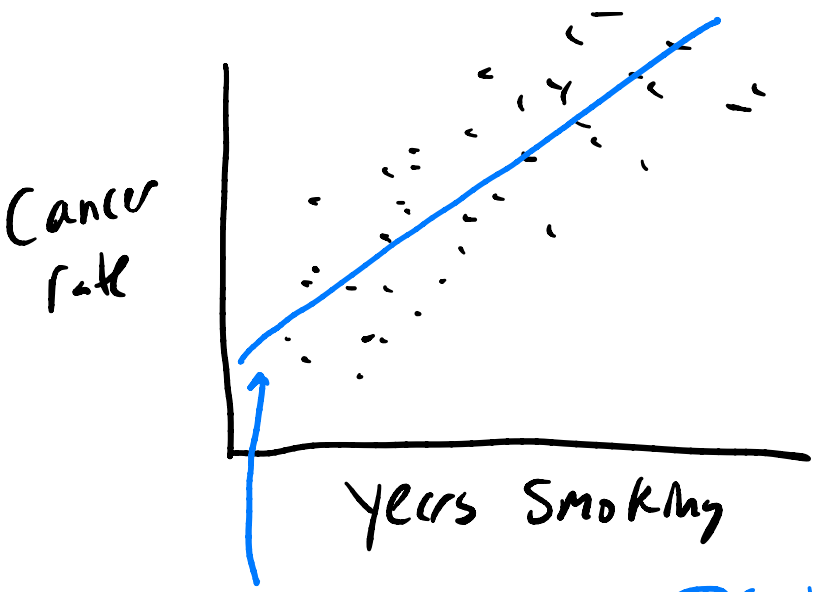
Lecture 11: Linear regression

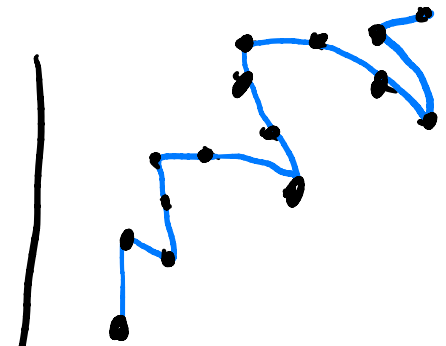


linear regression: quantify the strength of association between any 2 vars.

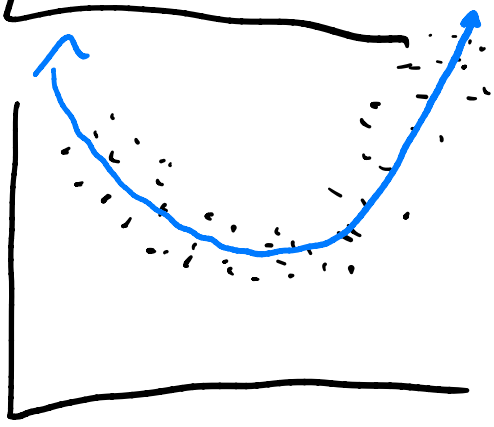
regression curve: Trying to get as close as possible to all the data samples.

← genetic protection from cancer

tolerates/good with noise, measurement error



← not doing this



← possible to fit non-linear curves to data

Fitting a Straight line to 2 Vars

Linear regression w/ 2 Vars: **Simple linear regression**

1) Independent Variable (x)

2) Dependent Var (y)

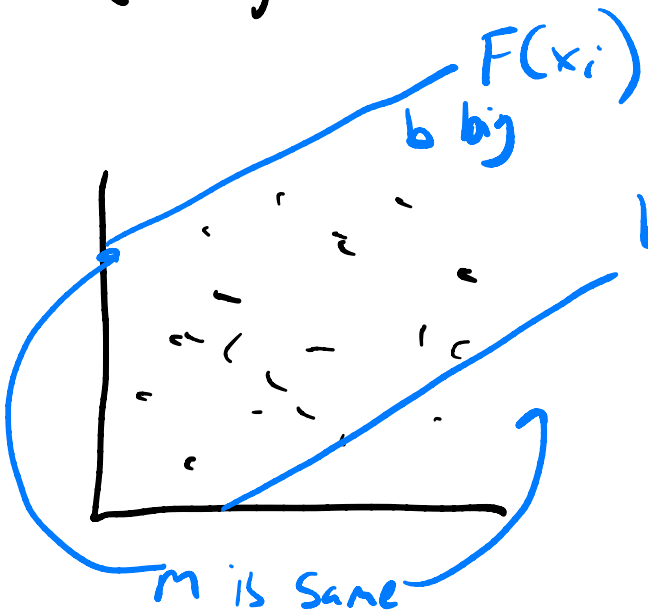
Liner regression equation:

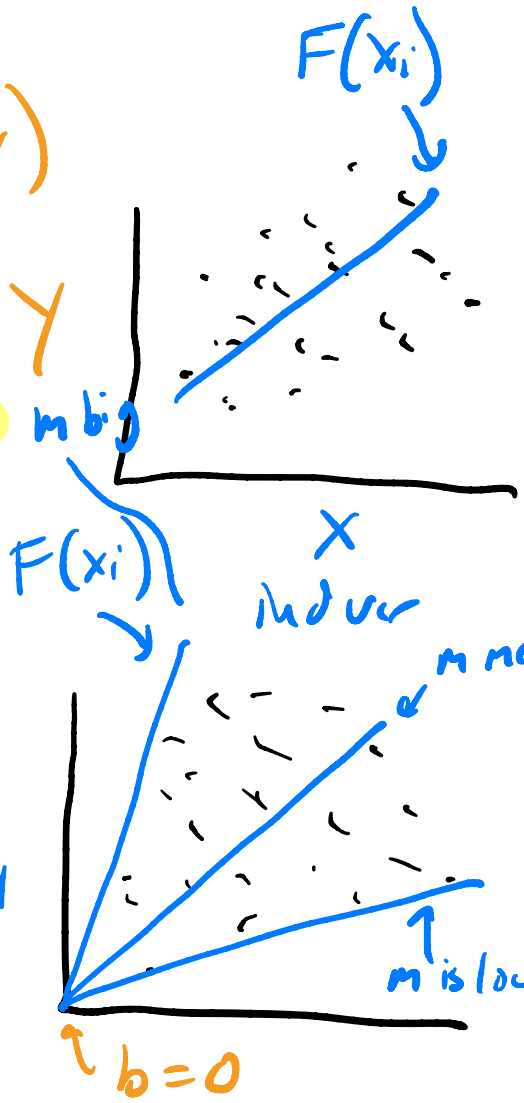
$$Y_i = F(X_i) = m X_i + b$$

dep Var: Y

slope: m

y-intercept: b

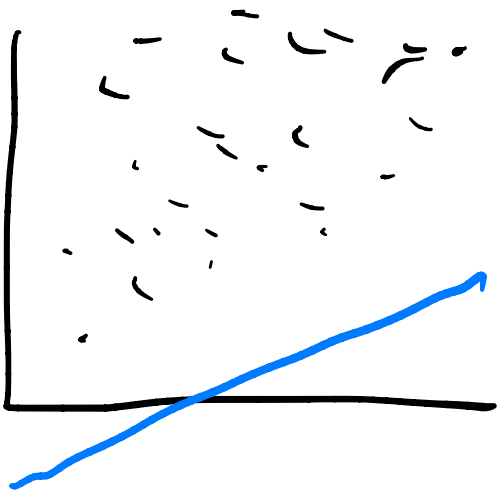




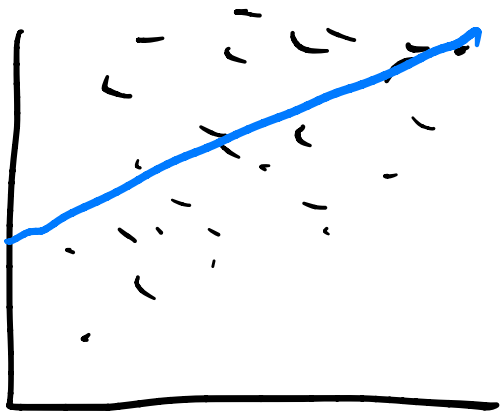
Goal of regression Find "best" values for m and b to fit data

