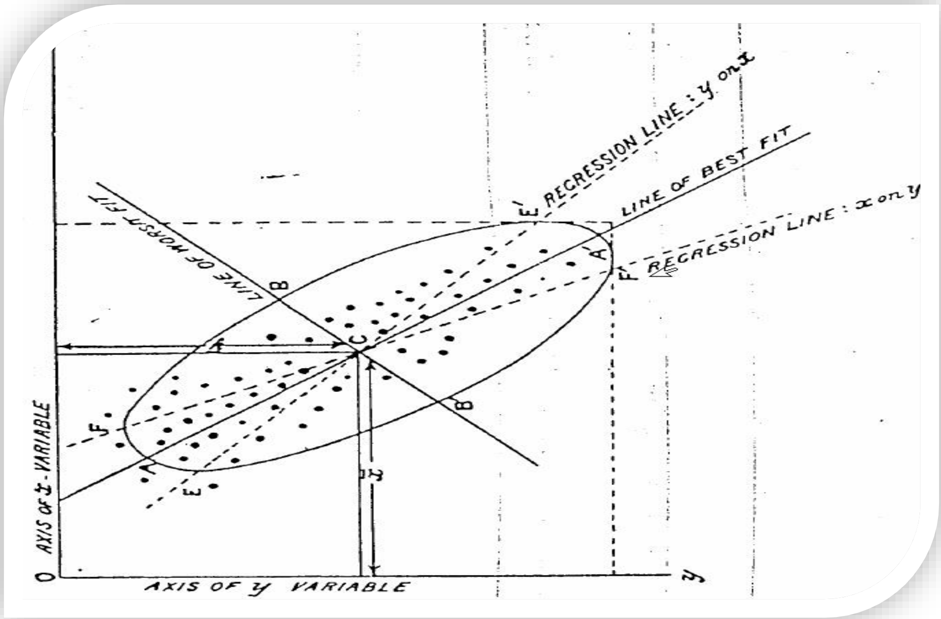
LINEAR REGRESSION

LINEAR REGRESSION is an approach used in Statistics & Machine Learning to model the relationship between a numeric output variable, and one or more explanatory/input variables.

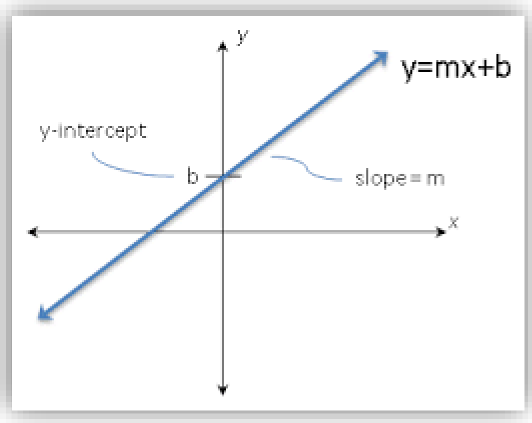
Regression is a method of modelling a target value based on independent predictors. This method is mostly used for forecasting and finding out cause and effect relationship between variables. Regression techniques mostly differ based on the number of independent variables and the type of relationship between the independent and dependent variables.

Simple linear regression is a type of regression analysis where the number of independent variables is one and there is a linear relationship between the independent(x) and dependent(y) variable.

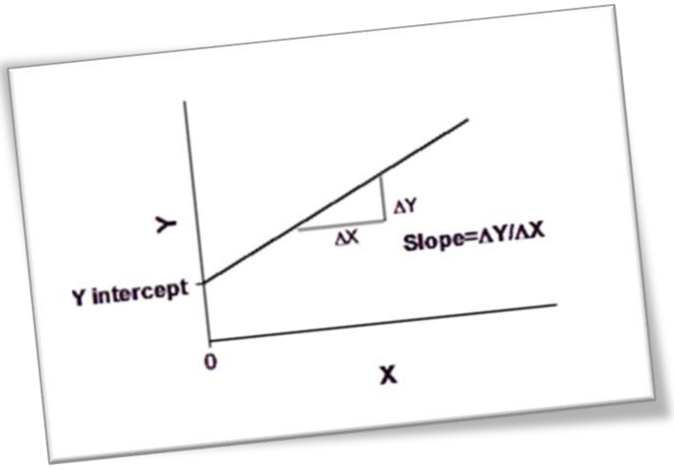


Linear Regression fits the best possible straight line through the data to generalize & represent this relationship.

