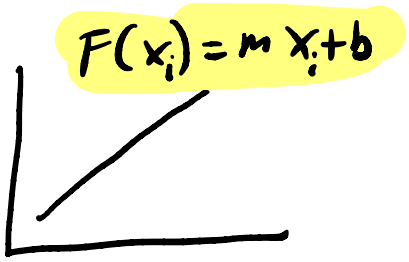
$$m = \frac{(\Sigma x_i)(\Sigma y_i) - N\Sigma x_i y_i}{(\Sigma x_i)^2 - N\Sigma x_i^2}$$

$$b = \frac{(\Sigma x_i)(\Sigma x_i y_i) - (\Sigma x_i^2)(\Sigma y_i)}{(\Sigma x_i)^2 - N\Sigma x_i^2}$$



Now that we know m & b that minimize the least squares errors — let's write the linear regression equation in matrix form:

Scaler Equation

$$F(x_i) = \underbrace{b + m x_i}_{\text{dot product between data \& coefficients } m \& b} = (1, x_i) \cdot (b, m)$$

dot product

data matrix A:

For intercept ⇒ like homogeneous coord!!

$$\underbrace{\begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ 1 & x_3 \\ \vdots & \vdots \\ 1 & x_N \end{bmatrix}}_{\text{independent variable}} \quad (N, 2)$$

Slope & intercept parameters:

$$\underbrace{\begin{bmatrix} b \\ m \end{bmatrix}}_{\vec{C}} \quad (2, 1)$$

dependent var vector

$$\underbrace{\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix}}_{\vec{Y}} \quad (N, 1)$$

$\Rightarrow \quad A\vec{c} = \vec{y}$

To set us up for generality later, let's rename

intercept $b \rightarrow c_0$

slope $m \longrightarrow c_1$

Our $\vec{c}$ becomes: $\vec{c} = \begin{bmatrix} c_0 \\ c_1 \end{bmatrix}$

and the linear regression model equation becomes:

$$F(x_i) = c_0 + c_1 x_i$$

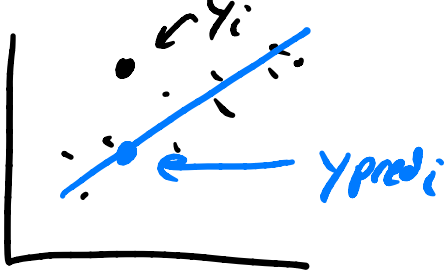
Regression work flow

- ok so now what? — we have m, b.

★ This gives us a predictive model of the data:

Use A & fitted $\vec{c}$ to compute $\vec{y}_{pred}$, the predicted y values:

$$A\vec{c} = \vec{y}_{pred}$$



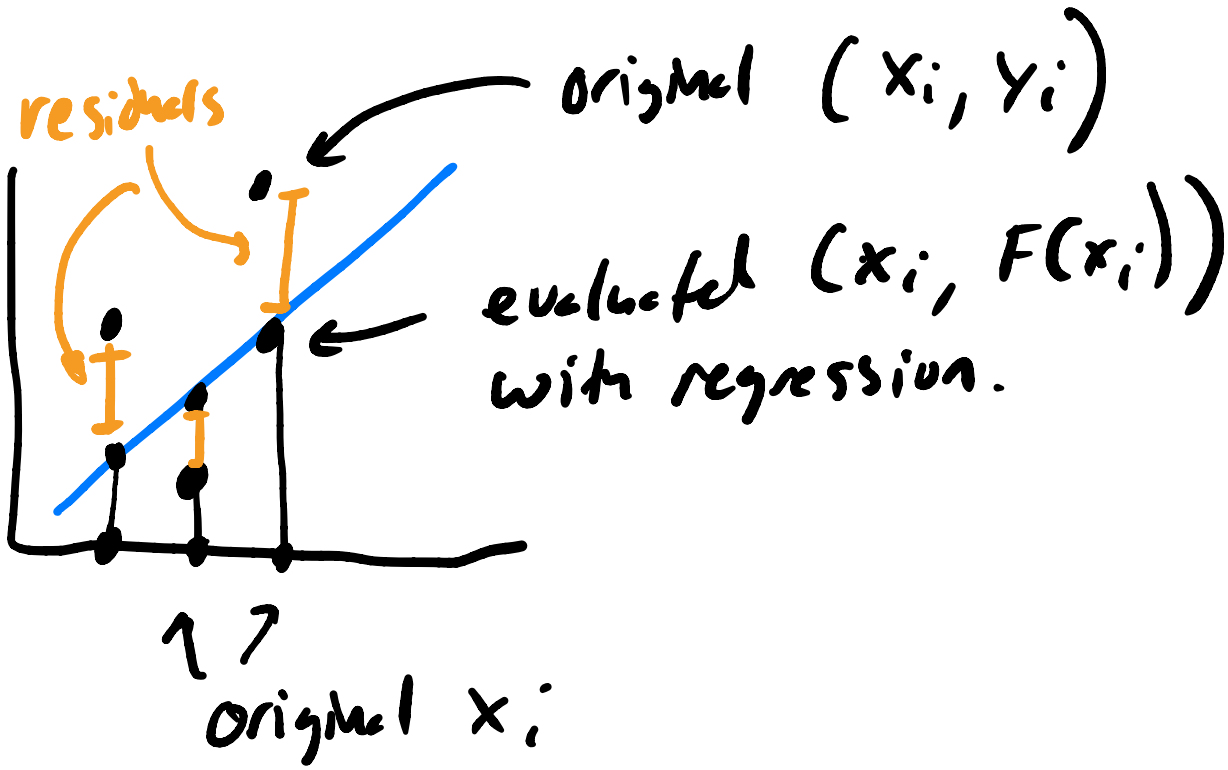
Options

1) plug in Same independent variable x_i values used to fit regression coefficients (m, b)

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

A in Same as used to solve for $\vec{c}$.

- can use to assess fit error (e.g. MSSE)

