

SUM OF SQUARES DECOMPOSITION (or PARTITION)

Imagine that we have a sample $(y_1, \dots, y_n)$, and we assume they are realizations of the random variables $Y_1, \dots, Y_n$.

We model the Y_i 's as independent and identically distributed following a normal distribution:

$$Y_i \sim N(\beta_1, \sigma^2) \quad \text{iid for } i=1, \dots, n$$

This corresponds to fitting the model

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

which gives the estimate $\hat{\beta}_1 = \bar{y}$

If we **predict** y with this model we obtain $\hat{y}_i = \bar{y}$ for all i .

In the absence of other information, the "best" way to explain the data is through the overall mean $\bar{y}$.

For each observation, the **prediction error** is $y_i - \hat{y}_i = y_i - \bar{y}$

Hence the best prediction that we can do, in the absence of additional information, is the overall mean.
Does it describe the data well? $\star$ It depends on the variability of y

$\star$ 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 do by predicting $(y_1, \dots, y_n)$ with $\bar{y}$ is much smaller

we can look at the quantity

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

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 $x = (x_1, \dots, x_n)$ is introduced, and we fit a simple linear model

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

With the inclusion of x , the **prediction** becomes

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

and the **prediction error** with this model is $y_i - \hat{y}_i = \varepsilon_i$

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 **variability** that is left unexplained by the model
→ **RESIDUAL (ERROR) SUM OF SQUARES: SSE**

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

$$\begin{aligned} \sum_{i=1}^n \varepsilon_i^2 &= \sum_{i=1}^n (y_i - \hat{y}_i)^2 = \sum_{i=1}^n (y_i - \hat{\beta}_1 - \hat{\beta}_2 x_i)^2 && \hat{\beta}_2 = \bar{y} - \hat{\beta}_1 \bar{x} \\ &= \sum_{i=1}^n (y_i - \bar{y} + \hat{\beta}_2 \bar{x} - \hat{\beta}_2 x_i)^2 = \sum_{i=1}^n ((y_i - \bar{y}) - \hat{\beta}_2 (x_i - \bar{x}))^2 = \\ &= \sum_{i=1}^n [(y_i - \bar{y})^2 + \hat{\beta}_2^2 (x_i - \bar{x})^2 - 2 \hat{\beta}_2 (x_i - \bar{x})(y_i - \bar{y})] = \\ &= \sum_{i=1}^n (y_i - \bar{y})^2 + \hat{\beta}_2^2 \sum_{i=1}^n (x_i - \bar{x})^2 - 2 \hat{\beta}_2 \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) && \text{recall that } \hat{\beta}_2 = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} \\ &= \sum_{i=1}^n (y_i - \bar{y})^2 + \hat{\beta}_2^2 \sum_{i=1}^n (x_i - \bar{x})^2 - 2 \hat{\beta}_2^2 \sum_{i=1}^n (x_i - \bar{x})^2 && \Rightarrow \hat{\beta}_2 \sum_{i=1}^n (x_i - \bar{x})^2 = \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) \\ &= \sum_{i=1}^n (y_i - \bar{y})^2 - \hat{\beta}_2^2 \sum_{i=1}^n (x_i - \bar{x})^2 && \Rightarrow \hat{\beta}_2^2 \sum_{i=1}^n (x_i - \bar{x})^2 = \hat{\beta}_2 \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) \end{aligned}$$

