Multivariate Analytics

DATA 514 - Unit 1

Danielle Senechal

September 10th, 2020

1. **Find and describe R implementations of at least two tests for assessing multivariate normality. Describe also those tests.**

One R implementation of a test for assessing multivariate normality is normal probability plots and quantile-quantile plots. A normal probability and quantile-quantile plots observe continuous variables to determine if there is normality. Multivariate normality can be determined from these plots if the points fall into a straight line. If the points do not form a straight line, but instead stray from the line in a systemic fashion, the sample being assessed can be deemed not normally distributed [1, p. 4]. The R function that can produce these plots is *mqqnorm,* which is installed from the *RVAideMemoire* package. The function is used as: *mqqnorm(x)*, where *x* is the data frame being used [2]. A multi-normal Q-Q plot is then drawn, and multivariate normality can then be inferred based on if the points fall into a straight line.

Another test for assessing multivariate normality is by using the Mardia’s statistic. This statistic measures how skewed and how peaked the distribution is (the kurtosis). When using the Mardia’s Multivariate Normality test, the reported p-value should be above the significance level of 0.05. A p-value below 0.05 suggests the sample is not normally distributed. To test the Mardia’s statistic in R, the *mardiaTest* command can be used. This command is from the *MVN* package and is used in the form: *mardiaTest(x)*, where *x* represents a data frame. The *qqplot = True* option can also be added to have R create a chi-squared Q-Q plot of the data [3]. Depending on the version of R being used, using the *mardiaTest* command may prompt R to output a message that asks for the *mvn* command to be used instead. The format for this command is: *mvn(x, mvnTest = c("mardia"))*, where *x* is the data and *mvnTest* specifies that the test being performed, which in this case is the Mardia’s test [4]. The output from these commands includes the p-value, which should fall above 0.05 so multivariate normality can be confirmed.

1. **Provide a short but informative description of a regression method that is often used for classification.**

A regression method that can often be used as a classification method is Logistic Regression. This method is widely believed to be a classification method because Logistic Regression can determine between two categories, which makes it considered to be a binary classification model [5]. An example of this would be determining if a subject is young or old. This kind of data is continuous, but the value of the observation determines what category the subject falls in. Logistic Regression uses zeros and ones to categorize the observations, and then a graph is made with values between zero and one on the y-axis. This can be seen in **Graph 1** which is shown below. The observation is either a zero or one based on where it falls on the plot. If the point is closer to the top of the graph (near a one), the observation is categorized as the category corresponding to one, and same for a point closer to the bottom of the graph (near zero) [6]. Logistic Regression is a regression model because the graph plots the quantitative data, but when a decision rule is made, this method can be used as a binary classification model to give each value a category.

An example of a Logistic Regression graph can be seen below in **Graph 1**. The line going through the data point is continuous, because the data is quantitative, not qualitive. It can also be seen that a clear distinction between the categories can been seen, as the points cluster in the top and bottom. With the assignment of two categories, each point can be grouped into the corresponding category.

A screenshot of a map

Description automatically generated

**Graph 1:** Example graph of Logistic Regression [5]

1. **Find a published journal paper reporting a predictive modeling experiment (either for classification or regression) that follows the misconception of performing unsupervised dimensionality reduction as a data preparation step for supervised modeling. Write a critical evaluation of the modeling process described in the paper. Provide also a pdf version of the paper.**

“Acute Coronary Syndrome Prediction Using Data Mining Techniques- An Application” by Tahseen A. Jilani, Huda Yasin, Madiha Yasin, and Cemal Ardil describes an experiment done to predict Acute Coronary Syndrome using Logistic Regression. The article discusses the data mining techniques used in order to reduce the size of their datasets. This data set is described by the authors as “noisy”, and they propose to use Principal Component Analysis (PCA) as a method to reduce the size of the data before performing the Logistic Regression. The authors then state that they performed PCA on ten independent variables and found that eight of the principle components were responsible for the majority of the variability. In the results section of the paper, they state that before they fit the model, they used a data reduction technique (PCA) to reduce the number of dimensions of the data [7]. This statement is a misconception, because PCA is an unsupervised dimensionality reduction technique, not a data reduction technique.

There are multiple occasions in the article when PCA is misconceived as a data preprocessing or data reduction technique. Data reduction is one of the methods of data preprocessing, which can also be called data preparation. Principle component analysis is an unsupervised dimensionality reduction technique, not a data reduction technique. PCA compresses the data and can reduce computation time and redundancy, which can easily be mistaken for a data preparation step [8]. The authors of this paper misconceive using PCA as a data reduction technique instead of a dimensionality reduction method. They accurately describe the methods behind using PCA, but they are unaware of the they made when describing how PCA was used.

