Project 3: Linear Regression

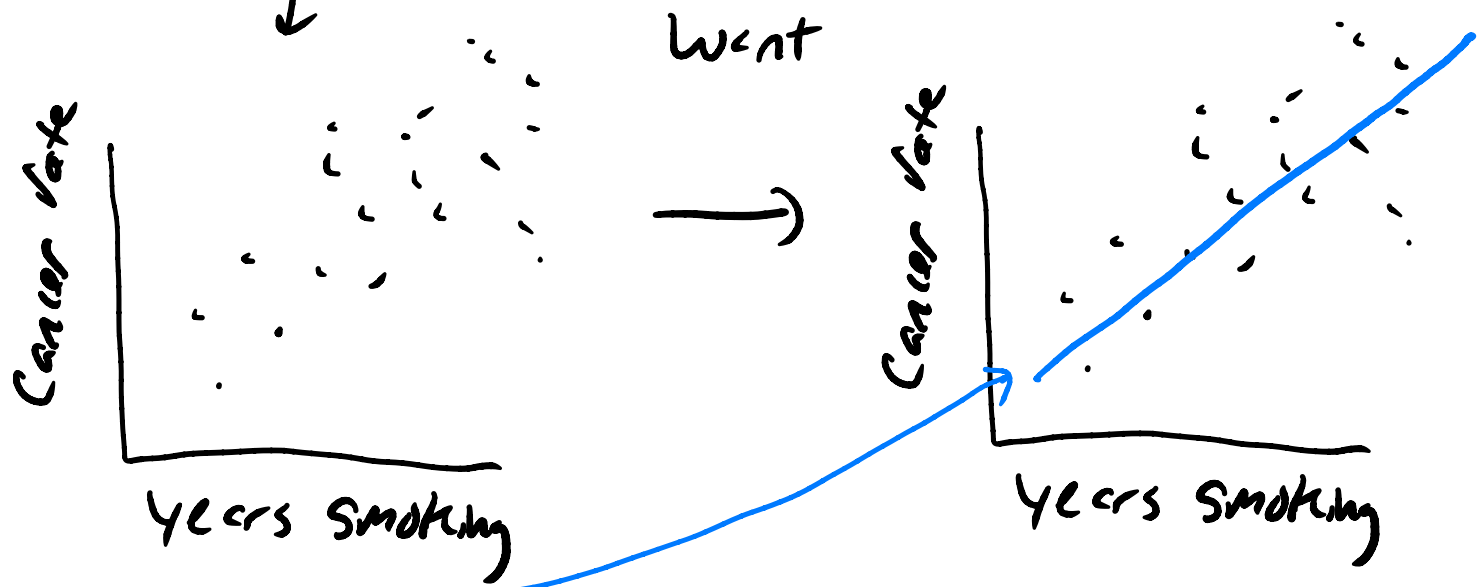
Goal Beyond just "looking" at data in a scatterplot, we want to quantify relationships between 2+ variables and the strength of association.

is Cancer associated with
- smoking?
- drinking?
- old age?
- ⋮

=

- weak?
- Strong?
- unrelated?

raw data ↓



regression: curve fit to data that tolerates error in measurement and noise

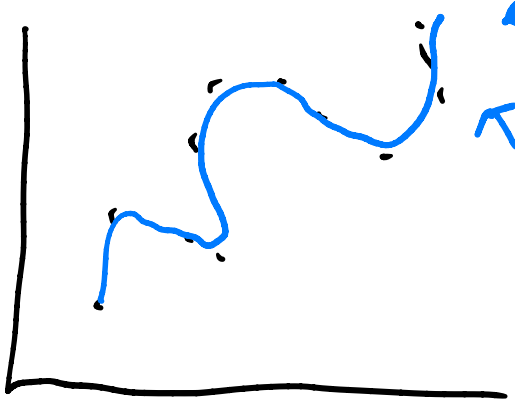
e.g. if cancer data comes from self-report survey, people

- make mistakes in filling out form
- misremember
- guess (b/c they forgot)

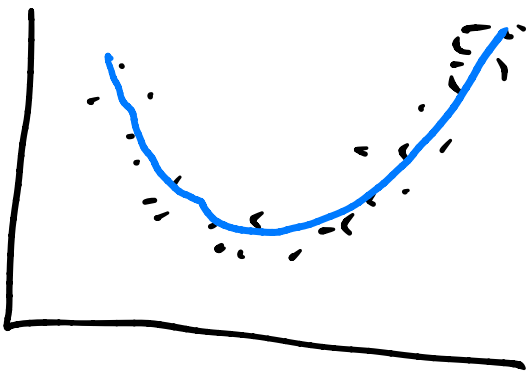
...

