8 oct - LEC 3

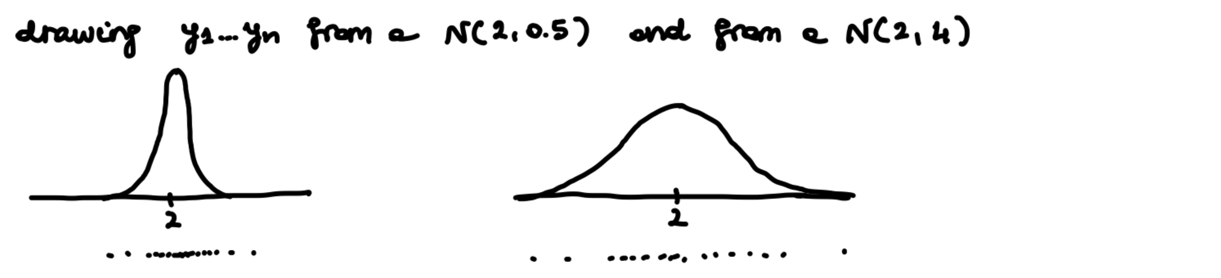
SUM OF SQUARES DECOMPOSITION (or PARTITION)

Imagine that we want to study (and make prediction) on a variable y, and we observe $(y_1, \dots, y_n)$.
In the absence of other information, the "best" way to explain the data is through the overall mean $\bar{y}$.
This corresponds to fitting the model

$$Y_i = \beta_1 + \varepsilon_i \qquad \varepsilon_i \sim N(0, \sigma^2) \qquad \text{"NULL MODEL"}$$

which gives the estimate $\tilde{\beta}_1 = \bar{y}$ ($\Rightarrow$ constant estimate)
If we predict y with this model we obtain $\tilde{y}_i = \bar{y}$ for all i.
Hence the best prediction that we can do, in the absence of additional information, is the overall mean.
Does it describe the data well? ⭐ It depends on the variability of y

⭐ example: imagine drawing $y_1 \dots y_n$ from a $N(2, 0.5)$ and from a $N(2, 4)$



In the first case, the error we commit by predicting $(y_1, \dots, y_n)$ with $\bar{y}$ is much smaller

we can look at the quantity

$$SST = \sum_{i=1}^{n} (y_i - \bar{y})^2 \qquad \text{TOTAL SUM OF SQUARES} \qquad (\text{"deviance"})$$

it tells us HOW MUCH VARIABILITY IS LEFT IN THE DATA AFTER WE SUMMARIZE THEM WITH THE OVERALL MEAN
(the "total amount of variability" of the data)

Imagine now that the additional variable $\underline{x} = (x_1, \dots, x_n)$ is introduced, and we fit a simple linear model

$$Y_i = \beta_1 + \beta_2 x_i + \varepsilon_i \qquad \varepsilon_i \sim N(0, \sigma^2)$$

With the inclusion of x, the prediction becomes

$$\hat{y}_i = \hat{\beta}_1 + \hat{\beta}_2 x_i$$

and the error that we commit is $y_i - \hat{y}_i = e_i$.
We want to understand how much the inclusion of x improves the prediction of y.

We want to partition the variability of y (SST) into two parts:

1. the ADDITIONAL VARIABILITY that is accounted for by the model
   (How much better is $\hat{y}_i$ compared to $\bar{y}$ at explaining y? or, equivalently: how useful is the linear model compared to "no model"?)
   → REGRESSION sum of squares: SSR
2. the variation that is left unexplained by the model
   → RESIDUAL (ERROR) sum of squares: SSE

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

$$\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (y_i - \hat{\beta}_1 - \hat{\beta}_2 x_i)^2 \qquad \hat{\beta}_1 = \bar{y} - \hat{\beta}_2 \bar{x}$$

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

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

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

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

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

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

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

$$\Rightarrow \hat{\beta}_2^2 \sum (x_i - \bar{x})^2 = \boxed{\hat{\beta}_2 \sum (x_i - \bar{x})(y_i - \bar{y})}$$