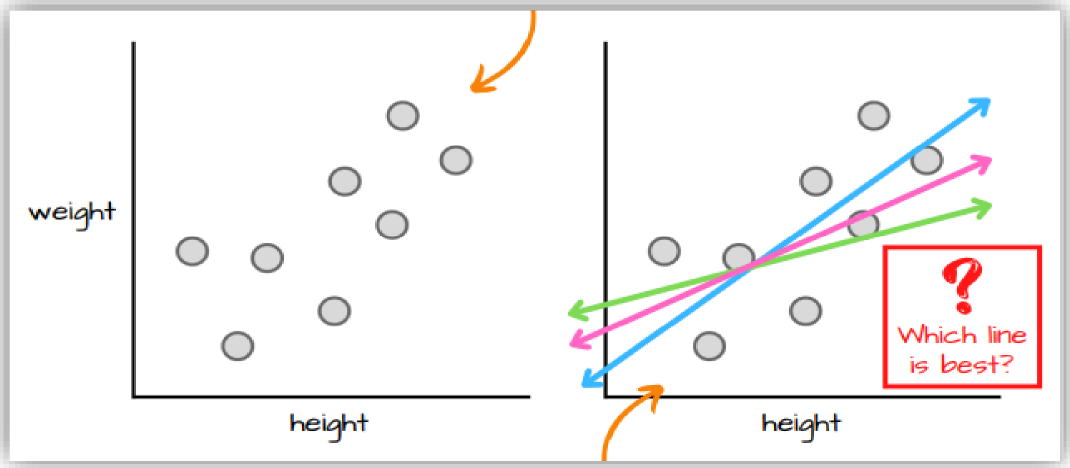
This straight line is defined by an intercept value, and a coefficient value for each input variable.



These values can be used to estimate or predict unknown output values in the future.



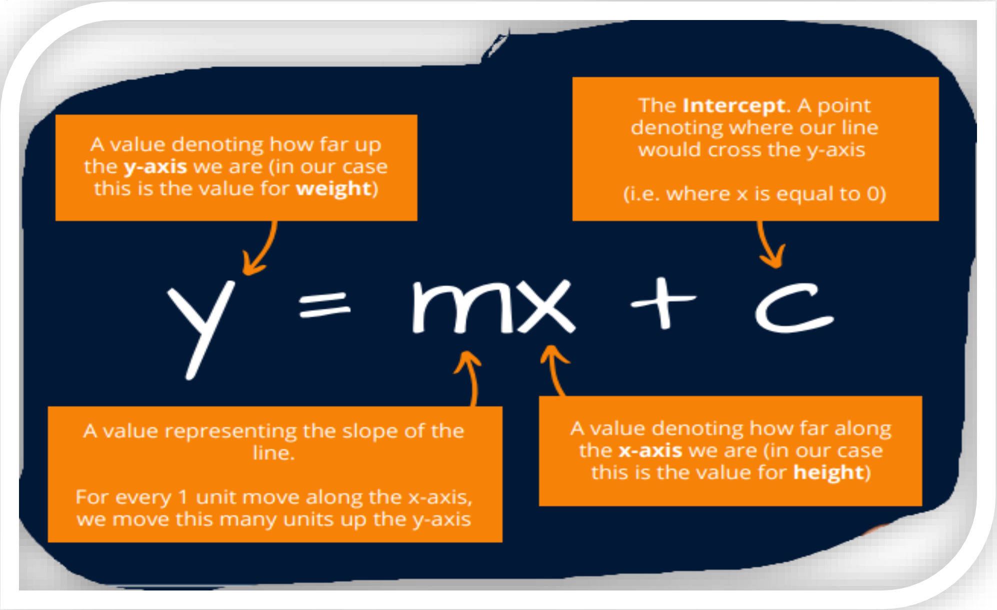
We have height and weight measurements for eight people. Our end goal is to explain weight in terms of height (in other words, how heavy would we expect someone to be at any given height?).