Now, we notice that

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

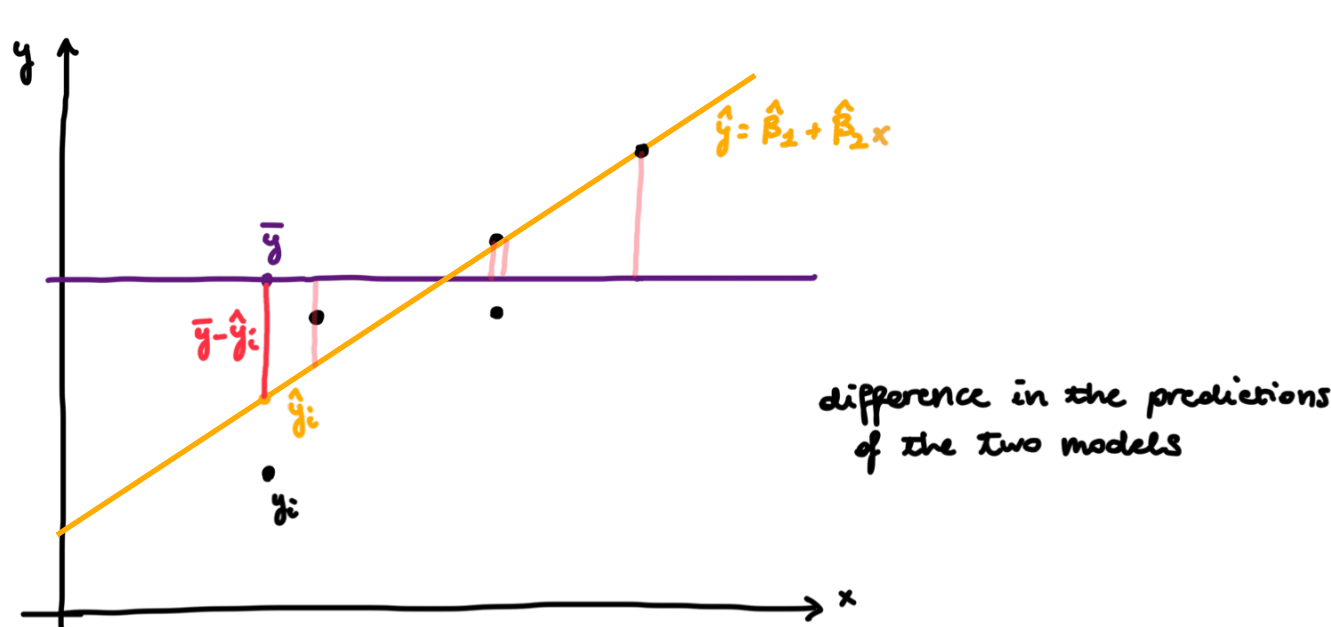
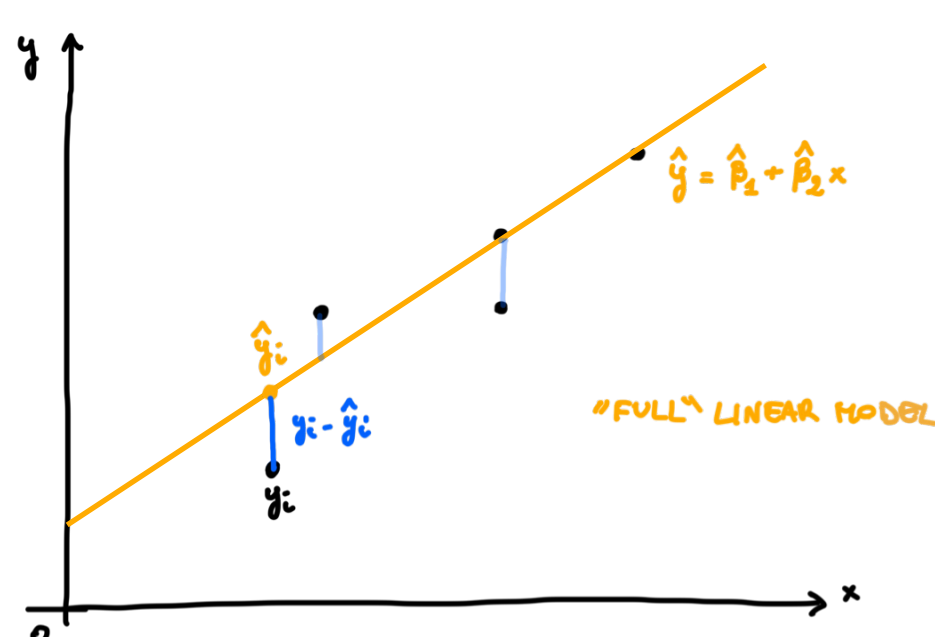
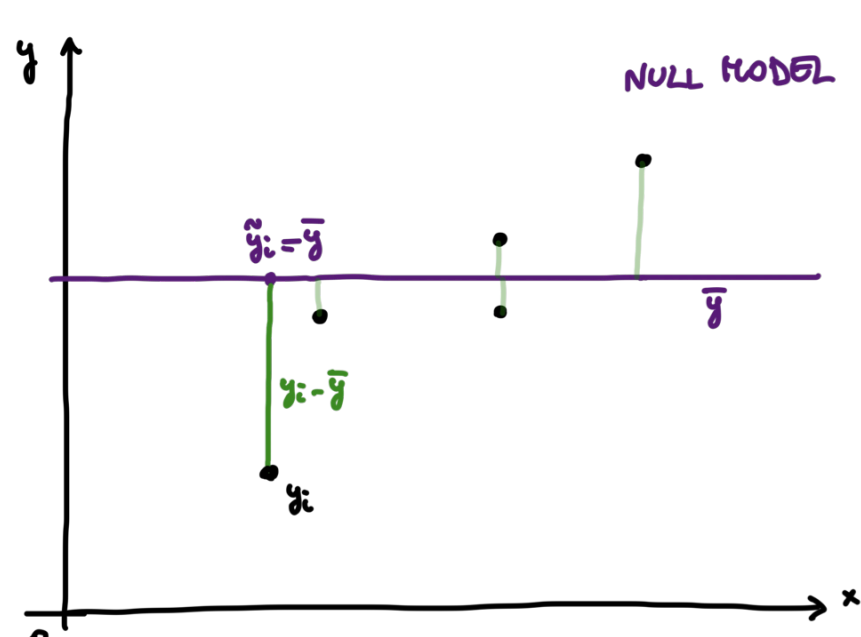
$$\begin{aligned} \text{Hence we obtain} \quad \sum_{i=1}^n \varepsilon_i^2 &= \sum_{i=1}^n (y_i - \bar{y})^2 - \sum_{i=1}^n (\hat{y}_i - \bar{y})^2 \\ \sum_{i=1}^n (y_i - \hat{y}_i)^2 &= \sum_{i=1}^n (y_i - \bar{y})^2 - \sum_{i=1}^n (\hat{y}_i - \bar{y})^2 \quad \text{or equivalently,} \end{aligned}$$

