# linear regression definition444444

Linear regression is a type of mathematical technique that is used to illustrate the correlation between a dependent variable and one or more independent variables. The dependent variable is the one you want to forecast, whereas the independent variable is the one used to predict its value. This technique is widely used in several fields, including economics, engineering, and social sciences, to make predictions and understand the relationship between variables (Cheusheva, 2023).

Using one or more independent variables that can most accurately predict the value of the dependent variable, this analysis determines the coefficients of the linear equation. The disparities between expected and actual output values are minimized using linear regression by fitting a straight line or surface (Cheusheva, 2023). The best-fit line for a set of paired data can be found using a linear regression calculator that employs the "least squares" approach. The independent variable (Y) can then be used to estimate the value of the dependent variable (X).

Linear regression can be performed on various programming environments such as R, MATLAB, Sklearn, Python, and Excel. By transforming large amounts of data into actionable information, linear regression can assist businesses in making better decisions. These insights can help to identify patterns and relationships that can be used to gain a competitive edge.

The equation for a simple linear regression model with one independent variable can be written as:

y = β0 + β1x + ε

Where y is the dependent variable, x is the independent variable, 0 represents the line's intercept, 1 represents its slope, and represents the error term or residuals. The change in y for a unit change in x is represented by the slope 1, while the intercept 0 reflects the expected value of y when x is zero. The discrepancy between the expected value of y and the actual value of y is represented by the error term. We employ the least squares approach to estimate the coefficients 0 and 1 (Dusane, 2023). Finding the values of 0 and 1 that reduce the sum of the squared discrepancies between the expected and actual values of y is the goal of this strategy. Once the coefficients have been computed, the regression line can be used to forecast the value of y for any given value of x (Dusane, 2023). For a successful linear regression analysis, some assumptions must be considered:

* Linearity: The relationship between the dependent and independent variables should be linear. This means that changes in the independent variable should correspond to changes in the dependent variable (Cheusheva, 2023).
* Normality: The data should be normally distributed. This means that the distribution of residuals should be normal (Cheusheva, 2023).
* Homoscedasticity: Homoscedasticity means that the variance of the residuals should be constant across all levels of the independent variable. This assumption ensures that the errors are randomly distributed and are not dependent on the values of the independent variable (Cheusheva, 2023).
* Independence: The observations should be independent of each other. This means that each observation should not be influenced by the values of other observations (Cheusheva, 2023).
* No Multicollinearity: The independent variables should not be highly correlated with each other. This assumption ensures that each independent variable contributes unique information to the model (Cheusheva, 2023).
* No Outliers: The data should not contain outliers, which are extreme values that can have a significant impact on the results of the analysis (Cheusheva, 2023).

Linear regression has some benefits and drawbacks. Benefits include simplicity, it’s a simple and easy-to-understand technique that requires minimal assumptions and easy to interpret, versatility, it can model a wide variety of relationships between variables, prediction, it can make predictions about the value of the dependent variable based on the values of the independent variables, quantification, the quantification of the strength and direction of the relationship between the dependent variable and the independent variables, and hypothesis testing, to determine whether there is a significant relationship between the dependent variable and the independent variables. Drawbacks include linearity assumptions, it assumes a linear regression and if this assumption is not met, the results may be inaccurate, overfitting and underfitting, overfitting means that the model makes non accurate predictions and underfitting means that the model makes accurate, but initially incorrect predictions.

# Why this dataset

The reason as to why this fish dataset was chosen is because it consists of 7 attributes or features (independent variables) of fish and their corresponding weight (dependent variable). The attributes include length of various body parts such as vertical length, diagonal length, and cross length, as well as the height and width of the fish. The dataset has a relatively large sample size of 159 fish observations, providing sufficient data for linear regression analysis. This data set is appropriate for analysis with linear regression because it involves a continuous dependent variable (fish weight) and several independent variables (environmental and biological factors) that may influence the weight of the fish. There are several reasons why the Fish dataset is appropriate for analysis with linear regression, Linearity: The dependent variable (weight) has a linear relationship with the independent variables, making linear regression a suitable choice for modelling the relationship between weight and the different fish attributes. Independence: Each observation in the dataset is assumed to be independent of the others, meaning that the weight of one fish should not be influenced by the weight of another fish or the fish attributes of another fish. Homoscedasticity: The variance of the errors in the model is assumed to be constant across all levels of the independent variables. This assumption can be checked by plotting the residuals against the fitted values and ensuring that there is no pattern or trend. Normality: The errors of the model are assumed to be normally distributed, which can be checked by examining the histogram of the residuals and performing a normality test. Adequate sample size: The dataset has a relatively large sample size of 159 observations, which provides enough data for accurate estimates of the regression coefficients. Linear regression can be used to identify the relationship between the dependent variable (fish weight) and each independent variable separately, as well as to determine which variables have the strongest impact on the fish weight. The regression model can also be used to predict the expected weight of a fish based on the values of the independent variables. Insights that can be obtained from analysing the fish data set with linear regression include identifying which variables have the strongest impact on the weight of the fish, understanding how changes in the independent variables affect the weight of the fish, and predicting the weight of fish under different environmental conditions.

# Analysis being conducted

In this dataset, the following analysis will be conducted for predicting the weight of the fish:

* A descriptive analysis that describes the overall data
* An explorative data analysis that will visualise the dataset in different formats and different types of graphs.
  + This includes pair plots, histogram, scatter plots, heat maps
* Analysing if there are any outliers
  + This uses box plots
* Doing a linear regression analysis using the train test function
  + Included in this is the linear regression points
  + Predicting model accuracy
  + Assessing whether the model meets linear regression assumptions
  + Testing model predictions with example data
  + Seeing whether the model is either overfitted or underfitted
* Plotting the predicted x and y values on a scatter plot showing
* Showing a table of predicted and actual weight and the difference between them.

# Reference

Pyae, A. (2021). Fish Market [Data set]. Kaggle.

*About linear regression* (no date) *IBM*. Available at: https://www.ibm.com/topics/linear-regression#:~:text=Resources-,What%20is%20linear%20regression%3F,is%20called%20the%20independent%20variable. (Accessed: April 28, 2023).

Cheusheva, S. (2023) *Linear regression analysis in Excel*, *ablebits.com*. Available at: https://www.ablebits.com/office-addins-blog/linear-regression-analysis-excel/ (Accessed: April 28, 2023).

Dusane, J. (2023) *Popular applications of linear regression for businesses*, *UNext*. Available at: https://u-next.com/blogs/business-analytics/popular-applications-of-linear-regression-for-businesses/#:~:text=Linear%20regressions%20can%20be%20used,forecast%20sales%20in%20future%20months. (Accessed: April 28, 2023).