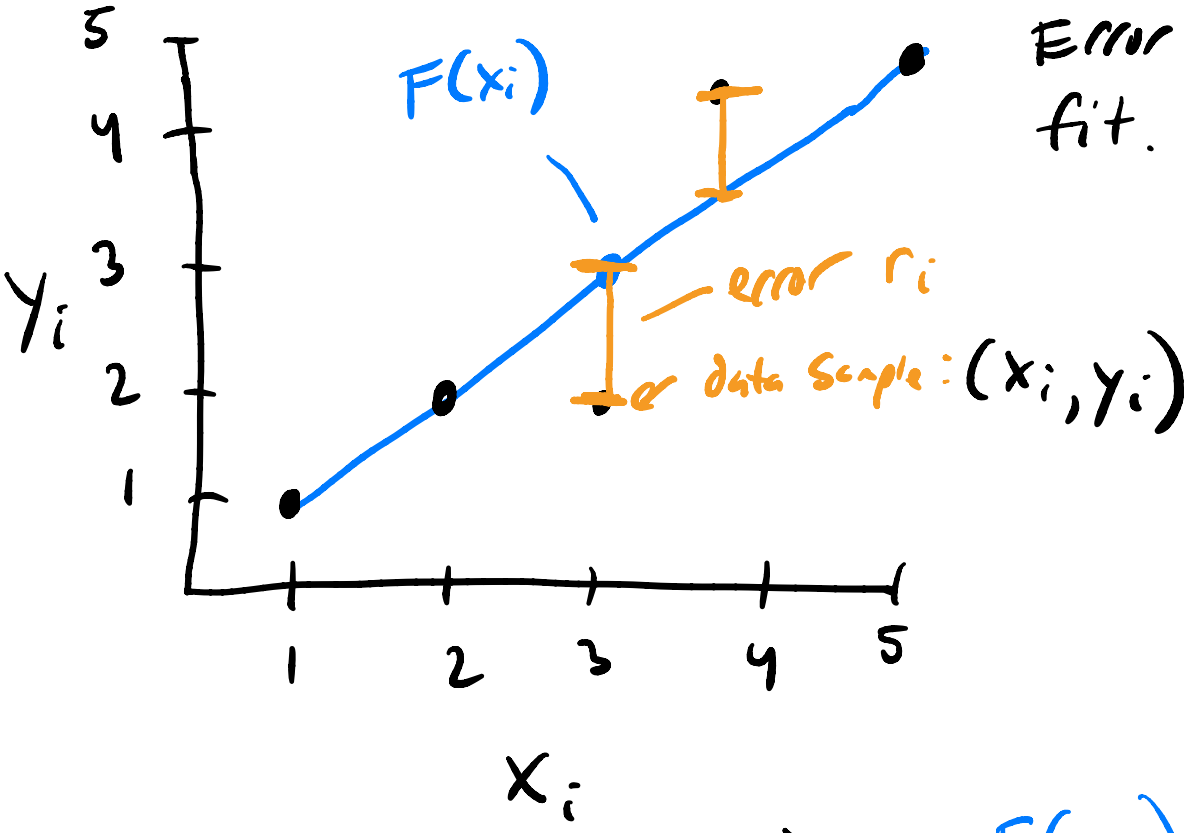
Example: Fix m. & Find best b

bad b



good b





Error in regression fit.

r_i = true value − fit value = $y_i - F(x_i)$

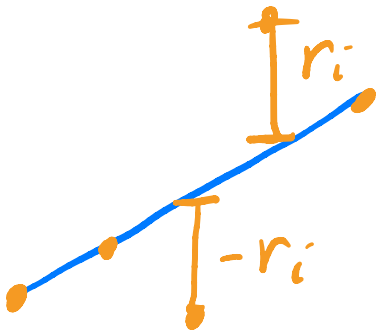
$= y_i - [mx_i + b]$ = error

r_i : residual error for one sample i

Want: Overall error across all data samples.

