**Linear Regression:**

Linear regression is a **linear model**, e.g. a model that assumes a linear relationship between the input variables (x) and the single output variable (y). More specifically, that y can be calculated from a linear combination of the input variables (x).

When there is a single input variable (x), the method is referred to as **simple linear regression**. When there are **multiple input variables**, literature from statistics often refers to the method as multiple linear regression.

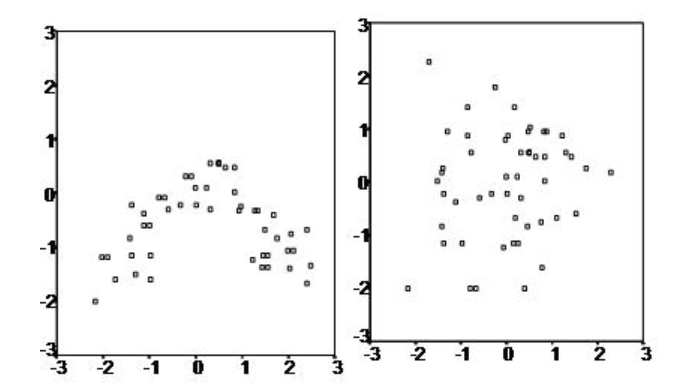
A linear regression line has an equation of the form *Y = a + bX*, where *X* is the explanatory variable and *Y* is the dependent variable. The slope of the line is *b*, and *a* is the intercept (the value of *y* when *x* = 0).

The regression has the following key assumptions:

* Linear relationship
* Multivariate normality
* No or little multicollinearity
* No autocorrelation

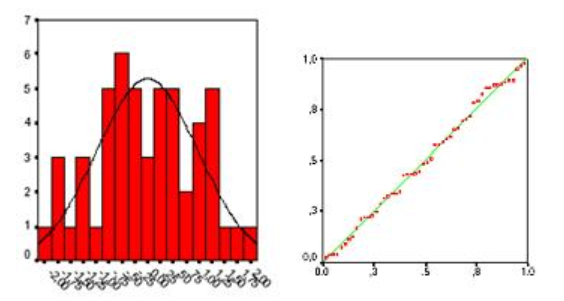
**Linear relationship:**

linear regression needs the relationship between the independent and dependent variables to be linear.  It is also important to check for outliers since linear regression is sensitive to outlier effects.  The linearity assumption can best be tested with scatter plots, the following two examples depict two cases, where no and little linearity is present.



**Multivariate normality:**

The linear regression analysis requires all variables to be multivariate normal.  This assumption can best be checked with a histogram or a Q-Q-Plot.  Normality can be checked with a goodness of fit test, e.g., the Kolmogorov-Smirnov test.  When the data is not normally distributed a non-linear transformation (e.g., log-transformation) might fix this issue.



**Histogram QQ Plot**

**No or little multicollinearity:**

Linear regression assumes that there is little or no multicollinearity in the data.  Multicollinearity occurs when the independent variables are too highly correlated with each other.

Multicollinearity may be tested with three central criteria:

1) Correlation matrix – when computing the matrix of Pearson’s Bivariate Correlation among all independent variables the correlation coefficients need to be smaller than 1.

2) Tolerance – the tolerance measures the influence of one independent variable on all other independent variables; the tolerance is calculated with an initial linear regression analysis.  Tolerance is defined as T = 1 – R² for these first step regression analysis.  With T < 0.1 there might be multicollinearity in the data and with T < 0.01 there certainly is.

3) Variance Inflation Factor (VIF) – the variance inflation factor of the linear regression is defined as VIF = 1/T. With VIF > 5 there is an indication that multicollinearity may be present; with VIF > 10 there is certainly multicollinearity among the variables.

If multicollinearity is found in the data, centering the data (that is deducting the mean of the variable from each score) might help to solve the problem.  However, the simplest way to address the problem is to remove independent variables with high VIF values.