Linear Regression aims to find the line of best fit (in other words the line which best generalizes the relationship between height and weight).

STRAIGHT LINE/ BEST FIT LINE:

Straight lines are often represented by the equation $y = mx + c$ But what does this all mean?



A value denoting how far up the **y-axis** we are (in our case this is the value for **weight**)

The **Intercept**. A point denoting where our line would cross the y-axis

(i.e. where x is equal to 0)

$$y = mx + c$$

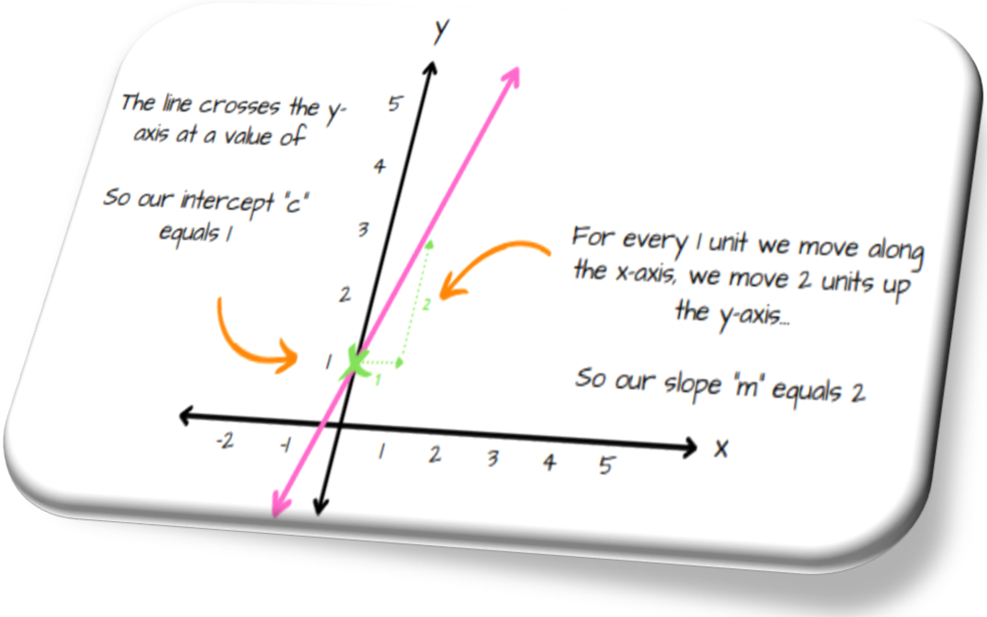
A value representing the slope of the line.

For every 1 unit move along the x-axis, we move this many units up the y-axis

A value denoting how far along the **x-axis** we are (in our case this is the value for **height**)

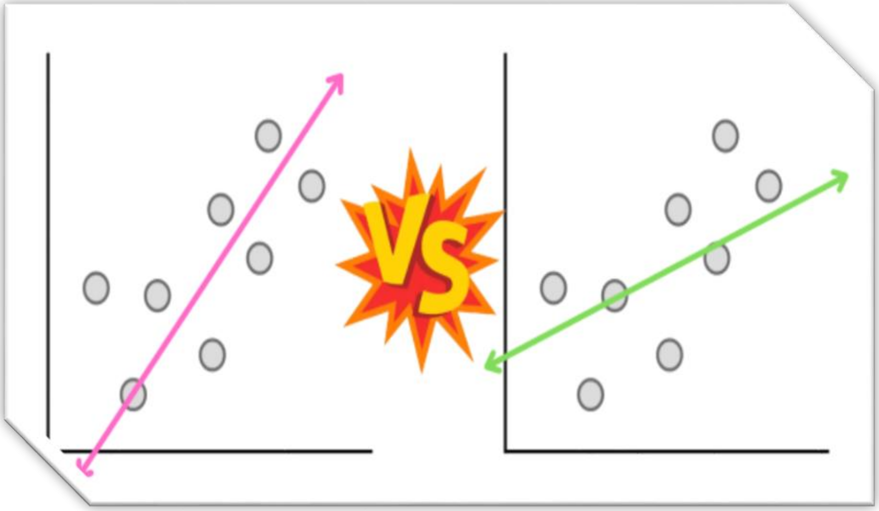
Since $m = 2$ and $c = 1$ our equation for this line is $y = 2x + 1$ we could put any value for x into this equation and be returned a value for y.

