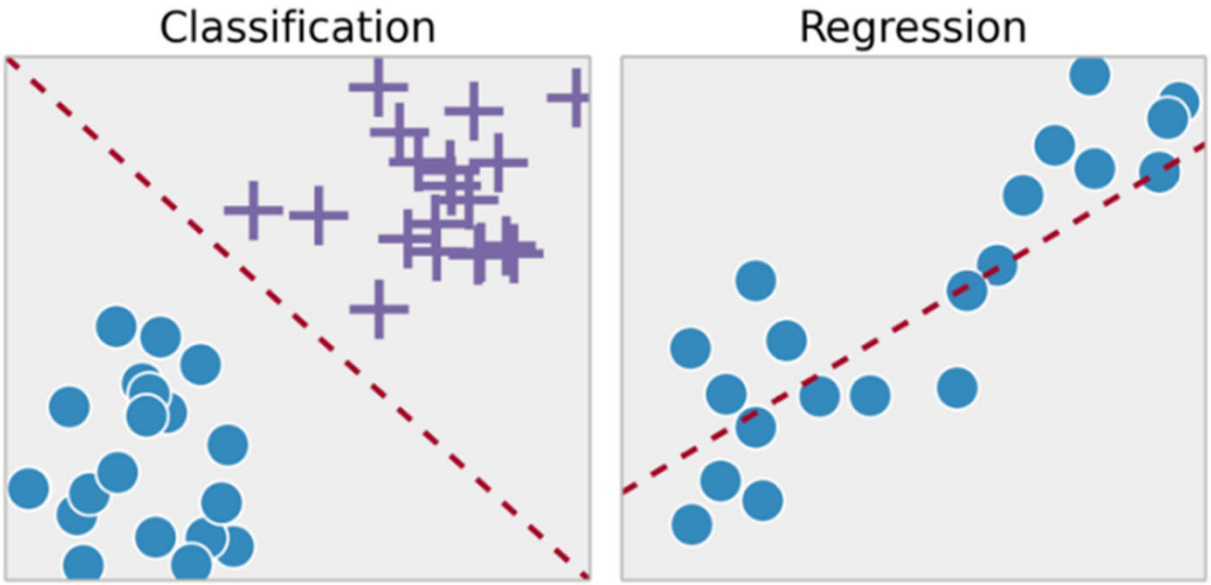
Definition

Learning a target function (f) that best maps input variables (X) to an output variable (Y):
Y = f(X)



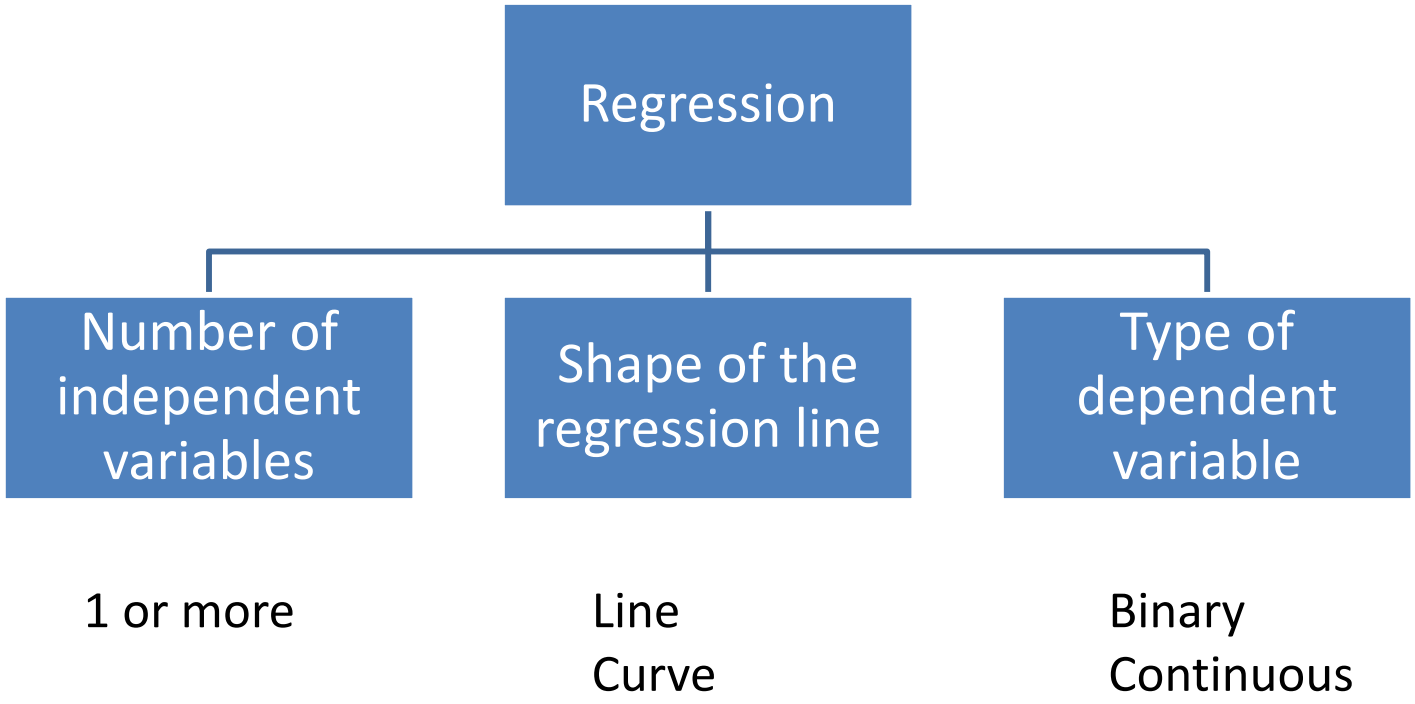
K-nearest neighbours
Decision trees

Linear regression

Regression

- Method of modelling a target value based on independent predictors.
- Mostly used for forecasting and finding out cause and effect relationship between variables.
- Regression techniques mostly differ based on the number of independent variables and the type of relationship between the independent and dependent variables.
- Benefits:
  - Significant relationships between dependent variable and independent variable.
  - Strength of impact of multiple independent variables on a dependent variable.
  - Compare the effects of variables measured on different scales, such as the effect of price changes and the number of promotional activities.
- These benefits help market researchers / data analysts / data scientists to eliminate and evaluate the best set of variables to be used for building predictive models.

4 MACHINE LEARNING ALGORITHMS



4.4. Supervised learning

Simple linear regression

Simple linear regression is used to estimate the relationship between two quantitative variables. Useful to know:

- How strong the relationship is between two variables (e.g. the relationship between rainfall and soil erosion).
- The value of the dependent variable at a certain value of the independent variable (e.g. the amount of soil erosion at a certain level of rainfall).

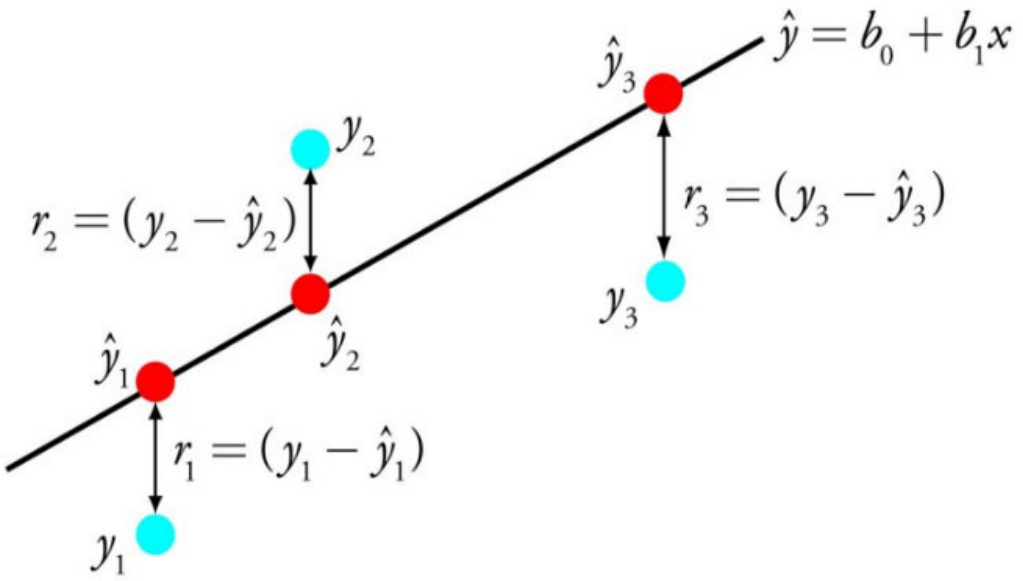
The dependent variable is continuous, the independent variable can be continuous or discrete, and nature of regression line is linear.

Simple linear regression

Assumptions:

1. Homogeneity of variance (homoscedasticity): the size of the error in our prediction doesn't change significantly across the values of the independent variable.
2. Independence of observations: the observations in the dataset were collected using statistically valid sampling methods, and there are no hidden relationships among observations.
3. Normality: The data follows a normal distribution.
4. The relationship between the independent and dependent variable is linear: the line of best fit through the data points is a straight line (rather than a curve or some sort of grouping factor).