Total error → Sum up all r_i values.

average them — error on average across all samples.



problem:

total error = $r_i - r_i = 0$

Fixes:

1) Take absolute values

$$\text{total error} = \sum_{i=1}^{N} |r_i|$$

More common

2) Squared error:

$$\text{error} = \sum_{i=1}^{N} r_i^2$$

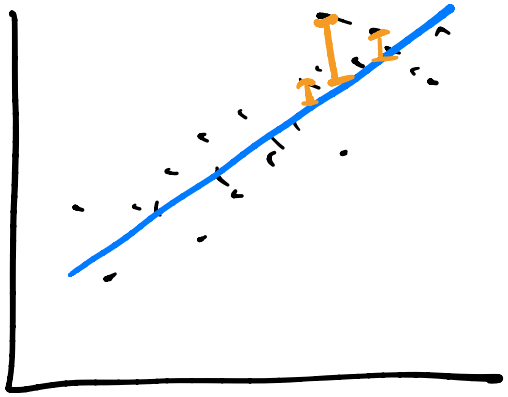
total average error = Mean Sum of Squared error

MSSE.

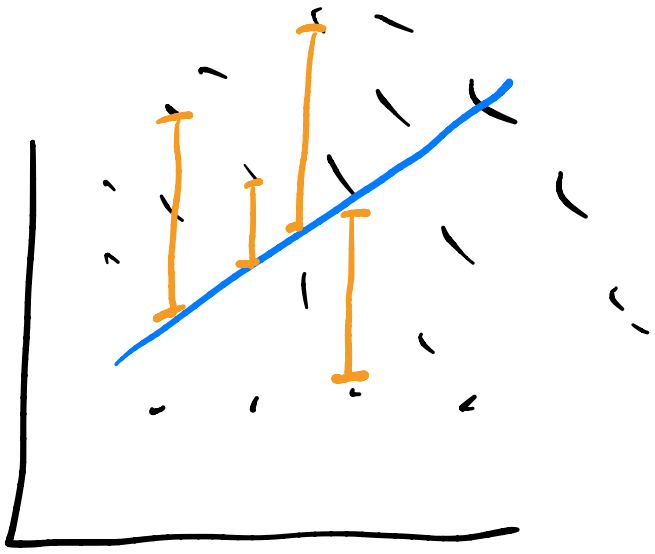
$$MSSE = \frac{1}{N}\sum_{i=1}^{N} r_i^2 = \frac{1}{N}\sum_{i=1}^{N}\left[y_i - (mx_i + b)\right]^2$$