☆ Try to get as close as possible to all samples with regression curve.

- Does not need to pass thru all data samples



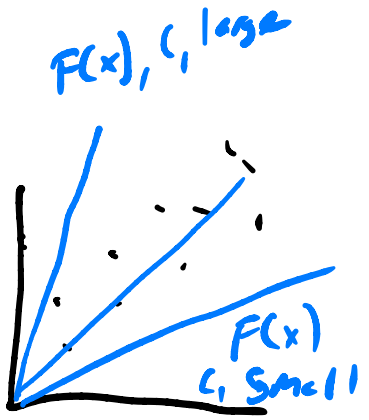
- Does not need to be a line



Fitting a Straight line to data

Let's start simple to gain Intuition:

- Regression between 2 variables:
  - 1 independent variable (x)
  - 1 dependent variable (y)

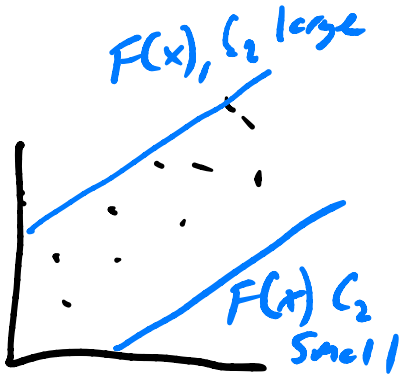


- Linear Fit function:

$$y_i = F(x_i) = C_0 + C_1 x_i$$

C_0: unknown intercept of regression line.

C_1: unknown slope of regression line

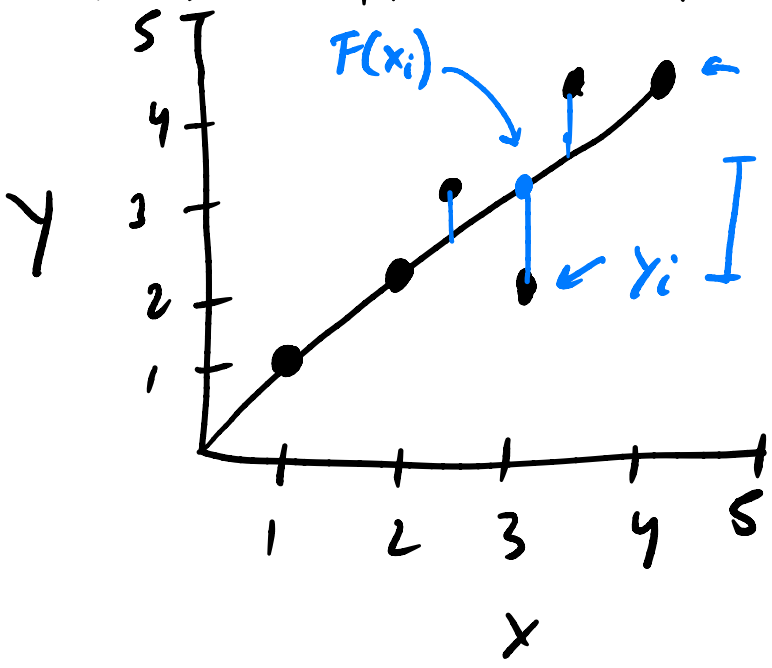


Let's rename $C_1 = m$, $C_0 = b$

$$\Rightarrow \quad y_i = F(x_i) = m x_i + b$$

Goal Find best m & b values to fit data.

- Error in regression fit to data → vertical distance between line fit at each x_i : $F(x_i)$ and the actual data y_i :



$F(x) = mx + b$ w/ specific m, b values.

$r_i = \text{true} - \text{fit}$

$r_i = y_i - F(x_i)$

$= y_i - (m x_i + b)$

r_i called the residual of the fit.

problem: Can't just sum r_i to get total error in fit —— why??

Some $r_i \geq 0$, Some $r_i \leq 0$

⇒ error would cancel!

Sum = 0, but there is error!

possible work arounds:

Absolute distance

$$\sum_i |r_i|$$

Squared distance

$$\sum_i r_i^2$$

- Squared dist usually selected b/c Math easier to compute best m, b.

