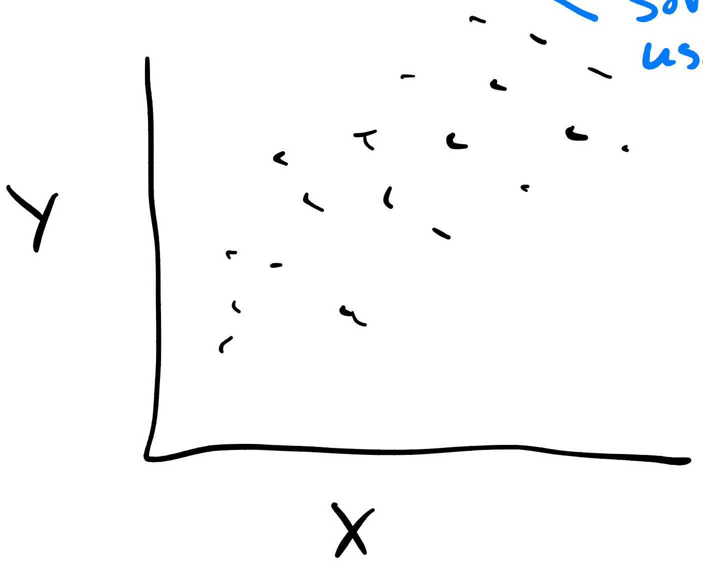
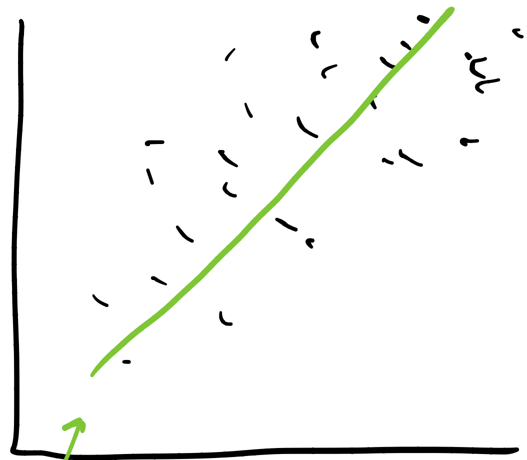
Lecture 12: Linear Regression

