A note on Statistical Methods (Initial draft)

Regression: It gives the relationship between a dependent and one or more independent variables. Regression analysis is primarily used for two conceptually distinct purposes. First, regression analysis is widely used for prediction and forecasting, where its use has substantial overlap with the field of machine learning. Second, in some situations, regression analysis can be used to infer causal relationships between the independent and dependent variables. *(Source: Wikipedia)*

Linear regression: It is a linear approach to modeling the relationship between a scalar response and one or more independent variables. The case of one explanatory variable is called simple linear regression; for more than one, the process is called multiple linear regression. *(Source: Wikipedia)*

Multivariate linear regression: Here correlated dependent variables are predicted, rather than a single scalar variable as in multiple linear regression. *(Source: Wikipedia)*

Logistic regression: (Logit model) It is a statistical model that in its basic form uses a logistic function to model a binary dependent variable. It is estimating the parameters of a logistic model (a form of binary regression). Mathematically, a binary logistic model has a dependent variable with two possible values, such as pass/fail which is represented by an indicator variable, where the two values are labeled "0" and "1". In the logistic model, the log-odds (the logarithm of the odds) for the value labeled "1" is a linear combination of one or more independent variables ("predictors"); the independent variables can each be a binary variable (two classes, coded by an indicator variable) or a continuous variable (any real value). The corresponding probability of the value labeled "1" can vary between 0 (certainly the value "0") and 1 (certainly the value "1"), hence the labeling; the function that converts log-odds to probability is the logistic function, hence the name. The unit of measurement for the log-odds scale is called a logit, from the logistic unit. *(Source: Wikipedia)*

Multinomial logistic regression: It is a model that is used to predict the probabilities of the different possible outcomes of a categorically distributed dependent variable, given a set of independent variables. It is a classification method that generalizes logistic regression to multiclass problems, i.e. with more than two possible discrete outcomes. *(Source: Wikipedia)*

Probit function: Probit regression, also called a Probit model, is used to model dichotomous or binary outcome variables. In the probit model, the inverse standard normal distribution of the probability is modeled as a linear combination of the predictors. *(Source: https://stats.idre.ucla.edu/stata/dae/probit-regression/)*. The word “probit” is a combination of the

words probability and unit; the probit model estimates the probability a value will fall into one of the two possible binary (i.e. unit) outcomes.

Tobit: The Tobit models are a family of statistical regression models that describe the relationship between a censored (or truncated, in an even broader sense of this family) continuous dependent variable y_i and a vector of independent variables x_i. The model was originally proposed by James Tobin (1958) to model nonnegative continuous variables with several observations taking value 0 (household expenditure). *(Source: https://link.springer.com/referenceworkentry/10.1007%2F978-94-007-0753-5_3025)*

Data cleaning and manipulation: (Data cleansing) It is the process of detecting and correcting (or removing) corrupt or inaccurate records from a recordset, table, or database and refers to identifying incomplete, incorrect, inaccurate, or irrelevant parts of the data and then replacing, modifying, or deleting the dirty or coarse data. Data cleansing may be performed interactively with data wrangling tools, or as batch processing through scripting. *(Source: Wikipedia)*

Panel data analysis: The data are usually collected over time and over the same individuals and then a regression is run over these two dimensions. Multidimensional analysis is an econometric method in which data are collected over more than two dimensions (typically, time, individuals, and some third dimension)

Conjoint analysis: It is a survey-based statistical technique used in market research that helps determine how people value different attributes (feature, function, benefits) that make up an individual product or service. The objective of conjoint analysis is to determine what combination of a limited number of attributes is most influential on respondent choice or decision making. A controlled set of potential products or services is shown to survey respondents and by analyzing how they make choices among these products, the implicit valuation of the individual elements making up the product or service can be determined.

Time series analysis: A time-series data or time series is a series of data points indexed (or listed or graphed) in time order. Most commonly, a time series is a sequence taken at successive equally spaced points in time. Thus it is a sequence of discrete-time data. Examples of time series are heights of ocean tides, counts of sunspots, and the daily closing value of the Dow Jones Industrial Average. Time series analysis is a statistical technique that deals with time-series data, or trend analysis. *(Source: Wikipedia)*