Simple linear regression



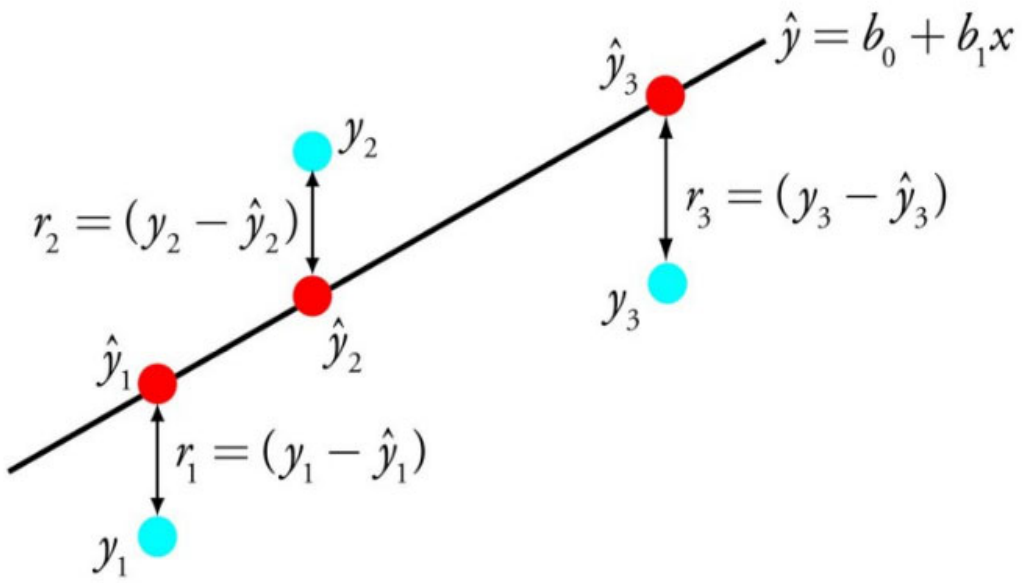
- y is the predicted value of the dependent variable (y) for any given value of the independent variable (x).
- B0 is the intercept, the predicted value of y when the x is 0.
- B1 is the regression coefficient – how much we expect y to change as x increases.
- x is the independent variable (the variable we expect is influencing y).

The goal is to find an equation that describes a line that best fits the relationship between the input variables (x) and the output variables (y), by finding specific weightings for the input variables called coefficients (B).

Simple linear regression

How to make a prediction?

Just substitute the new x value in the equation to get the estimate of y.



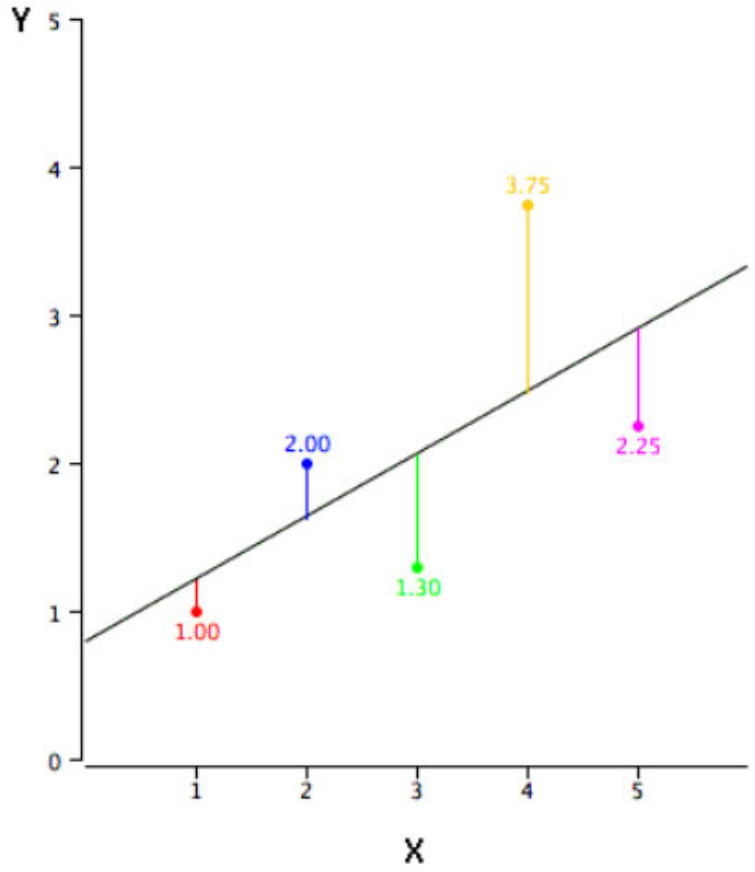
Can you predict values outside the range of your data?