In our example this would mean putting in a person's height (x) and be returned a generalized estimate (along the regression line) for their weight (y)!.


y
x
5
4
3
2
1
-2
-1
1
2
3
4
5
1
2
The line crosses the y-axis at a value of
So our intercept 'c' equals 1
For every 1 unit we move along the x-axis, we move 2 units up the y-axis..
So our slope 'm' equals 2

BEST FIT LINE:

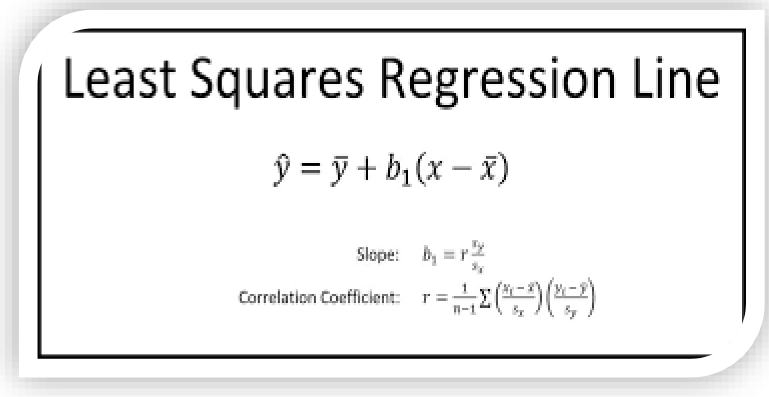
We now know the equation for a straight line, but how does the Linear Regression algorithm decide which line is the best line?



There could be an infinite number of possible lines that run through our data - so which is best, and how do we find it? .

LEAST SQUARES

"Least Squares is an approach used to approximate a line of best fit by minimizing the sum of the squared residuals".



Least Squares Regression Line

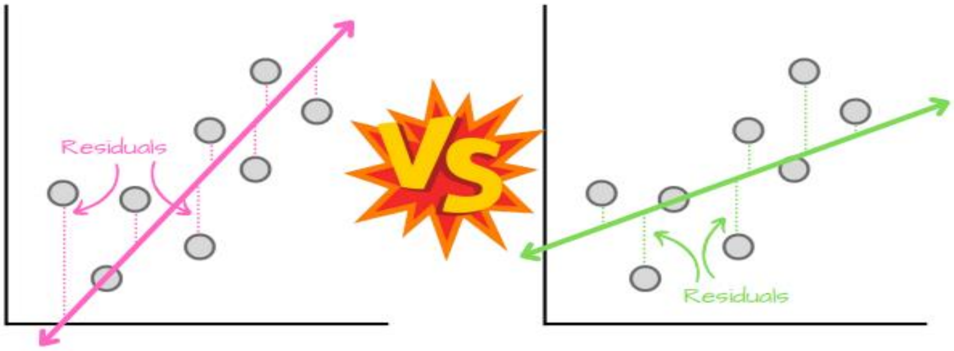
$$\hat{y} = \bar{y} + b_1(x - \bar{x})$$

Slope: $b_1 = r\frac{s_y}{s_x}$

Correlation Coefficient: $r = \frac{1}{n-1}\sum\left(\frac{x_i-\bar{x}}{s_x}\right)\left(\frac{y_i-\bar{y}}{s_y}\right)$