- With Squared dist, linear regression goal is to find coefficients m, b that minimize the average error over all data pts
  - ↳ Called mean sum of squares error (MSSE)

$$MSSE = \rho = \frac{1}{N}\sum_{i=1}^{N} r_i^2 = \frac{1}{N}\sum_{i=1}^{N}\left(y_i - F(x_i)\right)^2$$

- Minimize MSSE equation with respect to m, b — take partial derivatives set to 0, solve!
- Approach = least squares

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

$F(x_i) = m x_i + b$

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

$$\Rightarrow \begin{bmatrix} \Sigma x_i^2 & \Sigma x_i \\ \Sigma x_i & N \end{bmatrix} \begin{bmatrix} m \\ b \end{bmatrix} = \begin{bmatrix} \Sigma x_i y_i \\ \Sigma y_i \end{bmatrix}$$

x_i independent var $\rightarrow A$ | $\vec{c}$ ← unknown coefficients | $\vec{y}$ ← dependent variable

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

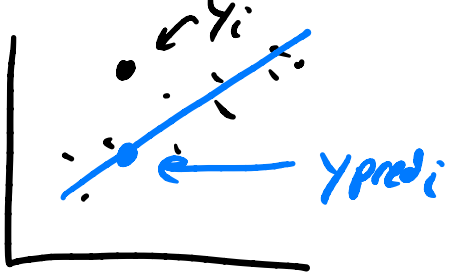
Regression work flow

- 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:

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



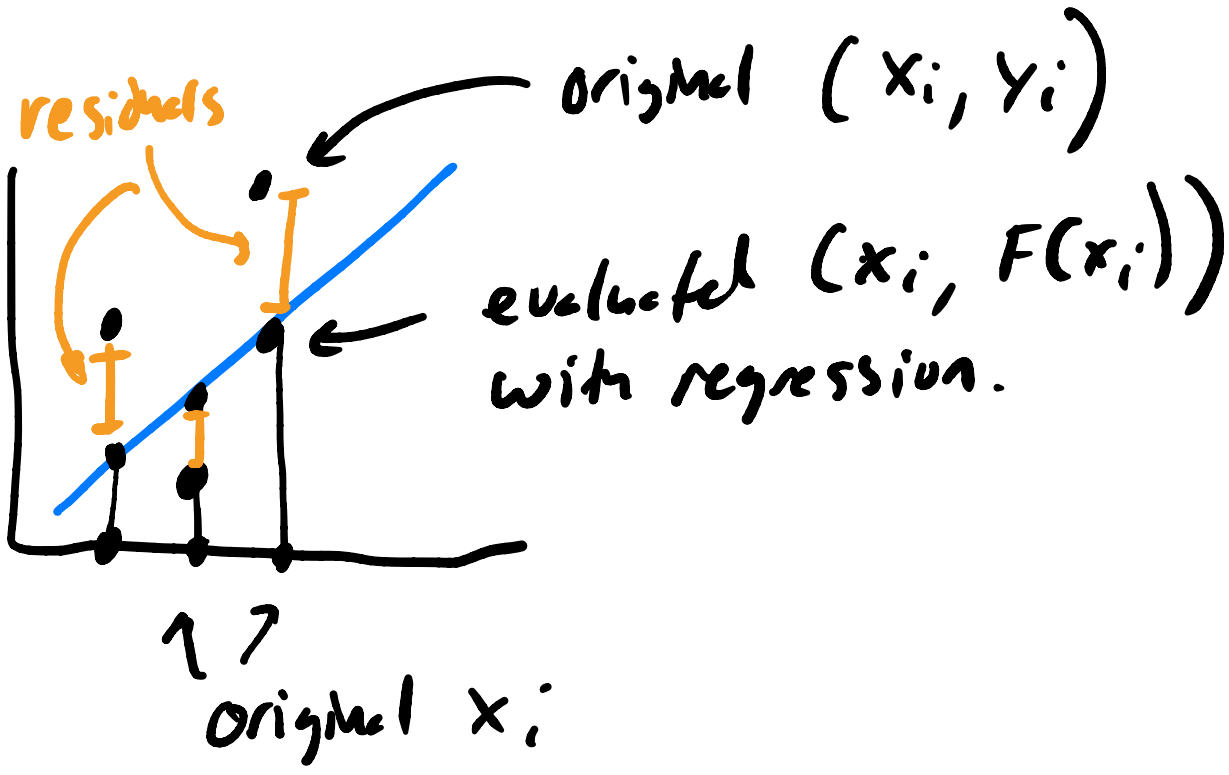
Options

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 different ind var x_i values — A_{new} — than those used to solve for $\vec{c}$.

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

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

★ Tests regression generalization to new data.