Statistical Equations

A guide by E. Walker (2022)

Below is a tabulated overview of central equations utilized in statistics. Each entry comprises of a full equation, and a summary.

|  |  |
| --- | --- |
| Equation | Explanation |
|  | A linear regressive model used to simulate phenomena which exhibit a linear relationship. occurs as a linear model which considers input variables from the environment (). Similar to , these models comprise of a constant y-intercept (), but one gradient for *each* environment variable (). This formula takes into consideration environment variables, and “dummy variables” – which are numeric representations of categorical data. |
|  | A polynomial regressive model used to simulate phenomena which exhibit a polynomial relationship. A quadratic regressive model would be an expression of degree , trinomial , etc. A standard linear regressive model would be degree . This formula considers environment variables. |
|  | A logistic regressive model used in non-binary (e.g., ordinal) classification. Outputs the probability of an environment variable yielding a response (a classification) .  Multivariate Logistic Regression (MLR) is also given for environment variables, in addition to Multiclass Logistic Regression (MCLR) for classes. |
|  | Sigmoid function. Exhibits an s-shaped trend. |
|  | Arithmetic mean of a collection of values. |
|  | Reductible error, given by the difference between the actual environmental response and our model . |
|  | Min-max normalization to ensure that all values within a collection reside in the same range/neighborhood of values between , , or an arbitrary range . |
|  | Mean scaling. Including the *standardization* variant which divides by the set’s standard deviation .  The z-statistic is also defined by the standardized form. |
|  | Mean Absolute Deviation (MAD). |
|  | Bowley Coefficient of Skewness. |
|  | Standard Deviation (SD), equal to the square root of **variance**. |
|  | Standard Error (SE) for observations of a population . |
|  | Mean Squared Error (MSE). Also applied as a cost (loss) function in Artificial Neural Networks. |
|  | Residual Sum of Squares (RSS). |
|  | Total Sum of Squares (TSS). |
|  | The t-statistic (or t-student test). Used when given a normally distributed population. |
|  | The F-statistic, calculated as the ratio between the *Mean Square Between (MSB)* and the Mean Squared Error (MSE). |
| *or* | The -statistic measures how well a model represents its environment. |
|  | Confidence interval of utilizing the sample mean , variance and an unknown mean . |
|  | Bayes Theorem considers the probability of an element belonging to class given a condition .  The prior probability , marginal likelihood , and the likelihood are utilized to determine the posterior probability . |
|  | Pearson’s correlation coefficient for continuous data. |
|  | Kendall’s (tau) rank correlation coefficient for categorical data. Suited for ordinal relationships, and comes in three forms – tau A, tau B, and tau C. |
|  | Spearman’s (rho) rank correlation coefficient for categorical data. |
|  | Bowley’s Coefficient of Skewness. |
|  | Weighted sum of variables with a given bias. |
|  | Threshold Activator Function. Included in Threshold Logic Units (TLUs). |
|  | Sigmoid Activator Function. |
|  | Rectifier Activator Function. |
|  | Hyperbolic Tangent Activator Function. |
|  | Softmax Activator Function for classes. |
|  | Identity activator (no alteration to input data) |
|  | Forward pass (neural network backpropagation). |
|  | Reverse pass (neural network backpropagation). |
|  | Ridge regression and ridge regression cost function. |
|  | Elastic net regression and elastic net cost function. |
|  | Lasso regression and lasso regression cost function. |
|  | Logistic regression cost function. |
|  | Softmax regression (logistic regression for classes) and softmax cost function. |
|  | Batch Gradient Descent. |
|  | Neural Network Matrix Form. |
|  |  |
|  |  |
|  |  |
|  |  |