|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Univariate statistics:**  Discover associations between a variable of interest and potential predictors. It is strongly recommended to start with simple univariate methods before moving to complex multivariate predictors.  Most of the univariate statistics are based on the linear model which is one of the main model in machine learning  **LINEAR MODEL:**  Given the n random samples () the linear regression models the relation between the observations and the independent variables is  The are the model parameters, the regression coefficients  is the intercept or the bias  are the residuals  **An independent variable .** It is a variable that stands alone and isn’t changed by the other variables you are trying to measure. For example, someone’s age might be an independent variable. Other factors (such as what they eat, how much they go to school, how much television they watch) aren’t going to change a person’s age. In fact, when you are looking for some kind of relationship between variables you are trying to see if the independent variable causes some kind of change in the other variables, or dependent variables. In Machine Learning, these variables are also called the **predictors**.  **A dependent variable**. It is something that depends on other factors. For example, a test score could be a dependent variable because it could change depending on several factors such as how much you studied, how much sleep you got the night before you took the test, or even how hungry you were when you took it. Usually when you are looking for a relationship between two things you are trying to find out what makes the dependent variable change the way it does. In Machine Learning this variable is called a **target variable**.  Step **1**: model the data on some hypothesis, for example: salary is a linear function of the experience  the slope or coefficients or parameter of the model  the intercept or bias is the second parameter of the model  the i-th error, or residual with  Step **2**: fit: estimate model parameters. The goal is to estimate  Minimises the mean squared error MSE/Sum squared error SSE/Ordinary Least Squares OLS  **F-TEST:**  **Goodness of fit:** of a statistical model describes how well it fits a set of observations. Measures of goodness of for typically summarizes the discrepancy between observed values and the values under the model in equation. We will consider the explained variance also known as the co-efficient of determination, denoted  The total sum of squares is the sum of squares explained by the regression, plus the sum of squares of residuals unexplained by the regression, also called SSE such that  The mean of y  The total sum of squares, also called the total squared sum of deviations from the mean : | The regression sum of squares, also called the explained sum of squares:  is the estimated value of given a value of experience  The sum of squares of residuals, also called the residual sum of squares RSS is:  is the explained sum of squares of errors. It is the variance by the regression divided by the total variance  **Test:** Let be an estimator of the variance of . The two in the denominator stems from the two estimated parameters: intercept and coefficient.      The single degree of freedom comes from the difference between  The fisher statistics of the ratio of two variances:  Using the F-distribution, compute the probability of observing a value greater than F under : the survival function (1Cumulative distribution function) at of the given F-distribution  **MULTIPLE REGRESSION:**  Multiple linear regression is the most basic supervised learning algorithm.  Given a set of regression, we assume the model that generates the data involves only a linear combination of the input variables.  Extending each sample with an intercept allows us to use a more general notation based on linear algebra and write it as a simple dot product:  Where is a vector of weights that defined the parameters of the model. From now we have P and the intercept.  Minimise the Mean Squared Error MSE loss:  be an metric of N samples of P inputs features with one column of one and let be be a vector of the N targets. The mean squared error MSE loss is  The that minimise the MSE can be found by:  **MULTIPLE REGRESSION categorical independent variable/factor:** **ANALYSIS of COVARIANCE (ANCOVA)**  Analysis of covariance (ANCOVA) is a linear model that blends ANOVA and linear regression. ANCOVA evaluates whether population means of a dependent variable (DV) are equal across levels of a categorical independent variable (IV) often called a treatment, while statistically controlling for the effects of other quantitative or continuous variables that are not of primary interest, known as covariates (CV). | **One way AN(C)OVA:**  ANOVA: one categorical independent variable, one factor  ANCOVA: ANOVA with some co-variates.  **Two way AN(C)OVA:**  with two categorical independent variables, two factors.  **MULTIPLE COMPARISONS:**  Note that under the null hypothesis the distribution of the p-values is uniform.  **Statistical measures:**  True Positive (TP) equivalent to a hit. The test correctly concludes the presence of an effect.  True Negative (TN). The test correctly concludes the absence of an effect.  False Positive (FP)equivalent to a false alarm, Type I error. The test improperly concludes the presence of an effect. at < 0.05 leads to FP. False Negative (FN)equivalent to a miss, Type II error. The test improperly concludes the absence of an effect.  **correction for multiple comparisons**  The correction is based on the idea that if an experimenter is testing 𝑃 hypotheses, then one way of maintaining the familywise error rate (FWER) is to test each individual hypothesis at a statistical significance level of 1/𝑃 times the desired maximum overall level.  So, if the desired significance level for the whole family of tests is 𝛼 (usually 0.05), then the correction would test each individual hypothesis at a significance level of 𝛼/𝑃. For example, if a trial is testing 𝑃 = 8 hypotheses with a desired 𝛼 = 0.05, then the correction would test each individual hypothesis at 𝛼 = 0.05/8 = 0.00625.  **The False discovery rate (FDR) correction for multiple comparisons**  FDR-controlling procedures are designed to control the expected proportion of rejected null hypotheses that were incorrect rejections (“false discoveries”). FDR-controlling procedures provide less stringent control of Type I errors compared to the familywise error rate (FWER) controlling procedures (such as the correction), which control the probability of at least one Type I error. Thus, FDR-controlling procedures have greater power, at the cost of increased rates of Type I errors.  **Brain volumes study**  The study provides the brain volumes of grey matter (gm), white matter and cerebrospinal fluid) of 808 anatomical MRI scans.  1. Fetch demographic data demo.csv and tissue volume data (gm.csv, wm.csv, csf.csv).  2. Merge tables.  3. Compute Total Intra-cranial volume.  4. Compute tissue ratios:  5. Descriptive analysis per site in excel file.  6. Visualize site effect of gm ratio using violin plot:  .  7. Visualize age effect of gm ratio using scatter plot:  8. Linear model ():  **Multivariate statistics:**  Multivariate statistics include all statistical techniques for analysing samples made of two or more variables. The data set is a collection of independent samples column vectors of length P |  |  |  |