No! Regression models can be used to predict the value of the dependent variable at certain values of the independent variable. However, this is only true for the range of values where we have actually measured the response.

Simple linear regression

$$\min_w ||Xw - y||_2^2$$

This task can be easily accomplished by Least Square Method. It is the most common method used for fitting a regression line. It calculates the best-fit line for the observed data by minimizing the sum of the squares of the vertical deviations from each data point to the line. Because the deviations are first squared, when added, there is no cancelling out between positive and negative values.



Simple linear regression: metrics

Mean Absolute Error (MAE): MAE measures the average magnitude of the errors in a set of predictions, without considering their direction. It's the average over the test sample of the absolute differences between prediction and actual observation where all individual differences have equal weight.

$$\mathrm{MAE} = \frac{1}{n}\sum_{j=1}^{n}|y_j - \hat{y}_j|$$

Root mean squared error (RMSE): RMSE is a quadratic scoring rule that also measures the average magnitude of the error. It's the square root of the average of squared differences between prediction and actual observation.

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{j=1}^{n}(y_j - \hat{y}_j)^2}$$

Both MAE and RMSE express average model prediction error in units of the variable of interest and can range from 0 to ∞ and are indifferent to the direction of errors. They are negatively-oriented scores, which means lower values are better. RMSE is more useful when large errors are particularly undesirable. However, MAE is easier to interpret.

RMSE gives us an idea of the average distance between the observed data values and the predicted data values

Simple linear regression: metrics

R2 is the coefficient of determination; it tells us how much variation the dependent variable has that can be predicted from the independent variable. In other words, how well the model fits our actual observations. The best possible value we have with R2 is 1 and the worst is 0.

When we use R2, all independent variables that are in our model contribute to its value. A disadvantage it has is that it assumes that every variable helps explain the variation in the prediction, which is not always true. If we add another variable, the value of R2 increases or stays the same but never decreases. This may lead us to believe that the model is improving, but this is not necessarily the case.

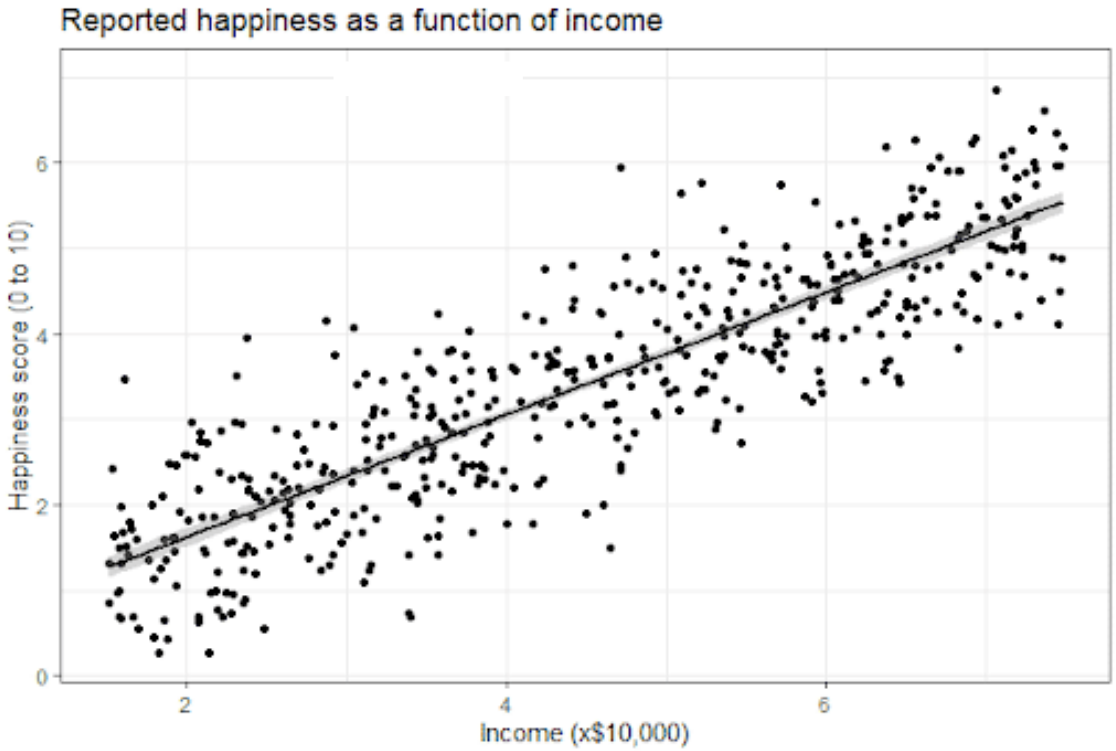
$$R^2 = 1 - \frac{SS_{RES}}{SS_{TOT}} = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \overline{y})^2}$$

