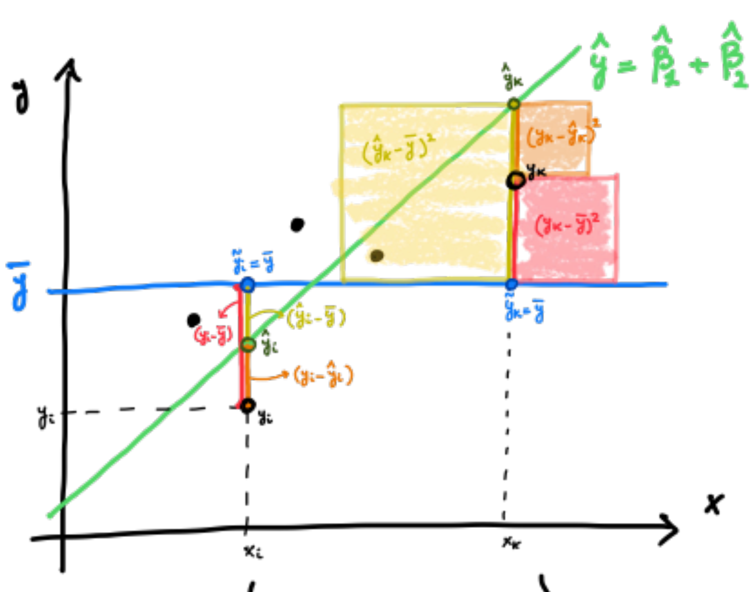
- in the absence of covariates, the model is $Y_i = \beta_1 + \varepsilon_i$ ⇒ the predicted values are $\bar{y}$ for all $i=1,\dots,n$

→ $\sum_{i=1}^{n}(y_i-\bar{y})^2$ is the total amount of variation in the data (SST)

- when I observe $x_1,\dots,x_n$, the model is $Y_i = \beta_1 + \beta_2 x_i + \varepsilon_i$ ⇒ the predicted values are $\hat{y}_i = \hat{\beta}_1 + \hat{\beta}_2 x_i$, $i=1,\dots,n$
- I still commit errors in my predictions: residuals $e_i = y_i - \hat{y}_i$

→ $\sum_{i=1}^{n}(y_i-\hat{y}_i)^2$ is the amount of variability that I can not explain : SSE
(how much the data vary around the predictions)

→ $\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2$ is the ADDITIONAL amount of variability that the model explains, compared to using $\bar{y}$
(how much the predictions vary around $\bar{y}$) : SSR



(1) compute the quantities
$(y_i-\bar{y})$
$(y_i-\hat{y}_i)$
$(\hat{y}_i-\bar{y})$
for all $i=1,\dots,n$

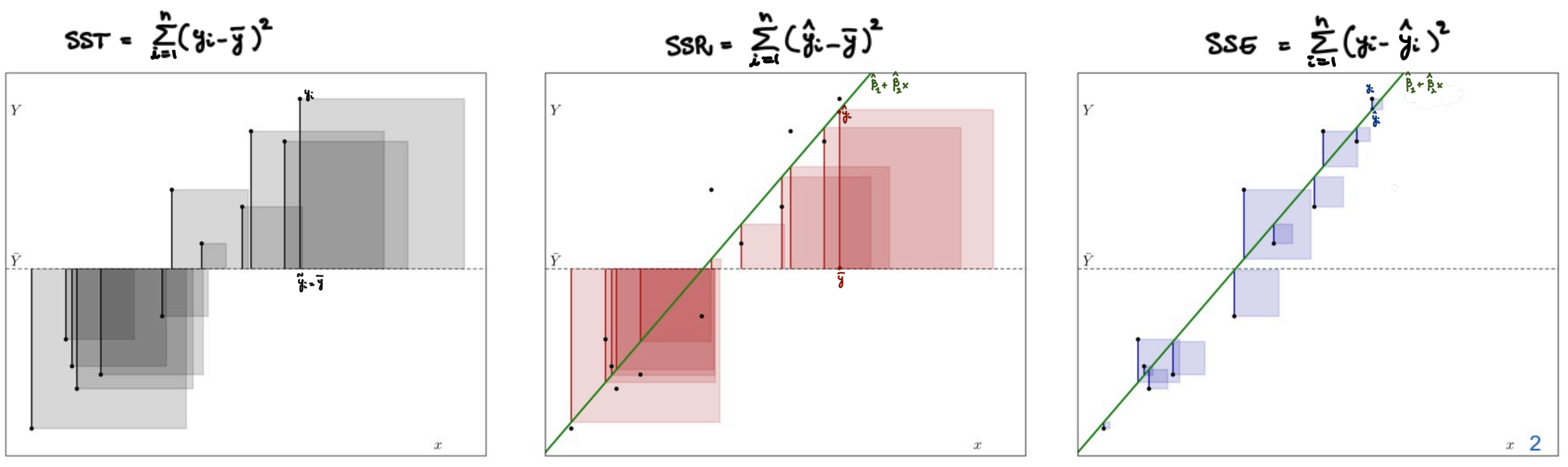
(2) compute the squares
$(y_i-\bar{y})^2$
$(y_i-\hat{y}_i)^2$
$(\hat{y}_i-\bar{y})^2$
for all $i=1,\dots,n$

(3) sum all the n points
$$\sum_{i=1}^{n}(y_i-\bar{y})^2 = \sum_{i=1}^{n}(y_i-\hat{y}_i)^2 + \sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2$$
↓
this decomposition does not hold pointwise

Nice graph found on Stack Overflow

https://stats.stackexchange.com/questions/524565/bit-confused-on-the-concept-of-deviance

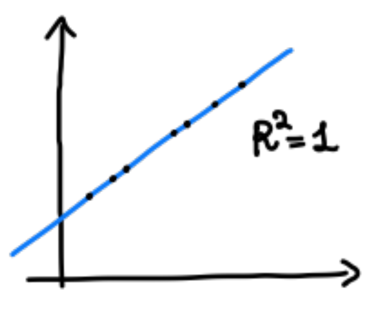
$SST = \sum_{i=1}^{n}(y_i-\bar{y})^2$ $SSR = \sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2$ $SSE = \sum_{i=1}^{n}(y_i-\hat{y}_i)^2$

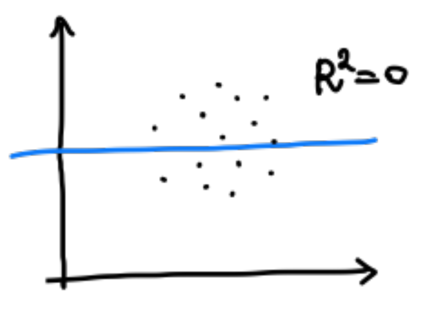


SST = total gray area (sum also the overlaps) SSR = total red area SSE = total blue area

If the model fit the data well, I expect SSR to be larger than SSE (w.r.t. the SST).
Hence I can study the ratio SSR/SST to understand how much variability is explained by the model.
The coefficient of determination R^2 ("R squared") is the proportion of variability of the dependent variable that is predicted by the covariate.

$$R^2 = \frac{SSR}{SST} = \frac{SST-SSE}{SST} = 1-\frac{SSE}{SST} \in [0,1]$$





In the case of the simple linear regression model, $R^2 = r_{xy}^2$
it measures the GOODNESS OF FIT (how adequate it is to summarise the relationship between x and y with the estimated model – i.e, a straight line)