"Least Squares" of error r_i

MSSE is Small



MSSE is big



1) Set-up data appropriately

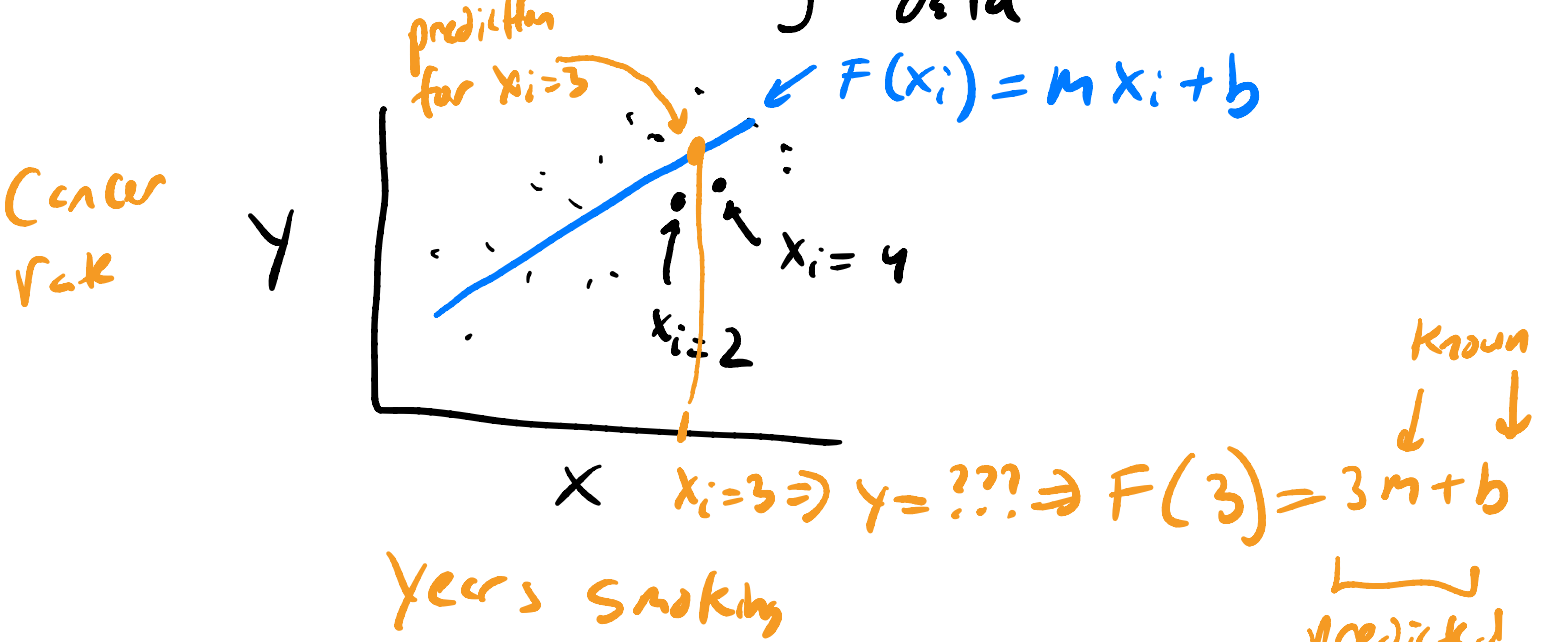
2) Scipy function to give us m + b.
"least squares" function

3) Given m, b values, we can calculate MSSE, plot regression line

Regression Work flow

Have m, b — now what?

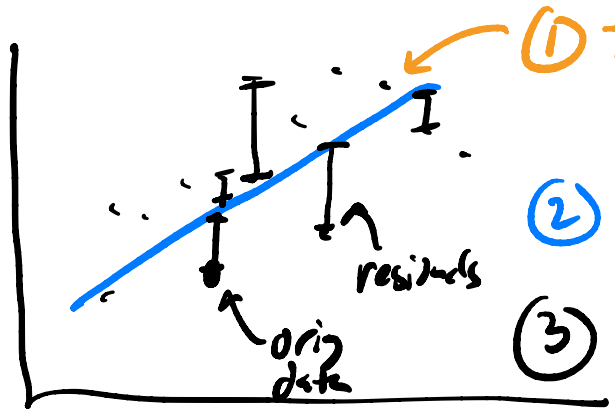
$$y = F(x_i) = m x_i + b$$] predictive model of data



Power: predict new values.

1) Plug in Same independent var values used to fit regression.

   $\Rightarrow$ Calculate how well regression curve fits original



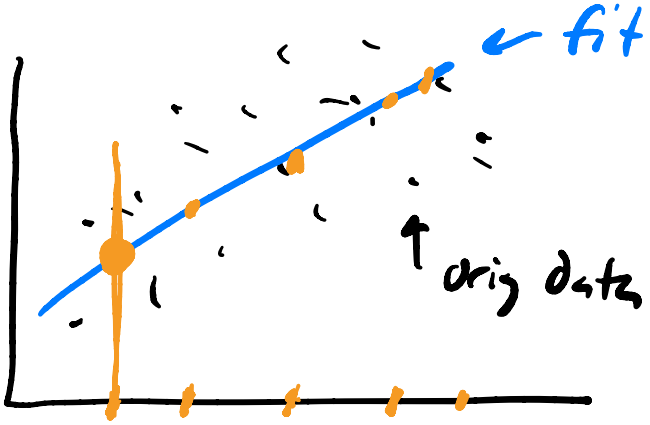
① take data matrix A, get m, b

② draw regression line

③ calc MSSE on <u>fit data</u>

how well did we do?

2) plug in <u>new</u> data into regression equation to get novel prediction



XNew: → predicted y values.