1. **Find two different (but reasonable) definitions of parametric and nonparametric learning algorithms. Describe and discuss the definitions. Find and discuss at least one supervised learning algorithm that is classified differently according to these definitions.**

A different definition of a parametric learning algorithm is that it has a fixed number of parameters and in these methods the distribution is known to be normal [9]. Another definition for a non-parametric model is that they use a flexible number of parameters, and these methods are independent of the population and are considered distribution-free [9].

Decision trees are nonparametric because they make no distribution assumptions when they are supervised. They also have unsupervised learning uses as well, because they can have a target variable or can be used for clustering [10]. Although, when a decision tree is used for clustering, it tries to identify clusters by assuming that the given data is associated with fabricated data that comes from a uniform distribution [11]. This contradicts that this is a nonparametric algorithm based on the above definition because the definition states that nonparametric algorithms do not make any distributional assumptions. Therefore, based on the above definitions, if a decision tree is supervised, it should be considered nonparametric, and if it is unsupervised, it should instead be parametric because of the distributional assumption made.

1. **Find and describe at least one R function that can be used to screen multivariate data for multicollinearity. Describe also the screening method.**

One R function that can be used to screen multivariate data for multicollinearity is the Variance Inflation Factor (VIF) statistic. This function comes from the *car* package in R, and is used as: *vif(x)*, where *x* represents a linear or logistic regression model [12]. This statistic looks at a multiple regression model and searches for multicollinearity among all of the independent variables. If the calculated VIF statistic is high, it can be determined that there is some collinearity between variables [13]. A VIF statistic is considered high and problematic if it is above the five to ten range [12].

To calculate this statistic, the r-squared value is needed. The r-squared value is subtracted from one and then that value is divided by one to produce the VIF statistic [12]. The output produced by running this command shows all the variables in the dataset and the values of their VIF statistics. To provide an example of the output, a multiple regression was performed on a data set of baseball statistics. This multivariate regression was performed as an assignment for a Data Analytics class. The vif command was used, and the output was as follows:

> vif(reg1) # VIF stats

G RBI SO yearID

4.536461 6.480098 6.048830 1.028254

In this example, it can be seen that the VIF associated with RBIs and strikeouts are above five, so multicollinearity can be detected in this model and can be considered problematic. Further steps should be taken to reduce multicollinearity in this model.

References:

[1] “4.4 - Multivariate Normality and Outliers | STAT 505,” *PennState: Statistics Online Courses*. https://online.stat.psu.edu/stat505/lesson/4/4.4 (accessed Sep. 08, 2020).

[2] “mqqnorm function | R Documentation.” https://www.rdocumentation.org/packages/RVAideMemoire/versions/0.9-77/topics/mqqnorm (accessed Sep. 08, 2020).

[3] “mardiaTest function | R Documentation.” https://www.rdocumentation.org/packages/MVN/versions/4.0/topics/mardiaTest (accessed Sep. 08, 2020).

[4] S. Korkmaz, D. Goksuluk, and G. Zararsiz, “MVN: An R Package for Assessing Multivariate Normality,” *R J.*, vol. 6, no. 2, p. 151, 2014, doi: 10.32614/RJ-2014-031.

[5] “Logistic Regression Explained. [ — Logistic Regression explained… | by Jaime Zornoza | Towards Data Science.” https://towardsdatascience.com/logistic-regression-explained-9ee73cede081 (accessed Sep. 08, 2020).

[6] “Understanding Logistic Regression - GeeksforGeeks.” https://www.geeksforgeeks.org/understanding-logistic-regression/ (accessed Sep. 08, 2020).

[7] T. A. Jilani, H. Yasin, M. Yasin, and C. Ardil, “Ardil C. Acute coronary syndrome prediction using data mining techniques: an application.”

[8] “Introduction to Dimensionality Reduction,” *GeeksforGeeks*, Jun. 01, 2017. https://www.geeksforgeeks.org/dimensionality-reduction/ (accessed Sep. 10, 2020).

[9] “Difference between Parametric and Non-Parametric Methods,” *GeeksforGeeks*, Feb. 08, 2020. https://www.geeksforgeeks.org/difference-between-parametric-and-non-parametric-methods/ (accessed Sep. 09, 2020).

[10] “Why decision trees is the best data mining algorithm | Mixotricha.” https://zyxo.wordpress.com/2010/09/17/why-decision-trees-is-the-best-data-mining-algorithm/ (accessed Sep. 09, 2020).

[11] A. Singh, “Can decision trees implement both supervised and unsupervised learning?,” *Medium*, Feb. 26, 2020. https://medium.com/@anupamasingh\_12727/can-decision-trees-implement-both-supervised-and-unsupervised-learning-f4dd3cad2e41 (accessed Sep. 09, 2020).

[12] “VIF function | R Documentation.” https://www.rdocumentation.org/packages/regclass/versions/1.6/topics/VIF (accessed Sep. 10, 2020).

[13] J. Chappelow, “Variance Inflation Factor Definition,” *Investopedia*. https://www.investopedia.com/terms/v/variance-inflation-factor.asp (accessed Sep. 10, 2020).