$$\begin{aligned} \sum_{i=1}^n (y_i - \bar{y})^2 &= \sum_{i=1}^n (\hat{y}_i - \bar{y})^2 + \sum_{i=1}^n (y_i - \hat{y}_i)^2 \\ \downarrow \text{SST} &= \downarrow \text{SSR} + \downarrow \text{SSE} \\ \text{TOTAL SUM OF SQUARES} &= \text{REGRESSION SUM OF SQUARES} + \text{RESIDUAL / ERROR SUM OF SQUARES} \\ \downarrow &\quad \downarrow \quad \downarrow \\ \text{how much the data vary around the overall mean?} &\quad \text{how much the predictions vary around the overall mean?} &\quad \text{how much the data vary around the predictions?} \end{aligned}$$

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, $\sum_{i=1}^n (\hat{y}_i - \bar{y})^2 = SSR$ is the additional amount of variability explained by the model compared to modeling the data only with their mean $\bar{y}$.
- I still have errors in my predictions: residuals $\varepsilon_i = y_i - \hat{y}_i \rightarrow \sum_{i=1}^n (y_i - \hat{y}_i)^2$ is the amount of variability that I can not explain



Nice graph found on Stack Overflow

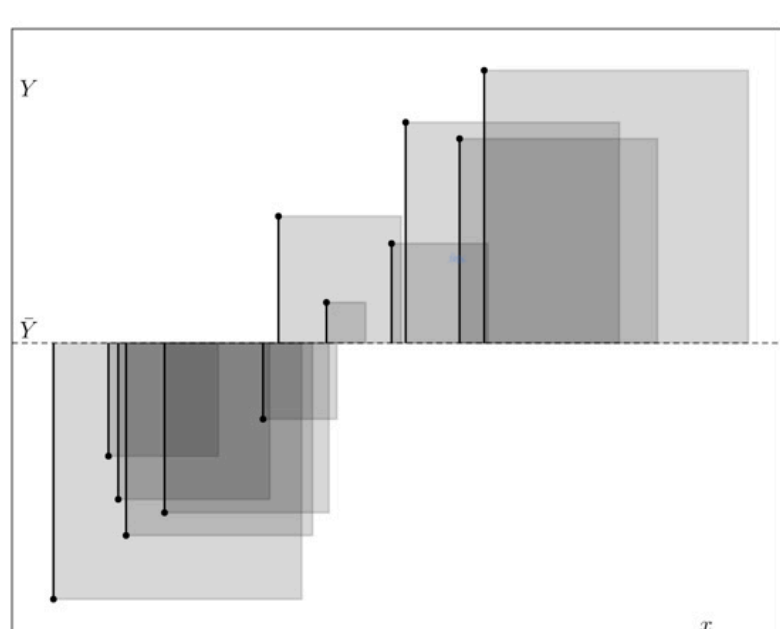
<https://stats.stackexchange.com/questions/524565/bit-confused-on-the-concept-of-deviance>

