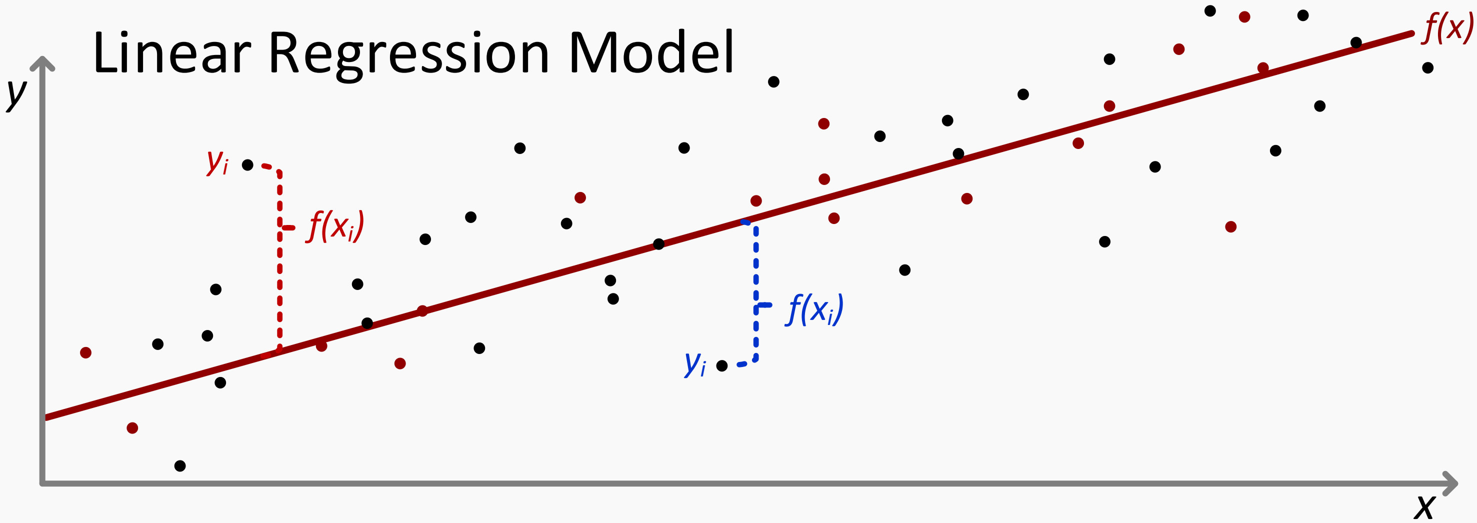
module2· regression

linear regression for machine learning

Linear Regression is used to predict real-valued outcomes (continuous variables)



The simplest form of **Linear Regression** is an equation that consists of a single linear coefficient and a constant fit the model data to a line; a function is applied in order to estimate $\boldsymbol{y}$ for each observation of $\rightarrow$ $\boldsymbol{f(x_i) = b_0 + b_1 x_i}$. In order to fit the best model possible to the data, the model needs to be adjusted in response to the error that can be measured.

Error in the above model can be measured as the distance from a point to the linear model prediction. However, there are two orientations of error as seen above: $\boldsymbol{y_i - f(x_i)}$ and $\boldsymbol{f(x_i) - y_i}$. This application is a fallacy considering that either way will not account for the error of the other. An alternative would be to use the absolute error $|\boldsymbol{f(x_i) - y_i}|$, however the latter remains inadequate. The solution to measure the model's error is the **Sum of Least Squares** error $\boldsymbol{(y_i - f(x_i))^2}$.

With the **SSE** applied to a linear model, the optimization objective of minimizing error is achieved.

Single Variable Linear Regression:

$$\boldsymbol{f(x_i) = b_0 + b_1 x_i}$$

Choose $\boldsymbol{b_0}$ and $\boldsymbol{b_1}$ to minimize the total error on the training set:

$$\mathrm{SSE}(f) = \sum_{i=1}^{n} \left(y_i - f(x_i)\right)^2 = \mathrm{SSE}(f) = \sum_{i=1}^{n} \left(y_i - (\boldsymbol{b_0 + b_1 x_i})\right)^2$$