Simple linear regression

Example:

You are a social researcher interested in the relationship between income and happiness. You survey 500 people whose incomes range from $15k to $75k and ask them to rank their happiness on a scale from 1 to 10. Your independent variable (income) and dependent variable (happiness) are both quantitative, so you can do a regression analysis to see if there is a linear relationship between them.

$$happiness = 0.128 + 0.725 * income$$



Simple linear regression

Example:

$$happiness = 0.128 + 0.725 * income$$

Interpretation: 0.725 of increase in happiness for each $10000 of income

We can predict the happiness level of a person with a certain income. For example, if a person has an income of 50k, her happiness level will be:

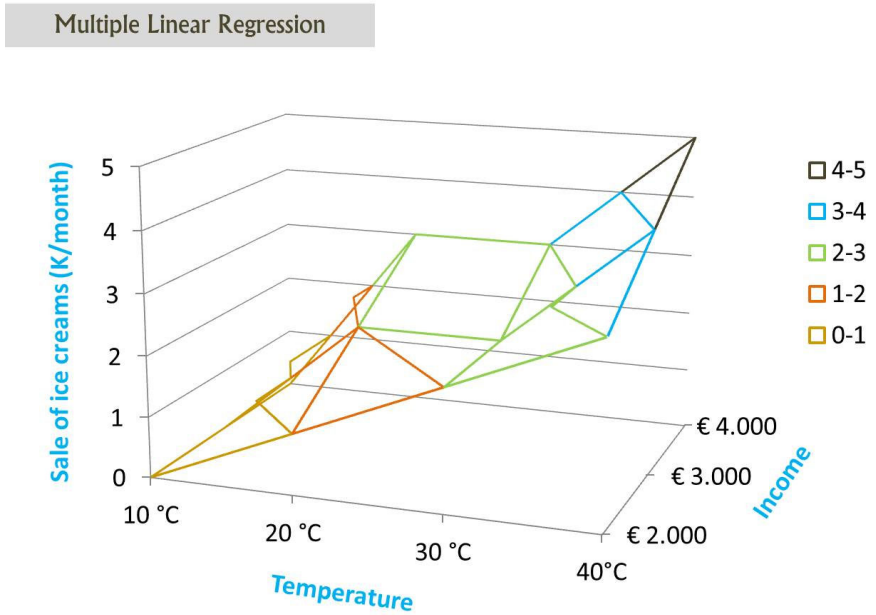
0.128 + 0.725*5 = 3.753

Mean Absolute Error: 0.617405060886751
Mean Squared Error: 0.583815358536913
Root Mean Squared Error: 0.7640781102437704
Mean Happiness: 3.392859263581404

Multiple linear regression

Multiple linear regression is used to estimate the relationship between two or more independent variables and one dependent variable. Useful to know:

- How strong the relationship is between two or more independent variables and one dependent variable (e.g. how rainfall, temperature, and amount of fertilizer added affect crop growth).
- The value of the dependent variable at a certain value of the independent variables (e.g. the expected yield of a crop at certain levels of rainfall, temperature, and fertilizer addition).



Multiple linear regression

$$y = \beta_0 + \beta_1 X_1 + ... + \beta_n X_n + \varepsilon$$

- y = the predicted value of the dependent variable
- B0 = the y-intercept (value of y when all other parameters are set to 0)
- B1X1= the regression coefficient (B1) of the first independent variable (X1) (a.k.a. the effect that increasing the value of the independent variable has on the predicted y value)
- … = do the same for however many independent variables you are testing
- BnXn = the regression coefficient of the last independent variable

(Same assumptions than in simple linear regression)

Multiple linear regression

Example:

You are a public health researcher interested in social factors that influence heart disease. You survey 500 towns and gather data on the percentage of people in each town who smoke, the percentage of people in each town who bike to work, and the percentage of people in each town who have heart disease.

Because you have two independent variables and one dependent variable, and all your variables are quantitative, you can use multiple linear regression to analyze the relationship between them.

$$\text{heart disease} = 15 + (-0.2*\text{biking}) + (0.178*\text{smoking})$$

Interpretation: 0.2% decrease in the frequency of heart disease for every 1% increase in biking, and a 0.178% increase in the frequency of heart disease for every 1% increase in smoking.