RESIDUALS

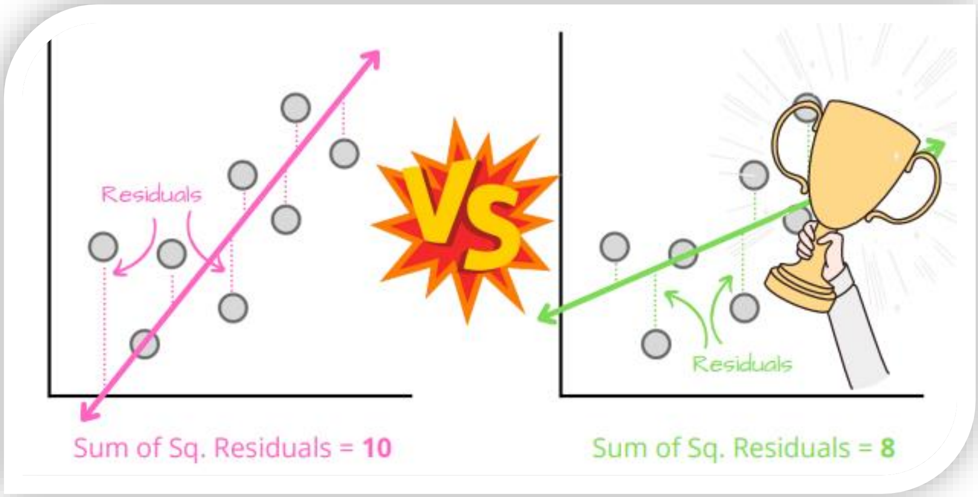
Residuals are the difference between an observed value (our data-points) and the value estimated from the regression line...



According to the definition of Least Squares - the line of best fit will be the one that minimizes the sum of the squared residuals (in other words the line where this value is lowest).

SUM OF SQUARED RESIDUALS

We square the residual values to ensure they are all positive, and thus we can add them all together (the sum) to measure the total deviation from the regression line...



The regression line with the smallest total value for this measure is considered the best fitting line - AKA the "least squares" At an overall level the line represents the observed values better than any other.

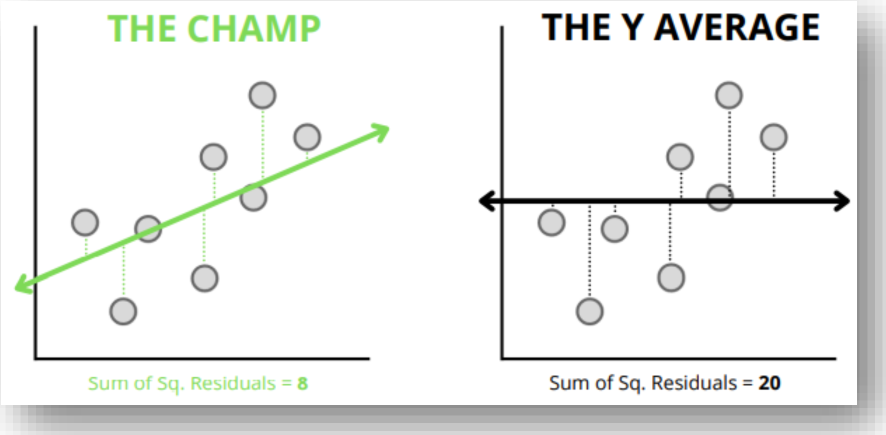
BUT IS THE BEST LINE ACTUALLY ANY GOOD?

We've found the best fitting line using Least Squares, but we still want to know how good that fit is...

The measure for this "goodness of fit" is known as R-Squared...

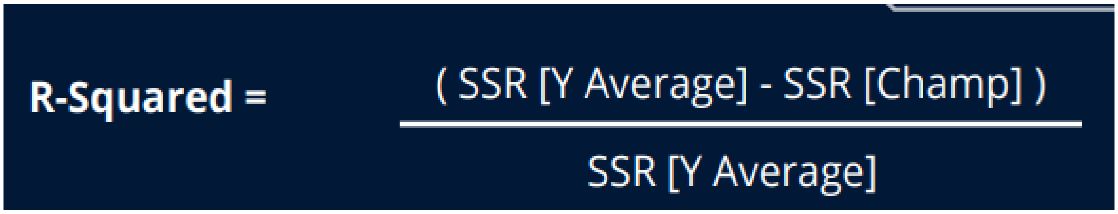
R-SQUARED:

R-Squared shows the percentage of variance in our output variable (y) that is being explained by our input variable(s) (x).