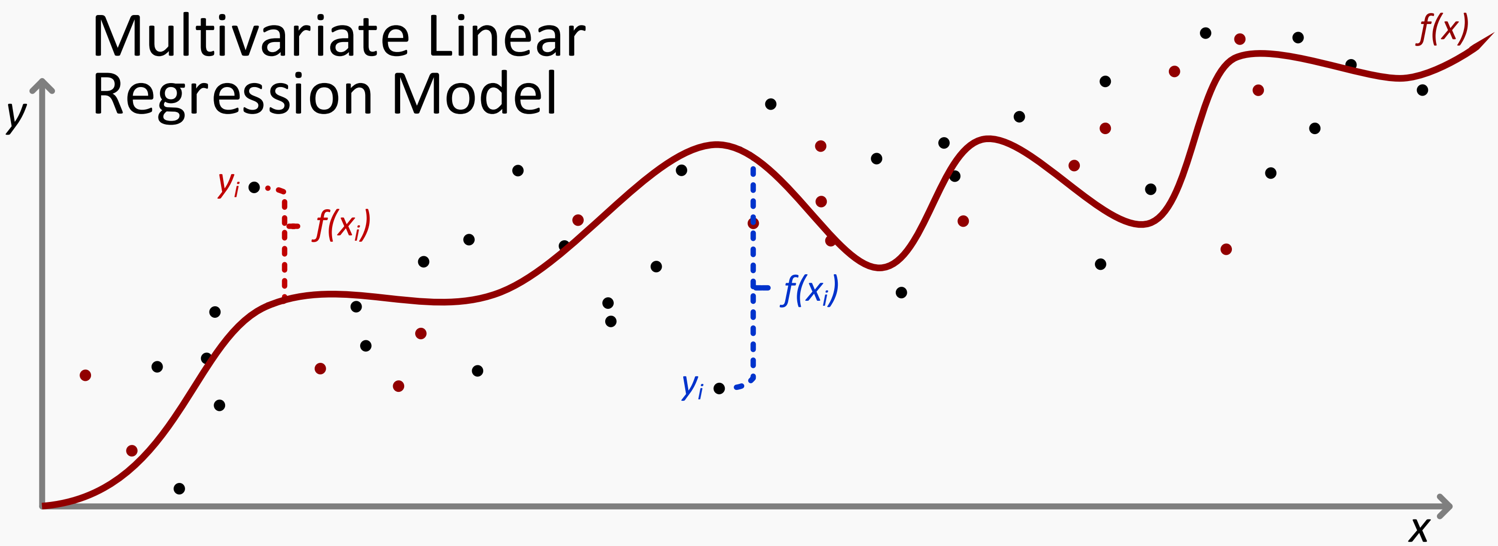
multivariate linear regression for machine learning

Multivariate Linear Regression expands upon simple linear regression when multiple weighted coefficients are added in perpetuity to the algorithm.

$$f(x_i) = b_0 + b_1 x_i \quad \rightarrow \quad \boldsymbol{f(x_i) = b_0 + b_1 x_{i,1} + b_2 x_{i,2} + \cdots + b_p x_{i,p}}$$

Method of Least Squares choosing coefficients to minimize:

$$\text{SSE}(f) = \sum_{i=1}^{n} \left(y_i - f(x_i)\right)^2 = \text{SSE}(f) = \sum_{i=1}^{n} \left(y_i - \left(\boldsymbol{b_0 + b_1 x_{i,1} + b_2 x_{i,2} + \cdots + b_p x_{i,p}}\right)\right)^2$$



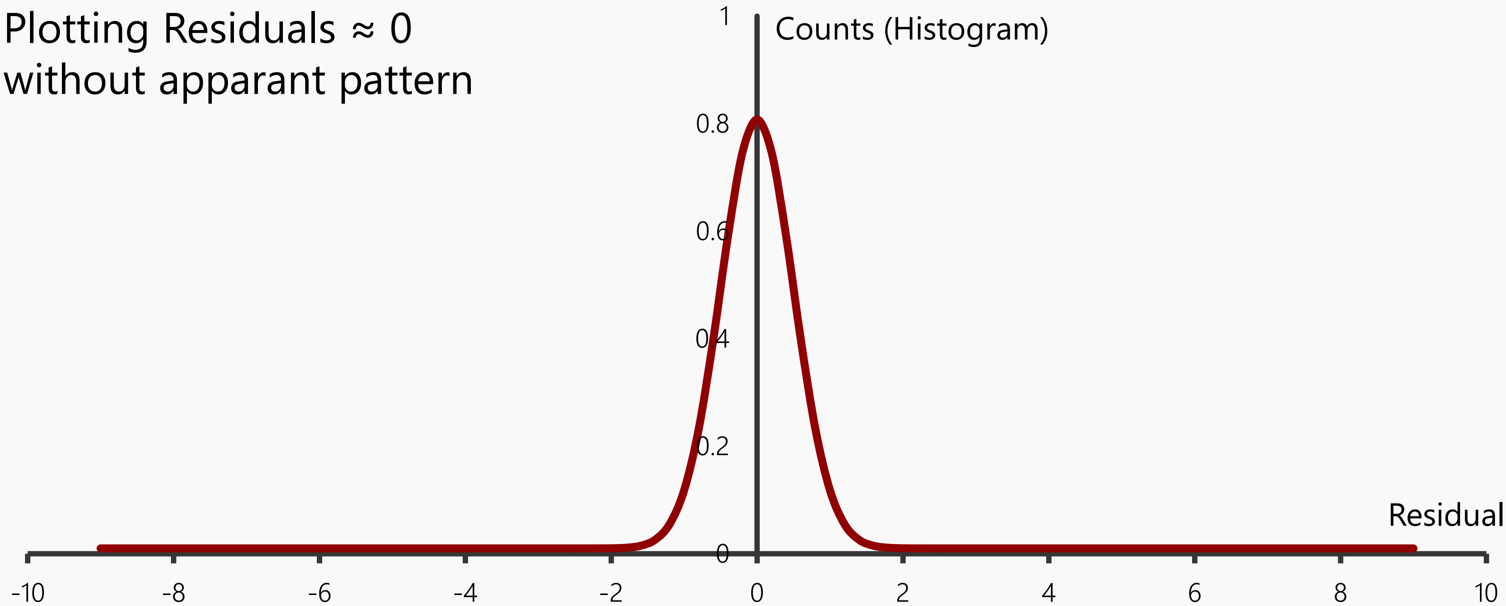
Multivariate Linear Regressive models can fit data in a much more dynamic manner as seen above. This is achieved through the use of polynomial, quadratic, trigonometric, and other various functions.

evaluation regression models

The **Residual** or **Error** is the difference between the predicted and actual values $\boldsymbol{f(x_i) - y_i}$.