$$SST = SSR + SSE$$

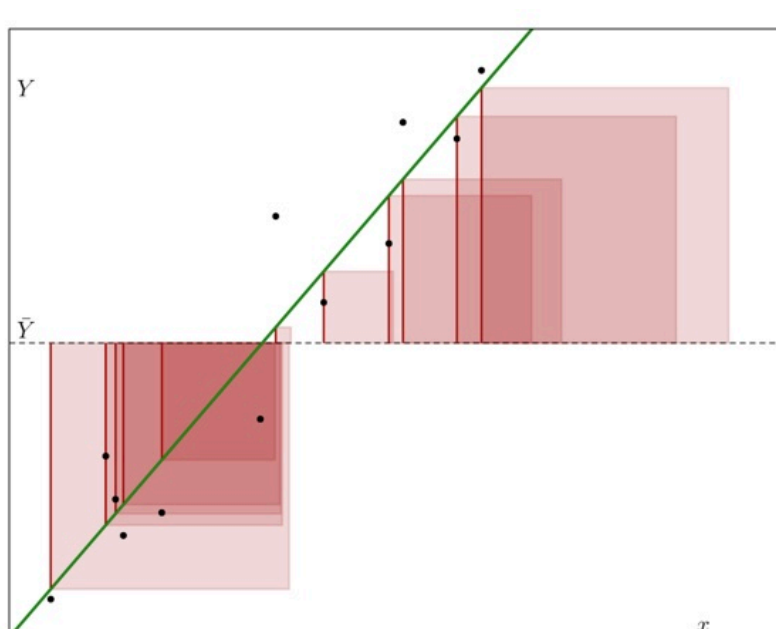
$$SST = \sum_{i=1}^n (y_i - \bar{y})^2$$

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

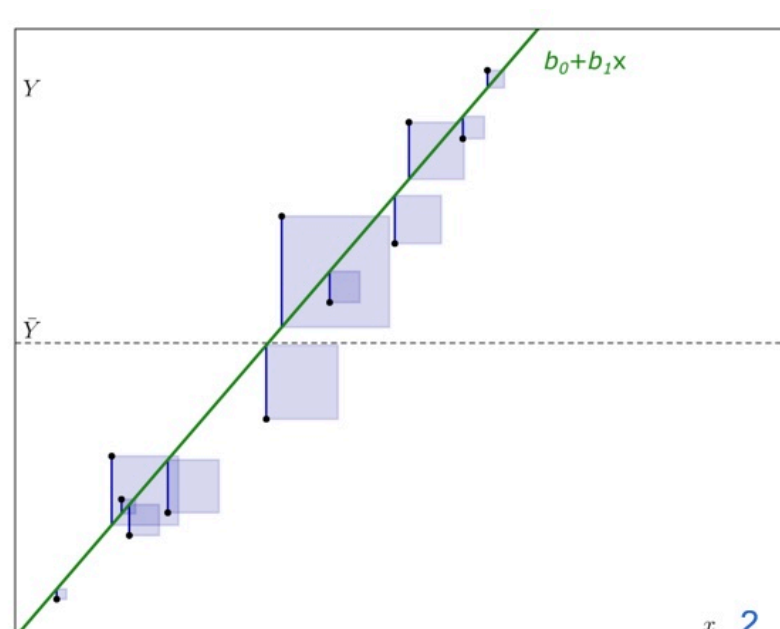
$$SSE = \sum_{i=1}^n (y_i - \hat{y}_i)^2$$



SST = total gray area (sum also the overlaps)



SSR = total red area



SSE = total blue area