Machine learning

Regression

Regression techniques are a category of machine learning algorithms that take labeled data and learn patterns in the data that can be used to predict a continuous output variable.

- How much carbon dioxide does a household contribute to the atmosphere?
- What will the share price of a company be tomorrow?
- What is the concentration of insulin in a patient's blood?

Statistics refresher - standard deviation

The **standard deviation of a variable** is a measurement of the amount of variability present. It is measured in the same units as the variable itself and tells us how spread out the instances of the variable are from the mean.

- if the standard deviation is low, the data points tend to be close to the mean;
- if a high standard deviation tells us to expect data points that are relatively far from the mean.

The standard deviation of a variable is normally expressed using the lowercase Greek letter sigma σ. In R we have **sd** function.

$$\sigma = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n}}$$

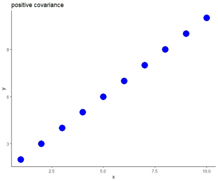
Statistics refresher - covariance

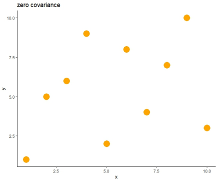
The **covariance** between two variables measures their joint variability. This is a measure of how strong the relationship is between those two variables, or how much one variable is likely to change in response to a change in the other variable.

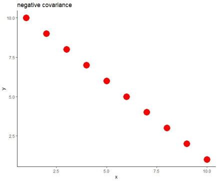
Covariance values range from $-\infty$ to ∞

$$cov(x, y) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n - 1}$$

Statistics refresher - covariance







- $cov(x, y) > 0$ (positive) - the x and y values tend to rise together (not how fast they rise or fall);
- $cov(x, y) < 0$ (negative) - the x rises as y falls (and vice versa);
- $cov(x, y) = 0$ - zero result (rarely happens with statistical data) just means the covariance does not let us know if x and y rise or fall together.

Correlation

Correlation is a statistical measure that expresses the extent to which two variables are linearly related (meaning they change together at a constant rate). It's a common tool for describing simple relationships without making a statement about cause and effect. Correlation also cannot accurately describe curvilinear relationships.

Pearson correlation coefficient

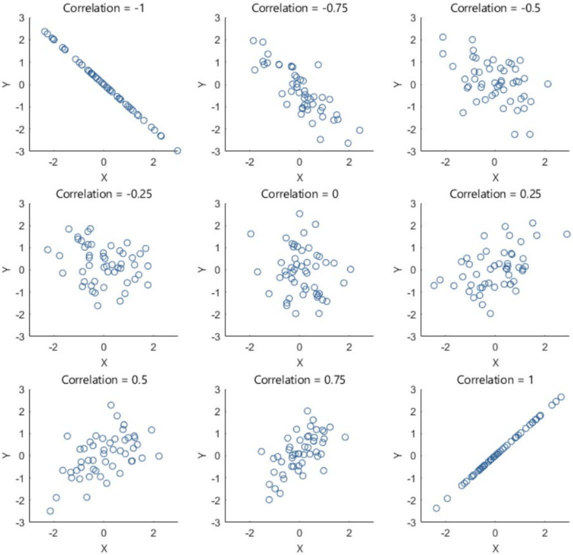
Pearson correlation coefficient (ρ) between two random variables x and y is denoted as follows

$$\rho_{x,y} = \frac{cov(x,y)}{\sigma_x, \sigma_y}$$

The values of Pearson's correlation coefficient range from -1 to +1.

- The closer ρ is to zero, the weaker the linear relationship.
- Positive ρ values indicate a positive correlation, where the values of both variables tend to increase together.
- Negative ρ values indicate a negative correlation, where the values of one variable tend to increase when the values of the other variable decrease.

Different correlation coefficient



Regression

Regression analysis is a family of statistical methods that are used to model complex numerical relationships between variables. In general, regression analysis involves three key components.

- A single numeric dependent variable, which represents the value or values that we want to predict. This variable is known as the response variable Y.
- One or more independent numeric variables X that we believe we can use to predict the response variable. These variables are known as the **predictors**.
- Coefficients β, which describe the relationships between the predictors and the response variable. We don't know these values going into the analysis and use regression techniques to estimate them. The coefficients are what constitute the regression model.

$$Y = f(X, \beta)$$

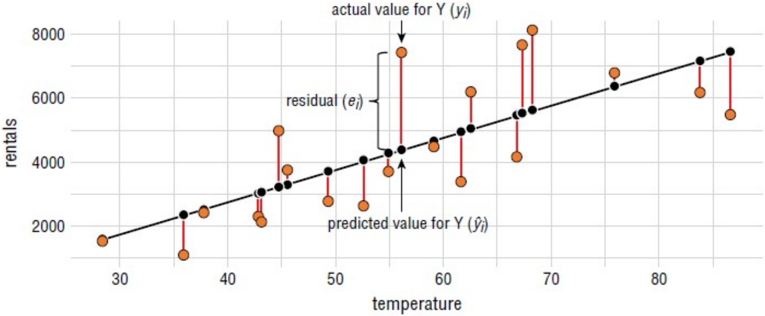
Simple linear regression

Linear regression is a subset of regression that assumes that the relationship between the predictor variables X and the response variable Y is linear. In cases where we have only a single predictor variable, we can write the regression equation using the **slope-intercept format.**

$$Y = \beta_0 + \beta_1 X$$

- β_0 - intercept - is the expected value for Y when $X = 0$;
- β_1 - slope - is the expected increase in Y for each **unit** increase in X.

Simple linear regression



Thank you for your attention!!!