① $F(x_i) = b + m X_i$ ← know this — data

solve for these unknowns using scipy

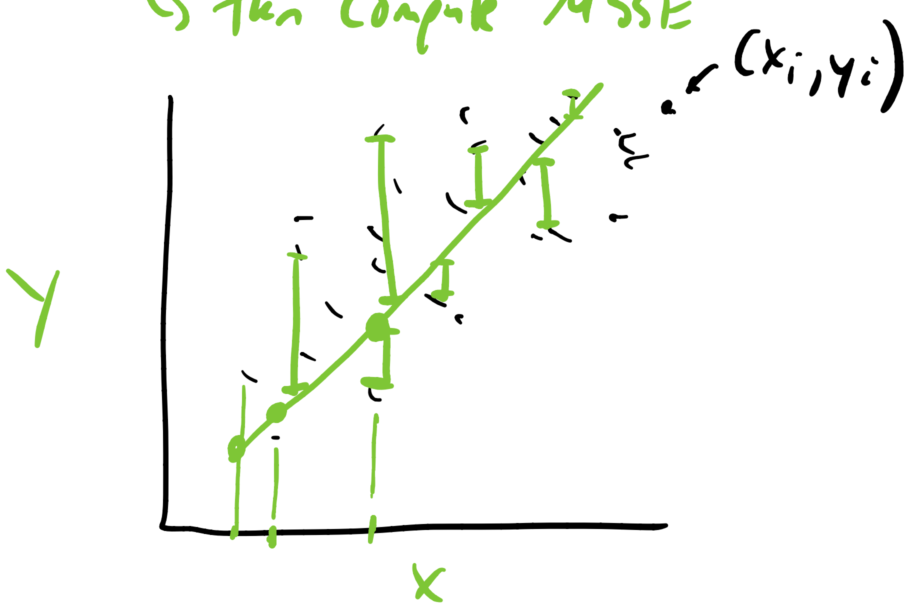




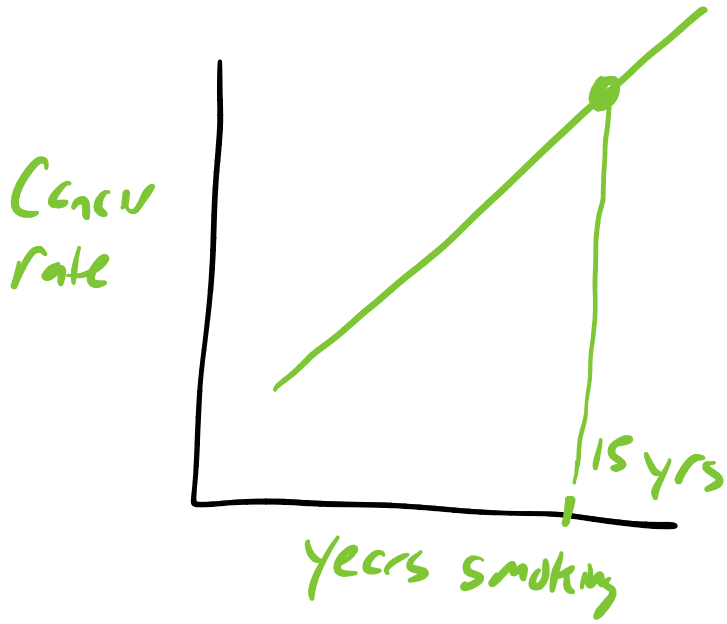
② Compute <u>MSSE</u>, regression line

<u>Option 1</u>: plug in X_i data used to find b & m into $F(X_i) = b + m X_i$ ← data used to fit regression

↳ then compute MSSE



Option 2: predict data values we've never seen before



$F(15) = b + m \cdot 15 \rightarrow$ prediction

$$\overbrace{F(x_i)}^{\vec{y}} = b + m\,\overbrace{x_i}^{\text{ind var}} \quad \Big] \text{ one sample}$$

$$\begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ \vdots \\ y_N \end{bmatrix} \qquad \qquad \underbrace{\begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ \vdots \\ x_N \end{bmatrix}}_{A}$$

$$b \cdot (1) + m X_i$$

$$\vec{Y} = A \vec{c}$$

"homogeneous coord" — col of 1s
⇒ used for the intercept b → applies it to X_i value.

$$\underbrace{\begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_N \end{bmatrix}}_{\text{shape: }(N,1)} = \underbrace{\begin{bmatrix} 1 & X_1 \\ 1 & X_2 \\ \vdots & \vdots \\ 1 & X_N \end{bmatrix}}_{A:\ (N,2)} \underbrace{\begin{bmatrix} b \\ m \end{bmatrix}}_{\vec{c}:\ (2,1)}$$

$$1 \cdot b + m X_1 = Y_1$$
$$1 \cdot b + m \cdot X_2 = Y_2$$

Rename:

$$b \longrightarrow C_0$$
$$m \longrightarrow C_1$$

$$F(x_i) = C_0 + C_1 X_i \qquad \vec{c} = \begin{bmatrix} C_0 \\ C_1 \end{bmatrix}$$

shape: $(2,1)$

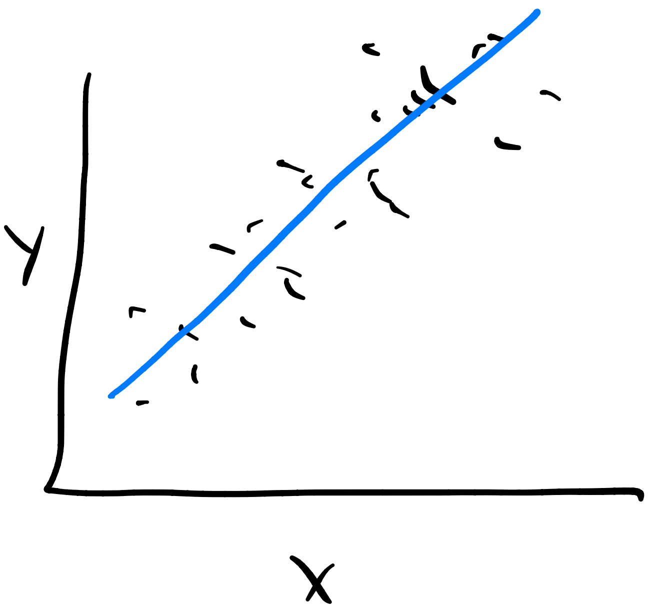
Quality of fit

Is bothering to do a regression worth the effort?

