📝 LINEAR REGRESSION

Concepts: 👌

- ✓ It is a statistical method that is used for predictive analysis.
- ✓ It shows a linear relationship between a dependent(y) and independent(x) variables.
- ✓ It finds how the value of the dependent variable is changing with respect to the change of independent variables.
- ✓ The slope represents the relationship between the variables

Numerical and Symbolism: 👌

$y = bo + ax + e$

⚠️ y: dependent variable

x: independent variable — a: slope

bo: constant — e: random error

TYPES OF LINEAR REGRESSION 🙁

1. Simple linear regression
2. Multiple linear regression

🗒️ SIMPLE LINEAR REGRESSION

- There is a single independent variable which will predict the dependent variables value.
- The dependent variable should be continuous.

$y = b_o + ax + e$

🗒️ MULTIPLE LINEAR REGRESSION

- There is more than one independent variable that will predict the value of the dependent variables.

$y = b_o + a_1x_1 + a_2x_2 + \ldots + a_nx_n + e$

⚠️ MAIN GOAL 🙁

To find the best fit line that means the error between the predicted values and actual values should be minimized and optimized. ✌️

EXAMPLE OF SIMPLE LINEAR REGRESSION

X	Y	$X-X_{mean}$	$Y-Y_{mean}$	$(X-X_{mean})^2$	$(X-X_{mean})(Y-Y_{mean})$	$(Y-Y_{mean})^2$
0	2	-2	-2	4	4	4
1	3	-1	-1	1	1	1
2	5	0	1	0	0	1
3	4	1	0	1	0	0
4	6	2	2	4	4	4
				= 10	= 9	= 10

⚠️ X_{mean}: mean of X

$$X_{mean} = (0+1+2+3+4)/5$$

$$= 10/5$$

$$= 2$$

⚠️ Y_{mean}: mean of Y

$$Y_{mean} = (2+3+5+4+6)/5$$

$$= 20/5$$

$$= 4$$

Wait .... Note done Yet 🤣 ⏭️

Slope: ☹️

$m = \sum((X - X_{mean})(Y - Y_{mean})) / \sum((X - X_{mean})^2)$

$m = 9/10$

$m = 0.9$

📝 *We need to find the constant value.*

$y = mx + b$

⚠️ *The line must pass through mean value of x, and y*

This means: $y = 4$, $x = 2$, $m(slope) = 0.9$

$y = mx + b$

$4 = 2 * 0.9 + b$

$b = 4 - 1.8$

$b = 2.2$

💥 Boom 💥

$y = 0.9x + 2.2$

X	Y(Yactual)	Yexpected	Yexpected-Yactual	$(Yexpected-Yactual)^2$	$Yexpected-Yexpected_{mean}$	$(Yexpected-Yexpected_{mean})^2$
0	2	2.2	0.2	0.04	-1.8	3.24
1	3	3.1	0.1	0.01	-0.9	0.81
2	5	4	-1	1	0	0
3	4	4.9	0.9	0.81	0.9	0.81
4	6	5.8	-0.2	0.04	1.8	3.24
				=1.90		=8.1

📝 Yexpected will be generated by using new equation we have derived earlier. ✌️ $y = 0.9x + 2.2$

😎 FINDING THE STANDARD ERROR USING LEAST SQUARE METHOD

$Error = \sqrt{(\sum(Yexpected-Yactual)^2)/(n-2)}$

$= \sqrt{(1.90/3)}$

$= 0.633$

📝 NB: smaller values are better because it indicates that the observations are closer to the fitted line. 👌

😎 FINDING THE STANDARD ERROR USING R^2 METHOD

$R^2 = \sum(Yexpected-Yexpected_{mean})^2/\sum(Y-Y_{mean})^2$

$Yexpected_{mean} = (2.2+3.1+4+4.9+5.8)/5$

$=(20)/5$

$= 4$

🔶 This shows that the mean of expected value and mean of actual value will be always the same.

$R^2 = \sum(Yexpected-Yexpected_{mean})^2/\sum(Y-Y_{mean})^2$

$= 8.1/10$

$= 0.81$

📝 NB: If the value of $R^2 = 1$ then the expected value is the same as the actual value. If the value of R^2 approaches to 0(zero), then there are huge gabs between the actual value and expected value.

Break time: ✌️ ✌️

📝 Most common terminologies:

1. **Predictor** |-> independent variable | explanatory
2. **Target** |->outcome | dependent variable | response
3. **Outlier** |-> Either very low or high value compared to other observed value.

🔶 Problem with **ALGORITHM**

a. **Underfitting** |-> The algorithm which doesn't work well with training dataset and testing dataset as well

b. **Overfitting** |-> The algorithm which works well with the training dataset

✍️ APPLICATION OF LINEAR REGRESSION

1. Analysing trends and sales estimates
2. Salary forecasting
3. Real estate prediction & etc.

💥 💥 Boom 💥 💥