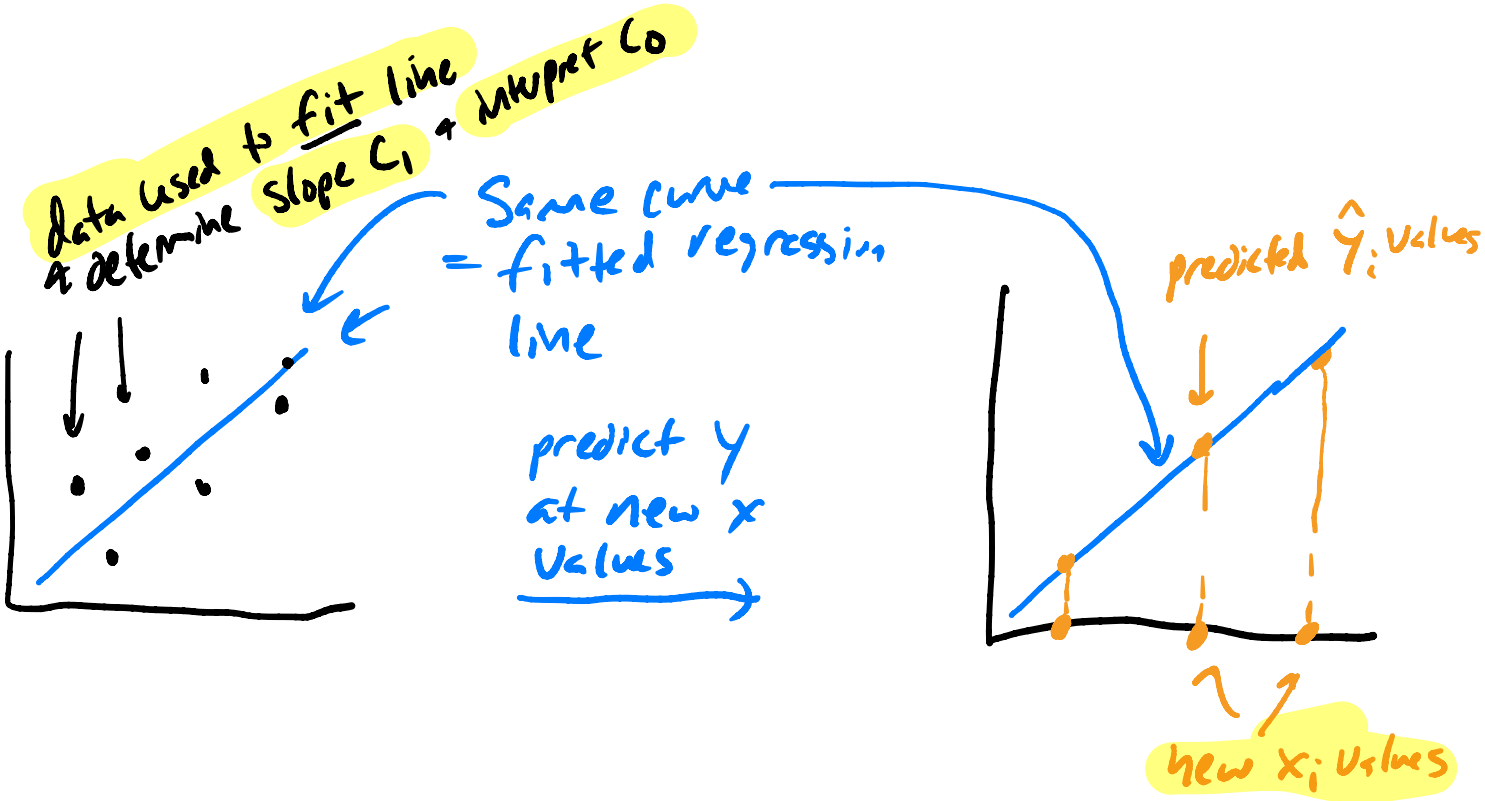


2) Plug in <u>different</u> ind var x_i values — A_{new} — than those used to solve for $\vec{c}$.

⇒ use $\vec{c}$ to get predictions — $\hat{y}_i$

$$A_{new}\,\vec{c} = \vec{y}_{pred}$$

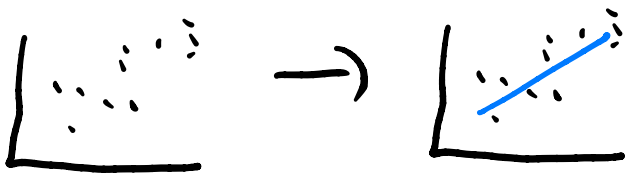
★ Tests regression generalization to <u>new data</u>.



Think of regression as a 2-step workflow:

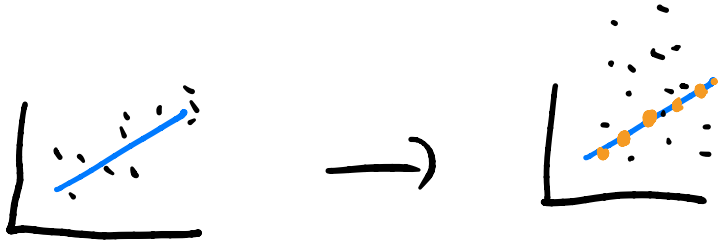
1) **Fit** regression model to data ⇒ get **model coefficients** (e.g. m, b)

★ Solve $A\vec{c} = y$ for $\vec{c}$.
ie [m, b]



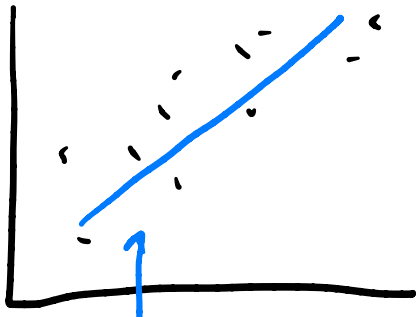
2) Use fitted coefficients to **predict** Y value of new data [not in original data set used to fit coefficients] — how well does regression generalize to new data?