4) Condition Index – the condition index is calculated using a factor analysis on the independent variables.  Values of 10-30 indicate a mediocre multicollinearity in the linear regression variables, values > 30 indicate strong multicollinearity.

**No autocorrelation:**

Linear regression analysis requires that there is little or no autocorrelation in the data.  Autocorrelation occurs when the residuals are not independent from each other.  In other words when the value of y(x+1) is not independent from the value of y(x).

While a scatterplot allows you to check for autocorrelations, you can test the linear regression model for autocorrelation with the Durbin-Watson test.  Durbin-Watson’s d tests the null hypothesis that the residuals are not linearly auto-correlated.  While d can assume values between 0 and 4, values around 2 indicate no autocorrelation.  As a rule of thumb values of 1.5 < d < 2.5 show that there is no autocorrelation in the data. However, the Durbin-Watson test only analyses linear autocorrelation and only between direct neighbours, which are first order effects.

**LOGISTIC REGRESSION:**

Logistic regression is named for the function used at the core of the method, the logistic function. Input values (x) are combined linearly using weights or coefficient values (referred to as the Greek capital letter Beta) to predict an output value (y). A key difference from linear regression is that the output value being modelled is a binary value (0 or 1) rather than a numeric value.

Below is an example logistic regression equation:

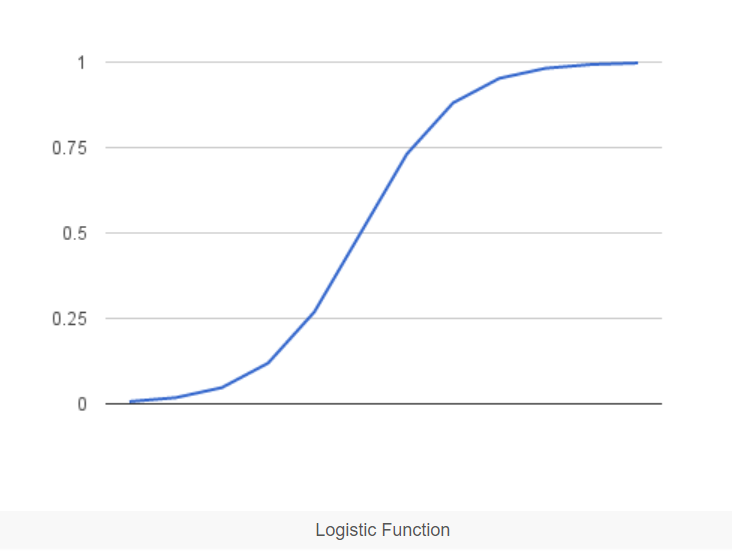
y = e^(b0 + b1\*x) / (1 + e^(b0 + b1\*x))

Where y is the predicted output, b0 is the bias or intercept term and b1 is the coefficient for the single input value (x). Each column in your input data has an associated b coefficient (a constant real value) that must be learned from your training data.

The actual representation of the model that you would store in memory or in a file are the coefficients in the equation (the beta value or b’s).

Assumptions:

1. Binary logistic regression requires the dependent variable to be binary and ordinal logistic regression requires the dependent variable to be ordinal.
2. Logistic regression requires the observations to be independent of each other.  In other words, the observations should not come from repeated measurements or matched data.
3. Logistic regression requires there to be little or no multicollinearity among the independent variables.  This means that the independent variables should not be too highly correlated with each other.
4. Logistic regression assumes linearity of independent variables and log odds.  although this analysis does not require the dependent and independent variables to be related linearly, it requires that the independent variables are linearly related to the log odds.
5. Logistic regression typically requires a large sample size.  A general guideline is that you need at minimum of 10 cases with the least frequent outcome for each independent variable in your model. For example, if you have 5 independent variables and the expected probability of your least frequent outcome is .10, then you would need a minimum sample size of 500 (10\*5 / .10).



**Logistic function**