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,

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

$$SST = SSR + SSE$$

TOTAL SUM OF SQUARES = REGRESSION SUM OF SQUARES + RESIDUAL / ERROR SUM OF SQUARES

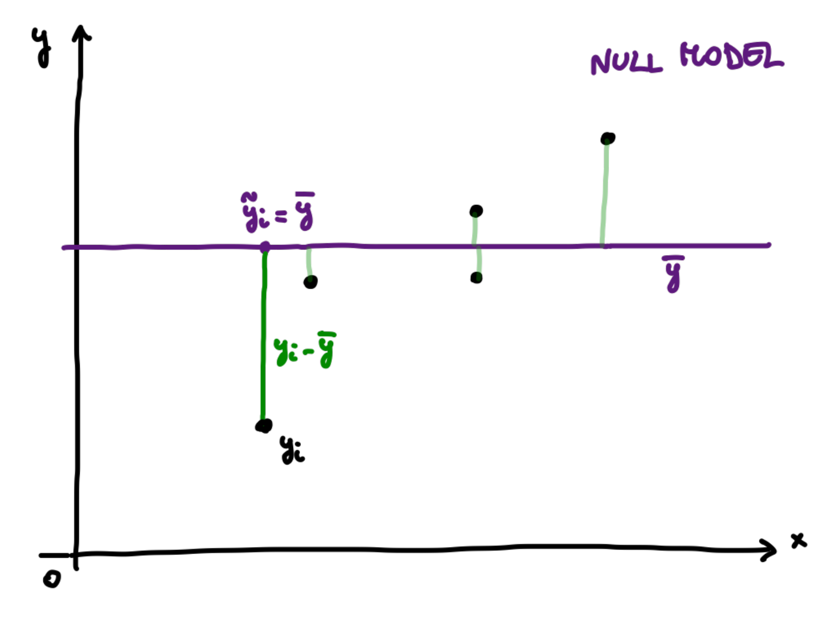
TOTAL SUM OF SQUARES: how much the data vary around the overall mean?

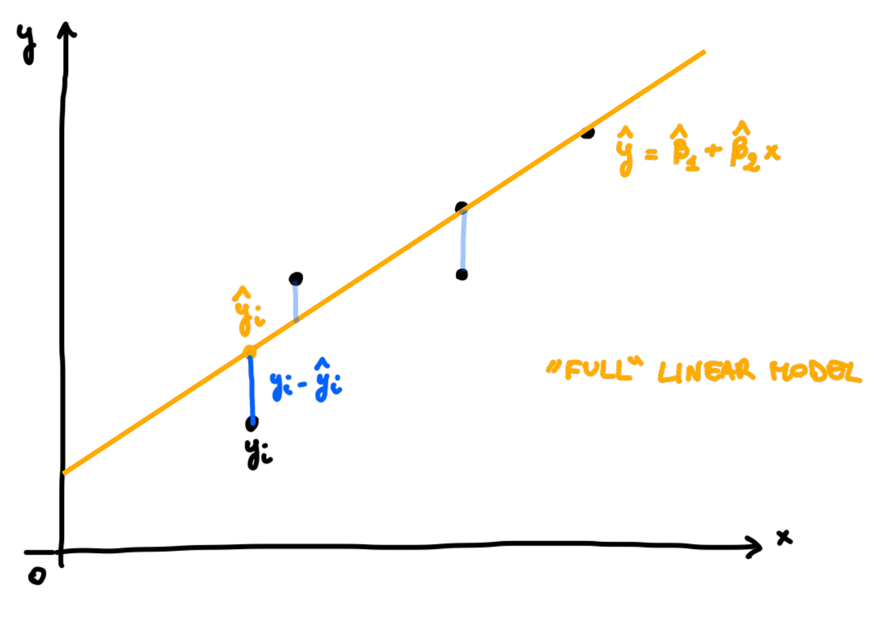
REGRESSION SUM OF SQUARES: how much the predictions vary around the overall mean?

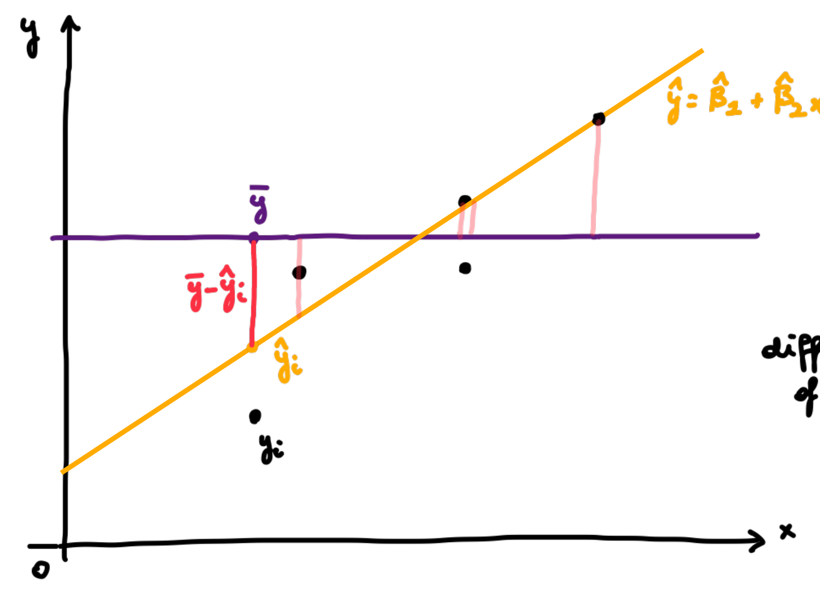
RESIDUAL / ERROR SUM OF SQUARES: how much the data vary around the predictions?

Recap:
We want to predict y.

