# REGRESSION

The focus this week is on **regression**, which is a statistical method used to model and analyse the relationships between a *dependent* variable and one or more *independent* variables

## Correlation

Measures the strength & direction of the relationship between two variables.

**Pearson correlation (r):** for sample data.
**Population correlation (ρ):** for entire population.

Values range from -1 to +1, with values closer to ±1 indicating a stronger relationship.

## Regression Model

$$\hat{y} = b_0 + b_1 x$$

Intercept (bo):
- Value of y, when x = 0
- Where the line crosses the y axis.

slope (b1):
- Defines how steep the line is
- Which direction the line goes (positive/negative).

## Sources of Variation

**$R^2$**

**Definition:** $R^2$ otherwise known as R-squared represents the proportion of variability in the outcome variable that is explained by the regression model.

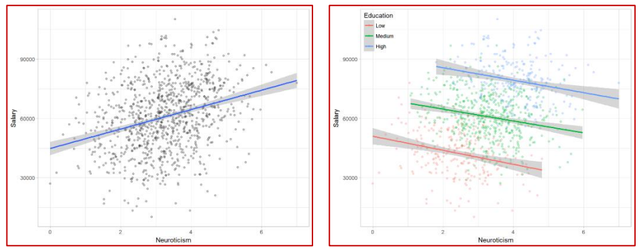
The Coefficient of Determination is the proportion of variation explained by the model.

**Range: 0 (no explanatory power) to 1 (full explanatory power).**

## Assumptions of Linear Regression

**L I N E**

**Linearity** of the relationship between x and y.
**Independence** of residuals (errors).
**Normality** of residuals.
**Equality** of variance of residuals (homoscedasticity)




## Hypothesis Testing in Regression

Focuses on the slope:
Null Hypothesis (no relationship). $H_0: \beta_1 = 0$

Alternative Hypothesis (relationship exists). $H_1: \beta_1 \neq 0$

Significance level (α): Often 5% threshold.

Tests for normal distribution, linear relationship, and constant variance of residuals.

## Simple Example for Regression

- Regression line predicts a score of Y for any given value of X
- The closer the observed scores are to the predicted scores, the better the model predicts y, the less variability around the line there is.
- The further away from the line (predicted scores) the observed scores are, the worse the model predicts y, the more variability around the line there is.
- Difference between predicted Y and observed Y for any given value of X is called a **residual.**

## Prediction

**Steps:**

1. **Verify assumptions**
2. **Check model significance and variance explanation.**
3. **Make predictions within the observed range of x.**