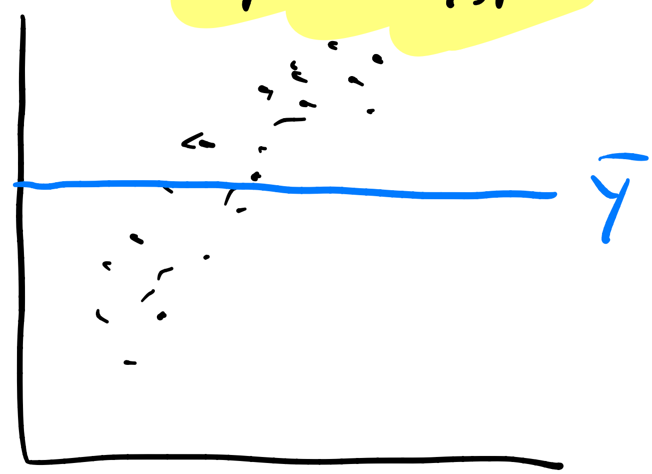
$\Rightarrow$ we compare doing regression with simply calculating the mean of the dependent var y : $\bar{y}$

$$\bar{y} = \frac{1}{N}\sum_{i=1}^{N} y_i$$

$F(x_i) = c_0 + c_1 x_i$



vs.



predicted value

predicted values from regression $\hat{y}_i = F(x_i) = \overbrace{c_0}^{b} + \overbrace{c_1}^{m} x_i$

Compare with mean of y_i values $\bar{y}$

How much of improvement $\hat{y}_i$ vs. $\bar{y}$

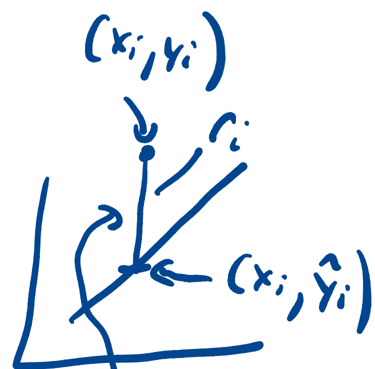
$$R^2 = \frac{\sum_{i=1}^{N}(\hat{y}_i - \bar{y})}{\sum_{i=1}^{N}(y_i - \bar{y})}$$

predicted vals from regression compared to mean

discrepancy of y_i values compared to their mean.

also written:

$$R^2 = 1 - \frac{\overbrace{\|r\|_2^2}^{\text{sum-of-squared errors (SSE)}}}{\sum_{i=1}^{N}(y_i - \bar{y})}$$



residuals

$\|r\|_2$ = distance

$$= \sqrt{r_1^2 + r_2^2 + r_3^2 + \dots + r_N^2}$$

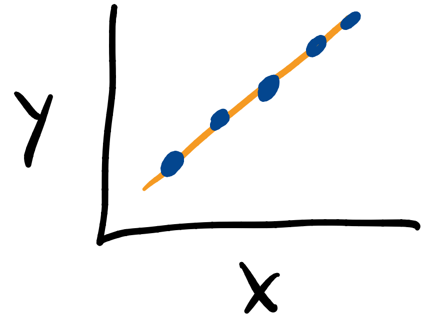
error

$$\|r\|_2^2 = r_1^2 + r_2^2 + r_3^2 + \dots + r_N^2$$

Sum of Squares error

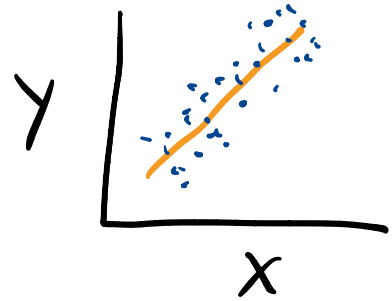
$\boxed{SSE}$

1) $\underbrace{\overbrace{\|r\|_2^2}^{\text{SSE}} \approx 0}_{\text{No error!}}$, $R^2 = 1$



2) $\overbrace{\|r\|_2^2}^{\text{error is small}}$ is much less than $\sum_{i=1}^{N} (y_i - \bar{y})$

$R^2 \approx 1$



3) $\|r\|_2^2 \approx \sum_{i=1}^{N} (y_i - \bar{y})$

$\underbrace{R^2 \approx 0}_{\text{regression not useful / worth it}}$