- in the absence of covariates, the model is the NULL model $Y_i = \beta_1 + \varepsilon_i$ $\Rightarrow$ 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 y
- when I observe $x_1, \dots, x_n$, the model is $Y_i = \beta_1 + \beta_2 x_i + \varepsilon_i$ $\Rightarrow$ the predicted values are $\hat{y}_i = \hat{\beta}_1 + \hat{\beta}_2 x_i$, $i = 1, \dots, n$
  $(\hat{y}_i - \bar{y})$ is the discrepancy between what I would have predicted in the absence of covariates and what I actually predict when I have them. Hence, $\sum (\hat{y}_i - \bar{y})^2 = SSR$ is the additional amount of variability explained by the model compared to modeling the data only with their mean $\bar{y}$.
- 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





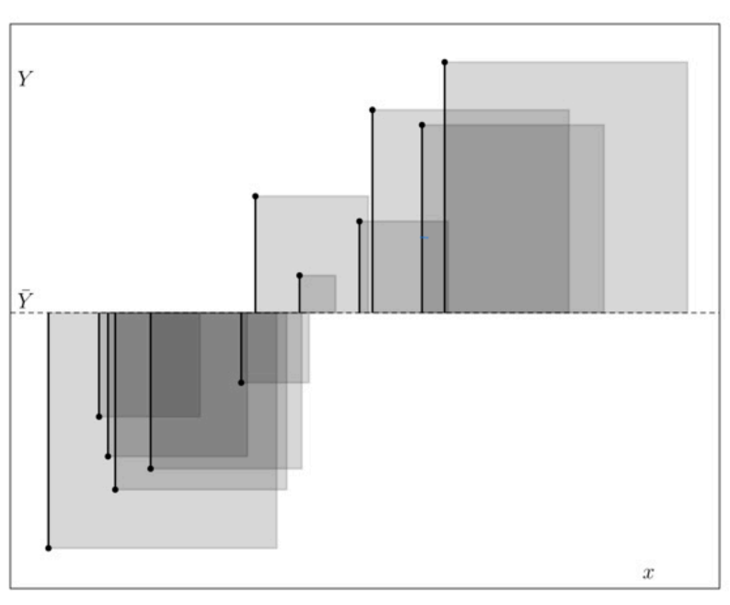


difference in the predictions of the two models

Nice graph found on Stack Overflow
https://stats.stackexchange.com/questions/524565/bit-confused-on-the-concept-of-deviance

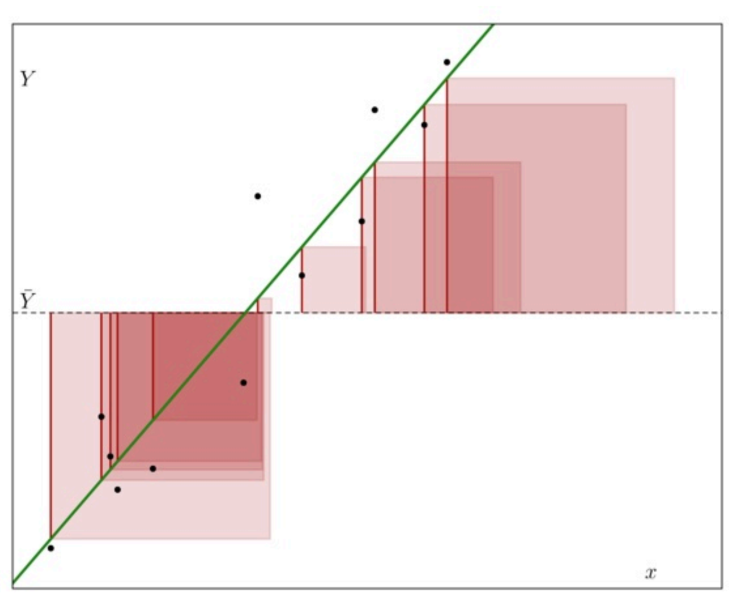
SST = SSR + SSE

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



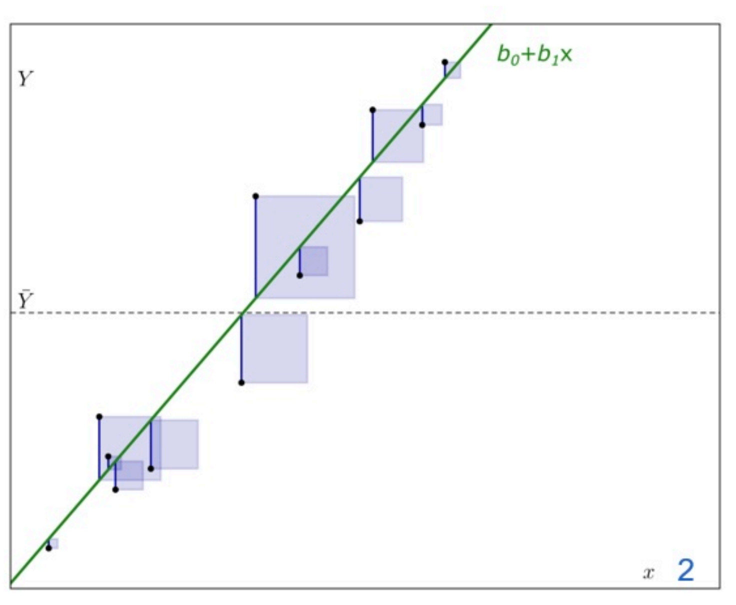
SST = total gray area (sum also the overlaps)

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



SSR = total red area

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



SSE = total blue area