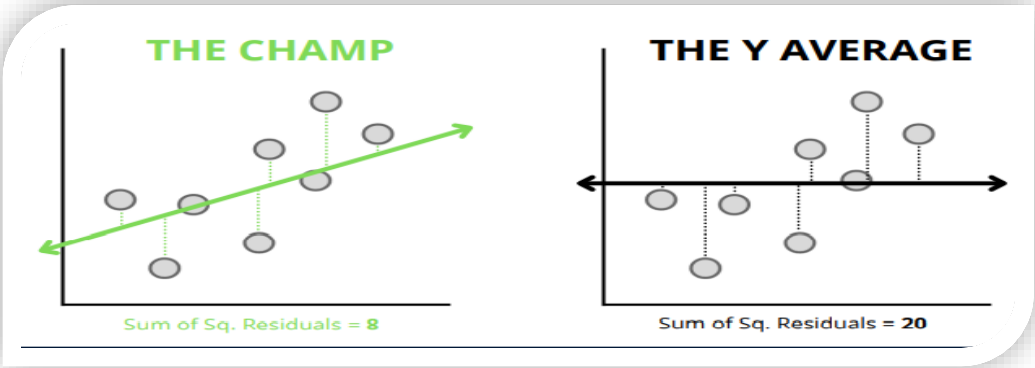


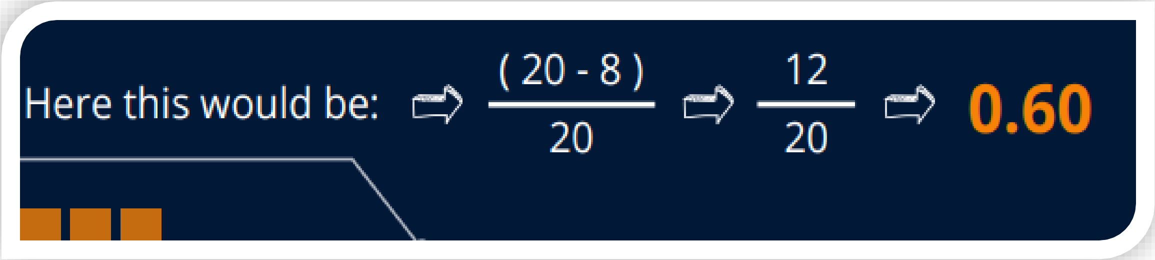
To calculate R-Squared we need both the SSR for our best fitting line, and for a line positioned at the average y value for our data-points.

R-SQUARED FORMULA:



$$\textbf{R-Squared} = \frac{\text{(SSR [Y Average] - SSR [Champ])}}{\text{SSR [Y Average]}}$$





Here this would be: ⇨ $\frac{(20 - 8)}{20}$ ⇨ $\frac{12}{20}$ ⇨ **0.60**

R-SQUARED MEANING…

Our definition was *"R-Squared shows the percentage of variance in our output variable (y) that is being explained by our input variable(s) (x)"*.

