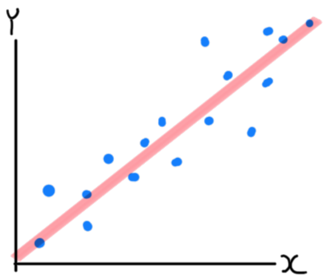
Linear Regression!

What is the main idea?

→ The output variable Models a ==Linear Relationship== Between Input variables



Types of Linear Regression

Simple Linear Regression (SLR)

1 input variable.
Changes in Y is caused by changes in X

Multiple Linear Regression (MLR)

The More Than 1 input

The is More Than

Variable

Terminology

Input variable

aka : predictor variable,
Explanatory variable
Independent variable

Output variable

aka : Response variable
Explained variable
dependen Variable

Note: Gotta Love Namings in Science Amirite?

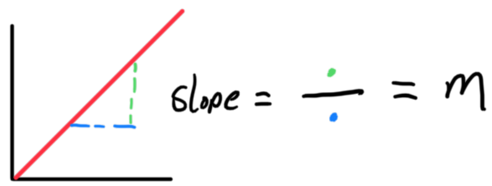
Okay, So what does Linear mean?


=
"I hereby declare linearity to be the property of a mathematical function that can be represented as a straight line"

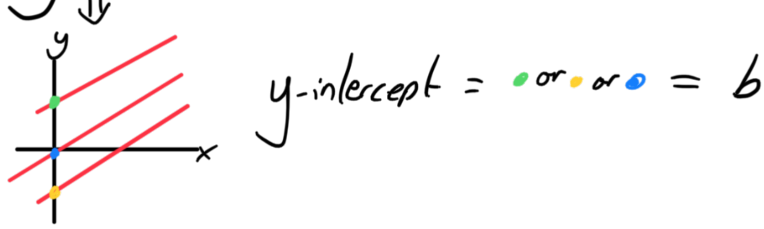
Maths God

Nice!

So a straight line has the formula of $\boxed{f(x) = mx + b}$ where <u>m</u> is the slope of the line



<u>b</u> is the y-intercept of the line



Now for Linear Regression

$$y = \beta_0 + \beta_1 x$$

β_0: y-intercept

β_1: slope

Note

This is for SLR

we will talk about MLR later.

Still the equation of a line!

Linear Regression Goals

Model relationship between an output variable and Input

Variable

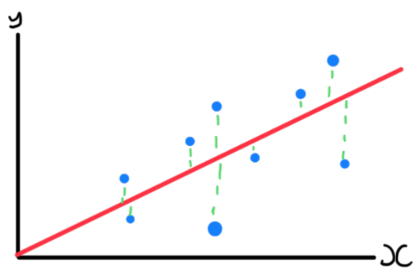
How do we make our Model?

So we know our model will look like $y = \beta_0 + \beta_1 X$ But what are our β_0 and β_1?

In a perfect world we could find β_0 and β_1 exactly but unfortunately for us we can only estimate them.

Our estimations are called $\hat{\beta}_0$ and $\hat{\beta}_1$

Also since we are estimating we will add an error term to our equation to account for the error.



$= \varepsilon =$ Error / Residual

$y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$

Final Model will be.

$y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$

$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$

How do we find a model that fits our data the best?

- Some lines fit better than others

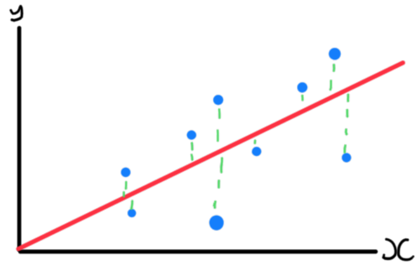
Well one way of checking how well our line fits our data is to check the Residuals.

(from our formula)
$\Downarrow$

$\varepsilon_i = y_i - \beta_0 + \beta_1 x_i$

actual y_i at x_i

line y_i at x_i



So to check how well our line fits we look at the residuals for all our points which we could say is $\sum_{i=1}^{n} \varepsilon_i = \sum_{i=1}^{n} y_i - \beta_0 - \beta_1$

this is wrong

However!

This ...

This formula can turn out to be zero
even if we have massive residuals because
Residuals should alternate in sign since they
are above **and** below the line.
To account for this we square the residual
to get a formula for how well our line
fits the data.

$$\sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$$

$\Downarrow$

- If this number is high our line is not a good fit.
- if this is close to zero it is a good fit
- if it equals zero it is a perfect fit.

Now that we have the formula for checking how well the line fits what can we do.

well with a bit of calculus we can see what happens when

$\sum (y_i - \beta_0 - \beta_1 x_i)^2$ is Zero and the Solve for β_0 and β_1.

This will give us the Best fit Parameters

when we do this:

$$\hat{\beta}_0 = \bar{Y} - \beta_1 \bar{X}$$

$$\hat{\beta}_1 = \frac{S_{XY}}{S_{XX}} = \frac{\sum_{i=1}^{n}(y_i - \bar{y})(x_i - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x}_i)^2}$$

But you ask: "Emm why the little hats?"

Well. These are just Estimates for β_0 and β_1 and thus the hats

---

Fitted Simple Linear Regression Model

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

iwoooo!

Now, A few things about our model.

- we assumed the Errors were Zero

(it was the best fit)

- We can get 75% confidence intervals for $\hat{\beta}_0$ and $\hat{\beta}_1$
- We can check if β_0 or β_1 is significantly different from 0 with a t-test or ANOVA

- SST = Total sum of squares
  ↳ $\sum_i^n (y_i - \bar{Y})$

- SSR: The sum squared deviation from mean. The variation in Y explained by the Regression Line
  ↳ $\sum_i^n (\hat{y}_i - \bar{Y})^2$

- SSE: Sum of Squared Errors. Variation in Y left unexplained
  ↳ $\sum_i^n (\hat{y}_i - y_i)^2$

Also:

$$SST = SSR + SSE$$

SST ↳ The overall variability in Y

SSR ↓ Explained by mode

SSE ↓ unexplained

Once we are finished we might wish to measure the strength of the linear relationship.

To do this we will use the coefficient of determination
$\Downarrow$
R^2

$\boxed{R^2}^{*} = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$ which is the fraction of the variation which is explained by the model.

*R^2 does not tell us if changes in x cause changes in y