Correlation & Regression

Created	@Jan 18, 2021 1:40 PM
Created By	K Khanh Vương
Last Edited By	K Khanh Vương
Last Edited Time	@Jan 18, 2021 1:57 PM
Module	Introduction to Machine Learning
Status	
Type	Introduction to Machine Learning

- To show the relation between two variables, we can use **Contingency Table**
- We must notice that, **Contingency Table** only applied to **Nominal** and **Ordinal** variables. For **Quantitative** variables, we must use **Scatter Plot**
- **Scatter Plot** shows us the relationship, but not the exact number. To know the exact number of the relationship, compute the **Pearson's R**

$$r = \frac{\sum Z_x Z_y}{n-1}$$

  - **Pearson's R** show the strength, and the direction of the **Linear Correlation** between two variables
  - We must **Standardized** the values x and y because we want the **Pearson's R** is a number between `-1` and `1`
- After having the **Scatter Plot,** we can plot a **Regression Line** to indicate the **Linear Correlation** between two variables. **Regression Line** is the line with the smallest sum of squared residuals
- But the line also just give us the overview. Instead, describe the line with a formula will help us get the exact number, and easy to predict the values:

$$y = ax + b$$

$$b = r(\frac{s_y}{s_x})$$

$$a = \mu_y - b\mu_x$$

  - a is the **Intercept** - A constant that predict the value of y when $x = 0$

    - b is the **Regression Coefficient** - The change of y when x increase in a unit number
- Finally, we can compute how the **Regression Line** fix the data with r^2
    - r^2 is nothing more than a number that tells you how much better a **Regression Line** predicts the value of a dependent variable than the mean of that variable.
    - It shows you how much of the variance in dependent variable is explained by independent variable.
    - For example: $r^2 = 0.69 \Rightarrow$ Prediction error is 69% smaller than when we use the μ