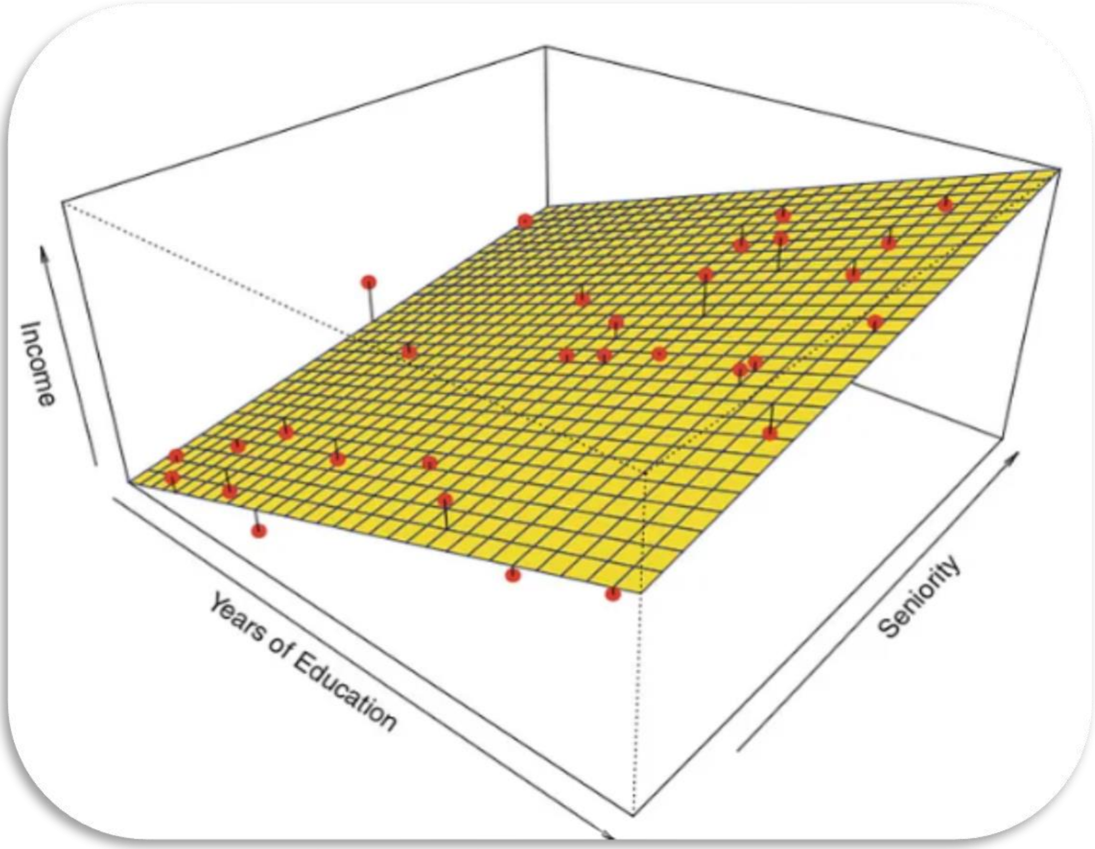
So here we are saying "our input variable (height) is explaining 60% of the variance in our output variable (weight)".

The y-average line is essentially telling us how well we could predict y based on only the y values themselves. Comparing this to our best regression line tells us the additional benefit we're getting by including the input variable(s) into the model.

MULTIPLE INPUT VARIABLE:

In this doc we have described **Linear Regression with only one input variable (called Simple Linear Regression)** - but we can indeed have **many inputs (called Multiple Linear Regression)**.

In the case of **Multiple Linear Regression** we no longer have a line of best fit but instead a plane of best fit across many dimensions (one per input).



Our equation for this plane of **best fit takes in a value for each input along with an accompanying coefficient (slope) value**.

In cases where we have many input variables it is often advised to use Adjusted r-squared rather than the standard r-squared which can become over-inflated. **Adjusted r-squared** compensates for additional input variables and only increases if the additional variable improves the model over and above what would be obtained by probability. **In other words adjusted r-squared scales r-squared by considering the number of input variables that have been included**.

INTERPRETING COEFFICIENT:

Each input variable in the model will have a coefficient value associated with it, which represents its effect on the line (or plane) of **best fit**.

In the case of **simple linear regression the coefficient was our slope value**. For **multiple linear regression the coefficient is essentially the slope value for that particular dimension in space.**

The coefficient value represents the change for the output variable (along the line/plane best fit) for every one unit increase in that input variable, with the proviso that everything else stayed constant.

For example,

If we were predicting house prices, and one input variable "house size" had a coefficient value of 240, this would be saying "for every one unit increase in house size, we would expect house price to increase up by $240.

Each coefficient is also often accompanied by a p-value which provides information around how confident we can be that this relationship between input & output truly exists.