★ Solve $A_{new}\vec{c} = \vec{y}_{pred}$ for $\vec{y}_{pred}$

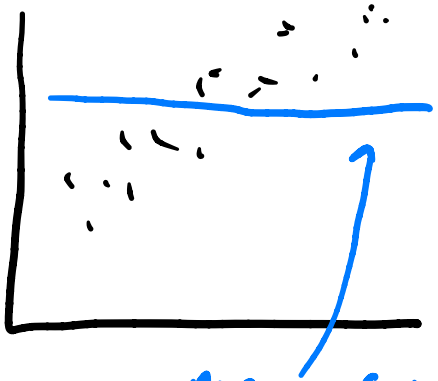


4 Quality of fit

- Is all our work worth it? Is the regression a better explanation of the data than the simple mean?



vs.



labour ions to compute!

Mean of y variable

Easy to compute!

- Let $\hat{y}_i$ be the predicted y-value by the regression model:

$$\hat{y}_i = C_0 + C x_i$$

- Let $\bar{y}$ be the $\vec{y}$ data mean (of the dependent variable)

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

regression model predictions $\hat{y}$

$\bar{y}$ mean

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

$\hat{y} = y$ value on fitted line

We define:

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

numerator: error of model over mean

denominator: error of data over mean

ie how much is our linear model buying us?

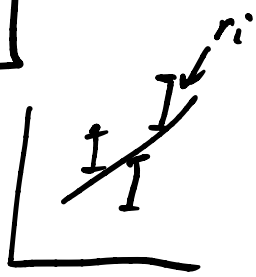
Can be written in terms of Sum-of-Squares error (SSE)

SSE = Sum of Squared error

$$R^2 = 1 - \frac{\|\vec{r}\|_2^2}{\sum (y_i - \bar{y})^2}$$

where $\vec{r} = (r_1, r_2, r_3, \ldots, r_N)$

residual errors



$\|\vec{r}\|_2$ = Euclidean distance

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

squared

$$\|\vec{r}\|_2^2 = r_1^2 + r_2^2 + r_3^2 + \ldots + r_N^2$$

↑ Sum of Squared errors (SSE)

— like MSSE — but total

(not mean) error

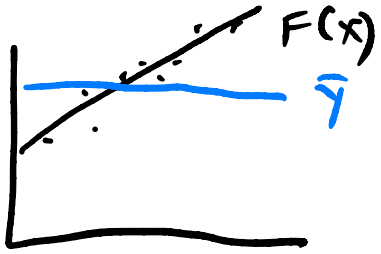
- If $\|r\|_2^2 \approx 0$, $R^2 = 1$ [line passes thru all pts]

  ($\|r\|_2^2$: Error)

- If $\|r\|_2^2 << \Sigma(y_i - \bar{y})^2$

  $R^2 \approx 1$.

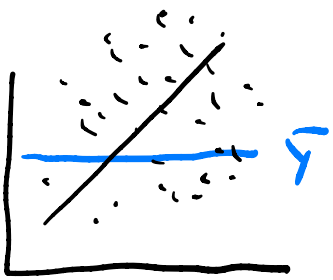
  Then big improvement over mean



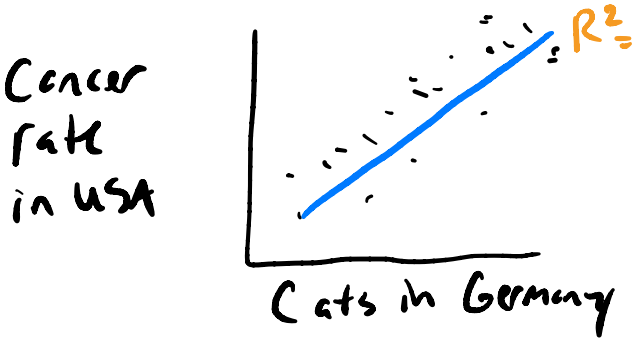
- If $\|r\|_2^2 \approx \Sigma(y_i - \bar{y})^2$

  $R^2 \simeq 0$.

  Then fit no better than mean



Important Just b/c $R^2 \approx 1$, does not mean independent Variable x caused dependent Variable y



Just b/c the Cat Population in Germany is correlated with the Cancer rate in USA does not mean the Cats caused Cancer far away!

⇒ Can be a coincidence.