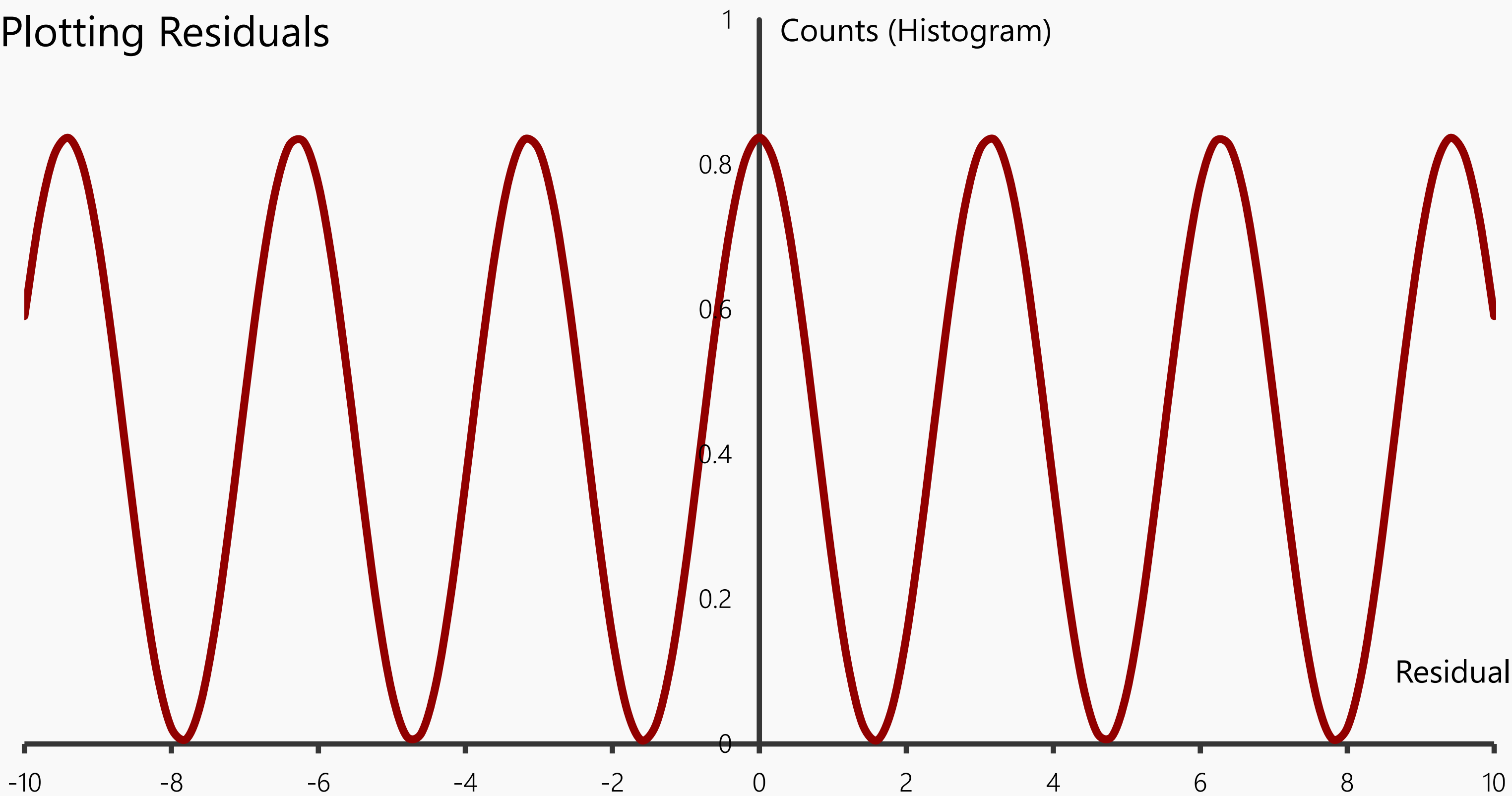
The reported Residuals are ideally ≈ 0 and will lack any specific structure or pattern amongst them:

Plotting Residuals ≈ 0
without apparant pattern



Plotted Residuals displaying an apparent pattern indicates missed properties in modeling the data:

Plotting Residuals



The above model could result from multiple linear features modeled inappropriately causing signal.

Plotted examples displaying a skew in either direction indicates a generally poorly performing model: