Can The National Benchmark Tests Be Used To Predict Academic Performance in Statistics 101?

A project submitted in partial fulfillment of the requirements for the degree of

Bachelor of Science Honours

in

Mathematical Statistics

Department of Statistics

Rhodes University

by

Nonhlanhla Linda Luphade

September 2019

Supervisor: Jeremy Baxter

Abstract

With the increasing diversity of students attending university there is a growing interest in the factors predicting academic performance. This study investigates whether the National Benchmark Tests, comprising the Academic Literacy, Quantitative Literacy and Mathematics tests, can be used to predict academic performance of the first year Rhodes University students registered for Statistics 101 test 1. Quantitative approaches, namely regression and neural networks, were proposed for this study. The result of the analysis is that the academic performance can be predicted but these models have very low predictive power. These models clearly indicated that other factors should be identified.

Keywords: **National Benchmark Tests, Statistics 101, Academic Performance, Prediction.**

Contents

List of Tables

List of Abbreviations

SC Senior Certificate

NSC National Senior Certificate

NBT National Benchmark Test

EDP Extended Degree Program

HESA Higher Education South Africa

SAT Scholastic Aptitude Test

ATAR Australian Tertiary Admission Rank

SAUVCA South African Universities Vice-Chancellors' Association

STA101 Statistics 101

EDM Educational Data Mining

GLM Generalised Linear Model

VIF Variance Inflation Factor

NN Neural Network

SW Shapiro-Wilk

FFNN Feed Forward Neural Network

MLP Multilayer Perceptron

Acknowledgements

Foremost, I would like to express my sincere gratitude to my supervisor Mr Jeremy Baxter for the continuous support in my honours thesis, for his patience, motivation, enthusiasm, and immense knowledge. His guidance helped me in all the time of research and writing of this thesis. I could not have imagined having a better supervisor. Besides my supervisor, I would like to thank the staff of the Department of Statistics at Rhodes University for their great support. My sincere thanks also goes to my family: Luphathe Nyathi, Lindiwe Nyathi, Sandisile Ncube, Thapelo Nyathi and Gareth Ndlovu for all the love and support throughout my life. I thank my fellow Statistics classmates for the stimulating discussions, for the sleepless nights we were working together before deadlines and for all the fun we have had in the last four years. Also I thank all my friends: Joregina Mthembu, Tavonga Mandava, Fadzai Matapuri and Chantell Murembeni. Most importantly I would like to thank God for the strength and guidance.

Chapter 1

The National Benchmark Test

1.1 Throughput in Higher Education in South Africa

Many higher education systems across the globe are struggling with the challenges of low throughput and high dropout rates (Moodley, P. and Singh, R. J., 2015). It is estimated that in South African higher education, 27% of entering students graduate in minimum time, that is the time allocated for the degree to be completed (Neethling, L. , 2015). 55% of those entering students never graduate and 40% of the graduates take up to two years more than the minimum time set for their degree studies (Prince, R., 2017). Academic support programs like the Extended Degree Programme (EDP) were introduced to assist students with proven potential but without adequate schooling to master their degree programmes. An additional year of study is added to a mainstream degree programme to form an EDP (Van Schalkwyk, S. and Bitzer, Eli and Van der Walt, C., 2010). This intervention was designed in the 1980s (**?**) to offer less prepared students access to university and to try facilitate their success. EDPs have been successful and hence most universities in South Africa are now offering some form of foundation or extended programme (Van Schalkwyk, S. and Bitzer, Eli and Van der Walt, C., 2010). The challenge is how best to identify students that would benefit from EDPs. In South Africa there are two assessments that contribute to this purpose. The National Senior Certificate (NSC), a certificate obtained after passing the Grade 12 exams (Spaull, N., 2013), is a requirement for university entry. The National Benchmark Tests (NBTs) are designed to help the universities make decisions about the most appropriate curriculum structures for students.

Scott, I. and Yeld, N. and Hendry, J. (2007) discuss increasing concerns about the meaning of the National Senior Certificate (NSC) results and the suitability of these results as a means of determining who qualifies and is ready for university entrance in South Africa. In 2008 the Senior Certificate, the certificate obtained after passing the Grade 12 exams (Scott, I. and Yeld, N. and Hendry, J., 2007), was changed to the NSC. This caused problems in the South African higher education sector, for example under-preparedness of the first-year students (Wilson-Strydom, M., 2012). These problems were highlighted by the poor throughput and

graduation rates. Students were not finishing the programs within the set time and in many cases the students dropped out. One of the conclusions drawn by Scott, I. and Yeld, N. and Hendry, J. (2007) was that the higher education sector should focus more on teaching and learning effectiveness. Higher Education South Africa (HESA) decided to introduce the NBTs as a means to gain extra information about the prospective students in an effort to assist the universities in placing the prospective students into appropriate programs.

1.2 The National Benchmark Tests

The National Benchmark Tests (NBT) are assessments for first year applicants into higher education institutions in South Africa. The NBTs are used to measure the applicant's understanding of Academic Literacy (AL), Quantitative Literacy (QL) and Mathematics (MAT) related to the demands of tertiary coursework (CETAP, 2018). The tests aim to address the following question: What are the academic literacy, quantitative literacy and mathematics levels of proficiencies of the school-leaving population who wish to continue with higher education, at the point prior to their entry into higher education. In essence these tests are an attempt to assess if it can be reasonably expected that these students will cope with the demands of higher education study (A. A. Wadee and A. Cliff, 2016).

In other countries similar tests are used for admissions to universities or colleges. The United Kingdom uses a similar testing system known as the Australian Tertiary Admission Rank (ATAR) (Wurf, G. and Croft-Piggin, L., 2015). ATAR is an aptitude test for school leavers. The United States of America utilise the Scholastic Aptitude Test (SAT) (McCormick, R. E. and Tinsley, M. , 1987). China has the National College Entrance Examination.

1.3 History of the National Benchmark Tests

Yeld (2007) suggests that there was widespread consensus in the higher education sector that the academic literacy of the students entering university was inadequate. This led to the National Benchmark Test Project (NBTP) being established in 2004 by the South African Universities Vice-Chancellors' Association (SAUVCA) and the Committee of Technikom Principals. Vice-Chancellor O'Connell emphasized the significance of the NBTP as follows:

> The NBT project represents an attempt to provide both schooling and higher education with important information on the competencies of their exiting (in the case of schools) and entering (in the case of universities) students: information that does not duplicate the essential information delivered by the school-leaving examination, but that provides an important extra dimension (Yeld, 2007).

The overall aim of the NBTP was to attempt to understand the applicants AL, QL and MAT components in order to assist the admission staff in making decisions about selection and placement. In addition this information could be used in determining appropriate curriculum interventions (Yeld, 2007).

In 2008 the Senior Certificate (SC) was replaced by the National Senior Certificate (NSC) which created discontinuity and doubt of credibility of the school leaving marks particularly in Mathematics (Yeld, 2007; Mahlobo, R., 2015). Parallel to implementation of the NSC, the NBT were commissioned by the Higher Education South Africa (HESA) now known as Universities South Africa (USAf) (Yeld, 2007; Mahlobo, R., 2015). The NBTs were designed to gain a better understanding of the university applicants (Cliff, 2015).

1.4 Purpose of the National Benchmark Tests

The NBT were primarily designed as a placement measure and to supplement the National Senior Certificate (NSC) results (CETAP, 2018). It was intended that the information would assist higher education institutions to understand the applicant and to provide appropriate curricular provision such as Extended Degree Programmes (EDPs). This extra applicant information was designed to assist institutions in accurately placing applicants entering higher education (Yeld, 2007). Although the NBT were initially created for placement, they are used in other universities for selection and placement in certain degree programmes such as medicine and law (A. A. Wadee and A. Cliff, 2016). In cases where the student performs badly in the school leaving assessments and performs well in the NBTs, the NBTs can be used for selection (CUT, 2015). NBT results are used by many universities as an extra admission criterion for applicants to their institutions (CUT, 2015).

1.5 How the National Benchmark Tests Work

The NBTs consist of two multiple choice tests each of three-hours duration. A total of two papers are written, consisting of AL and QL in paper 1 and MAT in paper 2. The NBT tests are managed by the Alternative Admissions Research Project at the University of Cape Town (CETAP, 2018).

1.6 The National Benchmark Test Domain

1.6.1 The Academic Literacy Test

The Academic Literacy (AL) test is designed to assess the ability of first-year students to cope with the typical language-of-instruction, academic reading and reasoning demands that they will likely encounter on entry to higher education (CETAP, 2018). The AL test is

further divided into nine subdomains, namely AL Cohesion, AL Communicative Function, AL Discourse, AL essential, AL Grammar, AL Inference, AL Metaphor, AL Test Genre and AL Vocabulary. The AL test was developed to assess the writer's ability to:

- Read carefully and make meaning from texts that are typical of the kinds that they will encounter in their studies;
- Derive meaningful words in context, knowledge and acquisition of vocabulary;
- Recognize, formulate and critique context and develop a logical academic argument.

1.6.2 The Quantitative Literacy Test

The Quantitative Literacy (QL) test is designed to test the writer's ability to manage situations and solve real world problems in a real context that is relevant to their higher education studies (CETAP, 2018). The QL test is divided into 6 subdomains, namely QL C, QL D, QL P, QL Q, QL R and QL S. The QL test is developed to assess the writer's ability to:

- Identify trends and patterns in various situations;
- Reason logically;
- Understand basic numerical concepts and information used in text and do basic numerical manipulations;
- Competently interpret quantitative information.

1.6.3 The Cognitive Academic Mathematics Proficiency Test (CAMP)

This test aims to assess the writer's ability related to mathematical concepts regarded as part of the secondary school curriculum (NSC Mathematics Paper 1 and 2) relevant for higher education studies (CETAP, 2018). The MAT test is divided into 5 subdomains, namely MAT M1, MAT M2, MAT M3, MAT M4 and MAT M5.MAT aims to assess candidates' ability with respect to several mathematical topics:

- Problem solving and modeling, requiring the use of algebraic processes, as well as understanding and using functions represented in different ways;
- Basic trigonometry, including graphs of trigonometric functions, problems requiring solution of trigonometric equations and application of trigonometric concepts;
- Spatial perception (angles, symmetries, measurements, etc.), including representation and interpretation of two- and three-dimensional objects; analytic geometry and circle geometry;

- Data handling and probability and competent use of logical skills.

It is not the intention of the MAT tests to replicate either the NSC or the Mathematics Olympiad. The NBT Mathematics test is explicitly designed to probe higher education competencies, that is the depth of understanding and knowledge within the context of the NSC curriculum (Prince, R., 2017).

1.6.4 The NBT Benchmarks

The NBT scores are divided into three categories, namely Proficient, Intermediate and Basic (CETAP, 2018). The description of these categories is shown in the table below:

Proficient	Performance in domain areas suggests that academic performance will not be adversely affected in cognate domains. If admitted, students should be placed on regular programmes of study (CETAP, 2018).
Intermediate	Challenges in domain areas identified such that it is predicted that academic progress in cognate domains will be affected. If admitted, students' educational needs should be met in a way deemed appropriate by the institution (e.g. extended or augmented programmes) (CETAP, 2018).
Basic	Serious learning challenges identified. Students will not cope with university study (CETAP, 2018).

Table 1.1: NBT overall benchmark descriptors

1.7 Research on the National Benchmark Tests

Many studies have been done on the importance of NBTs and similar tests for example the Health Sciences Placement Tests (HSPTs), developed for prospective medical students (A. A. Wadee and A. Cliff, 2016). Many researchers have done investigation on NBTs to answer certain questions as discussed in the following sections.

1.7.1 Can the NBTs be used to Assist Universities Understand Under-Preparedness of First-Year Students?

Wilson-Strydom, M. (2012) conclude that under-preparedness cannot be measured and it is best to view it as some students are prepared in some areas and under-prepared in other areas. This study encourages universities to review their curriculum for undergraduate studies and introduce various academic support programs that will improve learning and teaching. The authors conclude that the NBTs should not be used to measure under-preparedness and suggest a more comprehensive approach to assessing under-preparedness for higher education study.

1.7.2 Can the NBTs be used to Improve Placement and Admissions Procedures?

In February 2010 all the first-years in the University of the Free-State were required to write the NBTs (Wilson-Strydom, M., 2012). A multiple regression analysis was used to assess the extent to which the Admissions Point (AP) score and the NBT performance predict students' performance in selected first year modules. A total of 15 modules representing the Humanities, Economics, Management Sciences and Natural and Agricultural Sciences Faculties were considered. The dependent variable was the students' end of semester mark for each module. The independent variables were the different NBTs scores,namely the QL, MAT and AL and the Admissions Point (AP) score which is calculated using the NSC Grade 12 results only. The results from the multiple regression analysis show that combining the AP and the NBT scores provides a better means of admitting and placing students. The researchers suggest that both the NSC results and the NBT results provide sufficient information that aids admission committees in placement of first year students. This means that using NSC results and NBT results improves placement and admissions procedures.

1.7.3 Is there a Relationship Between the School-leaving Examinations and the University Entrance Assessments?

The NSC results are a requirement for entry into higher education and the NBTs were designed to help the universities make decisions about the most appropriate curriculum structures for students. These two assessments seem to be compatible and complementary. These two assessments are sufficiently different for them to be viewed as complementary sets of assessments (Prince, R., 2017). These results were obtained through correlation analysis, linear regression and Bland-Altman analyses. Prince, R. (2017) concluded that there is a sufficient difference between the school-leaving examinations and the university entrance assessments for them to be viewed as complementary sets of assessments.

1.7.4 Are the NBTs Better Predictors Of Performance than the NSC Results?

Due to the differences between the NSC results and the NBT results, studies have been conducted to investigate which one of the two is a better university academic performance predictor. Mahlobo, R. (2015) concluded that most students do well in the school-leaving examinations and fail the NBT examinations. The authors compared the two tests by comparing the NBT mathematics tests and the school-leaving mathematics exams. They found that both the NBT mathematics and NSC mathematics were not good predictors of first-years mathematics performance. As a result they suggested that there is a need to introduce other ways of assessing mathematical skills of prospective tertiary education students.

1.7.5 Are there Other Factors That Affect the NBT Results?

There are other factors that affect students' performance in the pre-admission tests for example whether the student did English as a home language or as a second language (A. A. Wadee and A. Cliff, 2016). A. A. Wadee and A. Cliff (2016) investigated if pre-admission tests of learning potential were predictors of academic success of first-year medical students. The results for this study provide evidence that the Health Sciences Placement Test results provide important information to selection committees about academic readiness for students from advantaged backgrounds or students from private schools. For disadvantaged students or students from government schools, the Health Sciences Placement Test scores provide alternative academic readiness information to the school-leaving examination results in particular how these students will cope with their studies. For example from an English medium-of-instruction teaching environment and about the extent and kind of academic support and curriculum responsiveness that might be indicated if such students are successful.

1.8 The Context of the Current Study

Pre-admission tests provide important information that helps admission committees in providing academic support programs and proving a curriculum that enhances learning and teaching in universities. As suggested from the literature reviewed these pre-admission tests do not take into consideration the language issue and the type of school the students attended which are important factors that affect performance. This paper aims to investigate whether NBTs results can be used to predict the performance of first-year students registered for Statistics 101 (STA 101), a course presented typically to Science Faculty students at Rhodes University. This study aims to investigate whether the NBTs can be used to predict a student's Statistics 101 test one result.

Chapter 2

Statistical Modeling and Learning

Data mining is a process of analysing data and extracting information that helps in decision making. Osmanbegovic, E and Suljic, M (2012) argue that prediction of student performance is important to higher education institutions as it helps in improving the quality of teaching. Higher education quality depends on offering services that meet the needs of students. These services and needs are assessed by collecting data which is integrated and utilised, by converting the data into knowledge such that it is helpful to the whole educational community. Educational Data Mining (EDM) is a data mining technique that is used for analysing data that is related to education. EDM is a process of transferring raw data collected from educational institutions into useful information that can be used to answer research questions. EDM uses many techniques, for example regression and correlation analysis, decision trees, neural networks etc. These techniques provide knowledge about the different associations, classifications and clustering of the variables or cases in the data set or study. This knowledge can be used to predict performance or other factors, such as identifying students that require additional assistance (Kumar, S. A. and Vijayalakshmi, M. N., 2011).

Statistical learning or data mining is a process of automatically searching big data to discover patterns and trends beyond simple analysis using advanced mathematical algorithms (Surampudi, S., 2015). These mathematical algorithms are used to split or cluster or classify the data to predict future events. Statistical learning is a guided process, meaning that the results depend on the formulation of the problem. Statistical learning is separated into two categories: supervised or unsupervised learning. Supervised learning is a process that is directed by prior knowledge of the dependent or target variable (James, G. and Witten, D. and Hastie, T. and Tibshirani, R., 2013). The goal of supervised learning is to explain the behavior of the target as a function of the set of independent variables, for example regression or supervised classification. In unsupervised learning there is no prior knowledge of the target, for example clustering using K-means algorithm (James, G. and Witten, D. and Hastie, T. and Tibshirani, R., 2013). Both processes are used in learning the structure of the data.

2.1 Statistical Modeling and Classification

Classification is typically, a supervised statistical learning or data mining process for predicting a qualitative or categorical response for an observation (James, G. and Witten, D. and Hastie, T. and Tibshirani, R., 2013). The goal of classification is to accurately predict the target class for each case in the data. In this context the classification is typically a discrete, mutually exclusive categorical variable.

2.1.1 Decision Trees

A decision tree is a classification algorithm that uses conditional statements called rules (Surampudi, S., 2015). These rules can be used to predict future observation class labels. This algorithm uses the tree presentation where each internal node is a condition of some attribute being examined and every branch is the outcome of the study. A decision tree predicts the targets by asking a series of question (Kumar, S. A. and Vijayalakshmi, M. N., 2011). At each instance a question is asked depending on the answers of the previous questions. These questions lead to the unique target. Decision trees are a popular technique because of their simplicity, how well they work with small or large data, their speed and their accuracy (Surampudi, S., 2015). Decision trees can be applied to both regression and classification problems (James, G. and Witten, D. and Hastie, T. and Tibshirani, R., 2013).

2.1.2 Regression

Regression is a predictive modeling technique which investigates the relationship between the dependent, or target variable and a set of independent, or predictor or feature, variables. It is a supervised process in that the target is known and typically continuous. Regression is a technique that helps in prediction and understanding of the relationship between variables. In regression the model is built by estimation of the target as a function of the predictors. This relationship is summarized and used to predict the target for data with unknown targets (Surampudi, S., 2015; Faraway, Julian J, 2014). Linear regression is the analysis of the linear relationship between the dependent and the independent or feature variable(s) (Uyanık, G. K. and Güler, N., 2013). This relation can be written as a linear function $y = \beta_0 + \beta_1 x_1 + \ldots + \beta_n x_n + \varepsilon$ where y denotes the dependent variable, X_i denotes the independent variables,β_i denotes the i^{th} parameter and ε denotes the error or natural variation term.

Simple linear is regression is regression analysis of one dependent variable and one independent variable while multiple regression is regression analysis of one dependent variable and more than one independent variables (Faraway, Julian J, 2014).

For inference based on a regression model to be valid it should conform to the following assumptions:

1. There is a linear relationship between the dependent variable and the independent variables. This means that a straight line relationship should exist between the predictors and the response (Faraway, Julian J, 2014). If the relationship is far from linear then the model prediction accuracy is reduced. Residual plots are a useful graphical tool for identifying non-linearity (Faraway, Julian J, 2014). Idealy the residual plot should show no recognisable pattern. The presence of a pattern indicates a problem with the linear model (James, G. and Witten, D. and Hastie, T. and Tibshirani, R., 2013).

2. The residuals are independent. The standard errors computed for the estimated regression coefficients are based on this assumption being true. If the residuals are correlated, then the estimated standard errors will tend to underestimate the true standard errors (James, G. and Witten, D. and Hastie, T. and Tibshirani, R., 2013). The Durbin-Watson test is used to check this assumption (Faraway, Julian J, 2014).

3. The residuals of the model are normally distributed. Q-Q plots are usually used to assess this assumption (Faraway, Julian J, 2014). The Shapiro-Wilk test can also be used to check this assumption (Faraway, Julian J, 2014). If this assumption is false a generalised linear model should be fitted (Faraway, J. J., 2016).

4. The errors have constant variance (homoscedasticity). This is an important assumption as the standard errors, confidence intervals and hypothesis tests associated with the linear model rely upon it (James, G. and Witten, D. and Hastie, T. and Tibshirani, R., 2013). The residual plots can be used to check this assumption, by assessing the shape the points take (Faraway, Julian J, 2014).

5. No outliers and no extremes. Cook's distance can be used to assess this assumption (Faraway, Julian J, 2014).

The assumptions are the same for multiple regression with one addition, namely checking for multicollinearity (James, G. and Witten, D. and Hastie, T. and Tibshirani, R., 2013). Multicollinearity is when the independent variables are highly related (Yoo, W. and Mayberry, R. and Bae, S. and Singh, K. and He, Q. P. and Lillard Jr, J. W., 2014). Multicollinearity causes inessential information, namely where a independent variable explains about the dependent variable is overlapped by another independent variable. A simple way to detect multicollinearity is to compute a correlation matrix of the independent variables (Faraway, Julian J, 2014). James, G. and Witten, D. and Hastie, T. and Tibshirani, R. (2013) suggest that the best way of assessing multicollinearity is by computing the variance inflation factor (VIF). The VIF is the ratio of the variance of the estimate β_j of the full model divided by the variance of the estimate β_j if fit on its own. The minimum VIF value is 1, which indicates the complete absence of collinearity. A VIF value that exceeds 5 or 10 indicates the presence of collinearity. Multicollinearity can be solved by using one of the independent variables that are highly related or by dropping the problematic variables (James, G. and Witten, D. and Hastie, T. and Tibshirani, R., 2013).

There are three main criteria concerned with evaluating a multi linear regression model (Faraway, J. J., 2016):

- The goodness of fit line : The goodness of fit is assessed by determining how far the residuals deviate from the trend line. The residual is the true value minus the predicted value.
- The significance levels : In R the significance of a variable is easy to recognise as this is denoted by ***, **, *, etc. The presence of three stars *** indicates that an independent variable is highly likely to be significantly related to the dependent variable. Common practice is to use a significance level of 0.05 to denote a statistically significant variable.
- The coefficient of determination or the multiple R-squared statistic: The closer the R-squared value is to 1 the better the fit of the model to the data. In models with a large numbers of independent variables the Adjusted R-squared statistic is used to compensate for model dimensionality. The Multiple R-squared statistic is not affected by an increase in the number of independent variables.

A multiple regression model is typically used to assess certain important questions (James, G. and Witten, D. and Hastie, T. and Tibshirani, R., 2013):

1. Is at least one of the independent variables useful in predicting the dependent variable? Researchers may investigate this by looking at the F-statistics and the p-value obtained after computing the summary of the linear model.
2. Do all the indepenent variables explain the dependent variable Y? Variable selection may be used to answer this question.
3. How well does the model fit the data? Common measures of how well the model fits are the relative standard error (RSE) and R^2 statistic, the fraction of variation explained.
4. How accurate is the prediction of the dependent variable given a set of independent variables? The data can be into two sets, the training data and the testing data. The model is fitted using the training data then the testing data is used for prediction.

2.1.3 Generalised Linear Models

Generalised Linear Models (GLMs) provide flexibility in modeling as they allow the response variables to have an error distribution other than the normal distribution (Faraway, J. J., 2016). GLM is a generalisation of ordinary linear regression. A GLM is specified by two components, namely that the response should be a member of the exponential family of distributions and the link function which describes how the mean of the response is related to a linear combination of the predictors.

The response Y, is from the exponential family if it can be written in the form below

$$f(y|\theta,\phi) = exp\left[\frac{y\theta - b(\theta)}{a(\phi)} + c(y,\phi)\right].$$

θ is called the canonical parameter which represents the location or mean while ϕ is called the dispersion parameter which represents the scale. Various members of the family are defined by specifying a, b and c. The linear predictor is defined as $\eta_i = \beta_0 + \beta_1 x_{1i} + \ldots + \beta_p x_{pi}$. The link function that describes the mean, $E(Y_i) = \mu_i$, depends on the linear predictor $g(\mu_i) = \eta_i$. The variance function describes how the variance, $var(Y_i)$, is related to the mean, namely $var(Y_i) = \phi V(\mu)$, where ϕ is the dispersion parameter.

The normal linear model is a special case of the GLM. For the normal linear model where $\epsilon \sim N(0, \sigma^2)$ where the linear predictor, has link function $g(\mu_i) = \mu_i$ and variance function $V(\mu_i) = 1$. Examples of common GLMs are shown in the table:

Family	Link	Variance Function
Normal	$\eta = \mu$	1
Poisson	$\eta = log(\mu)$	μ
Binomial	$\eta = log(\frac{\mu}{1-\mu})$	$\mu(1-\mu)$
Gamma	$\eta = \mu^{-1}$	μ^2
Inverse	$\eta = \mu^{-2}$	μ^3

Table 2.1: Common GLM model specifications.

2.1.4 Neural Networks

Neural Networks (NNs) are a supervised learning method. Neural Networks are algorithms that mimic the brain that consist of interconnected nodes that exchange information in the same way as neurons, dendrites and axons (Osmanbegovic, E and Suljic, M, 2012). NNs learn by observing different examples, similar to how children learn skills from observing parents. The only difference is that children can learn and regconise objects after observing once while NNs require many observations to obtain a sufficient predicting capacity (Gerritsen, L., 2017). An advantage of NNs is their ability to obtain answers for complex and vague data. Disadvantages are that NNs are abstract, complex and require a lot of data (Gerritsen, L., 2017).

2.1.5 Clustering

Clustering is an unsupervised data mining process for finding natural groupings in the data (Surampudi, S., 2015). Cluster analysis is the process of finding groups of observation that are similar. The goal of clustering analysis is to find clusters that have high similarity within groups and low similarity between groups (Surampudi, S., 2015). Similarity is measured by calculating a distance metric for example a euclidean distance for quantitative variables (Baxter, 2019). Clustering is the same as classification in that it is used to segment data, the difference is knowledge about the target. Clustering is an unsupervised process, that is there is no prior knowledge of the target (James, G. and Witten, D. and Hastie, T. and Tibshirani, R., 2013). Clustering is useful for exploring data (Surampudi, S., 2015). It is also useful in determining outliers and grouping similar objects on which supervised models can be built (Baxter, 2019). Cluster analysis is divided into two categories: hierarchical methods which consist of single linkage, complete linkage, average linkage and Ward and non-hierarchical methods such as the K-means algorithm. K-means is a simple and efficient distance based clustering algorithm that separates data into clusters (James, G. and Witten, D. and Hastie, T. and Tibshirani, R., 2013). It is a distance based algorithm because it relies on the distance function to measure similarity between data points (**?**).

Chapter 3

Can the NBTs Predict Academic Performance in Statistics 101?

Based on the literature reviewed in chapter one, it is reasonable to propose that the academic performance of Statistics 101 students can be predicted based on the students' NBT scores, namely Academic Literacy, Quantitative Literacy and Mathematics scores. This chapter investigates this hypothesis.

3.1 Study data

126 Statistics 101 students' test one scores were provided by Rhodes University. The NBT results for 317 first year students registered in the Faculty of Science at Rhodes University were provided by the Centre of Educational Testing for Access and Placement. 80 out of the 317 of the students were registered for Statistics 101. Excel sort and filter functions were used to merge the files. Only 65 out of the 80 registered Statistics 101 students had both the Statistics 101 test one score and the NBT test scores.

The final data has 65 student's Statistics 101 test one scores and NBT test results. In this analysis, the Statistics 101 results are treated as the dependent variable and the NBT results are the independent variables. The final data set has 65 observations of the variables comprising of Statistics 101 test one score, the three NBT test scores (AL, QL and MAT) and the associated NBT subdomains.

3.2 Descriptive Statistics

Most statistical tests rest upon the assumption of normality (Mendes and Pala, 2003). Graphical methods have been used to assess the normality assumption by inspection of the distribution the data. This approach is not reliable and does not guarantee that the distribution is normal. This graphical approach gives the readers an opportunity to judge the distribu-

tion by themselves (Ghasemi, A. and Zahediasl, S., 2012). Graphical assessments are not sufficient to conclude that this data is a random sample from a normal population, since graphical assessment is dependent on what an individual sees (Mendes and Pala, 2003). Supplementing the graphical methods with other statistical analysis tests will result in a reliable conclusion (Mendes and Pala, 2003). The Shapiro-Wilk (SW) test is the most powerful test of normality (Mendes and Pala, 2003). SW test was developed by Shapiro and Wilk and it is a test that is preferred in most situations (Mendes and Pala, 2003). The SW compares the scores in the sample to a normally distributed set of scores with the same mean and standard deviation. The null hypothesis states that the variable is a random sample from a normally distributed population and it works best when the data is large (Mendes and Pala, 2003). The alternative hypothesis states that the variable is a random sample from a population that is not normally distributed (Mendes and Pala, 2003).

3.2.1 Statistics 101 Test 1 Results

The Statistics 101 test one results (%) are numerical data represented in R as doubles. The *stats.desc* R function from the *pastecs* library was used to summarise the data, as per the script file in Appendix A, and the results are shown in a table below.

	Minimum	Maximum	Mean	Median	Variance	Standard Deviation	n
Percentage (%)	18.0000	94.2000	54.8077	57.0000	281.2588	16.7708	65

Table 3.1: Summary of the Statistics 101 test 1 results.

The graphical approach was used to investigate the shape of the data. A function called describeData() was used to plot the histogram, box plot and Q-Qplot and to assess the normality assumption using Shapiro-Wilk's test, see figure 3.1

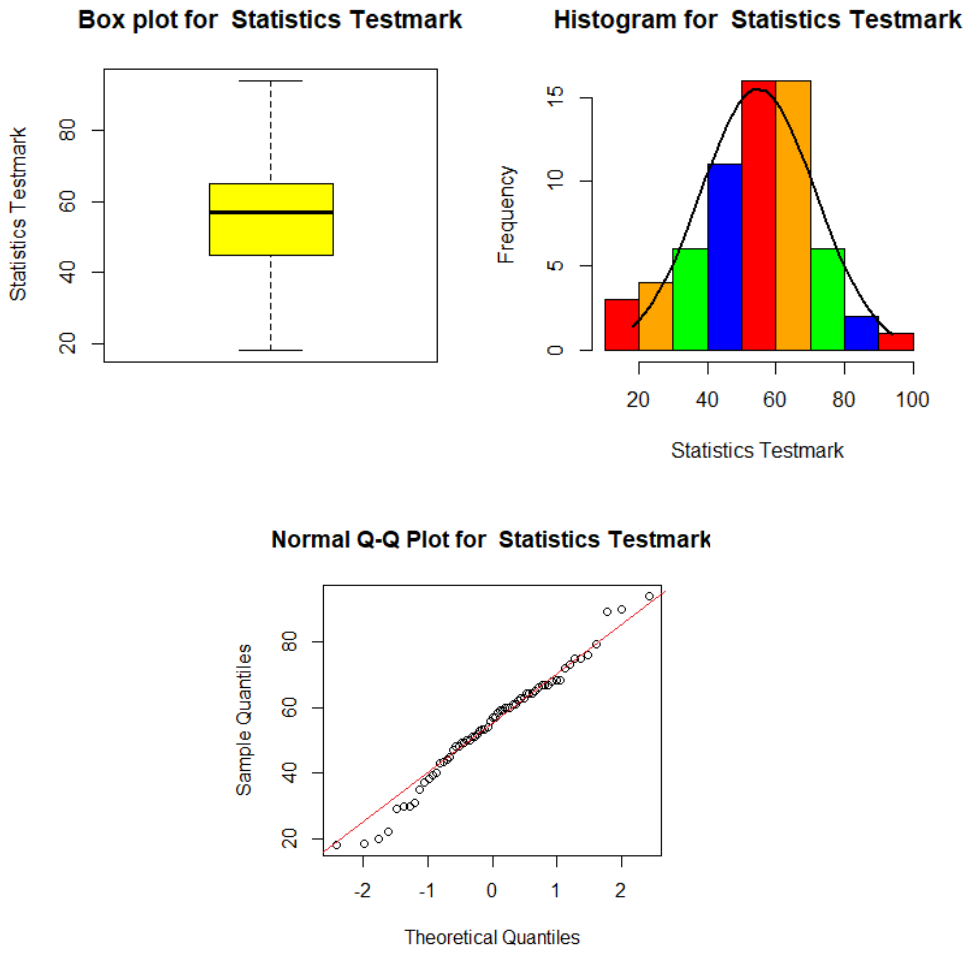

Figure 3.1: Box plot, histogram and Q-Q plot of the Statistics 101 test 1 results.

The histogram produced has an almost bell-shaped shape which means the data may be from a normal population. The problem with a histogram is that it is dependent on the choice of bins therefore it is unreliable (Ghasemi, A. and Zahediasl, S., 2012). The box-plot describes the distribution better than the histogram. The box-plot seems symmetric. The median line is approximately in the center of the box and the whiskers seem to be symmetric. According to Mendes and Pala (2003) Q-Q plots are the most valuable visual diagnostics tool. The function describeData() was used to plot a Q-Q plot for the Statistics test marks. From the Q-Q plot it can be seen that the data is close to the trend line, which is a good indicator that the population is roughly normal. The Shapiro-Wilk normality test indicated that these data provide insufficient evidence that the population from which this variable was selected is not normally distributed ($W = 0.98113$, p-value$= 0.423$). This means that these data may be from a normal population.

A bar graph was plotted to show the number of students that passed and the number of students that failed, see figure 3.2

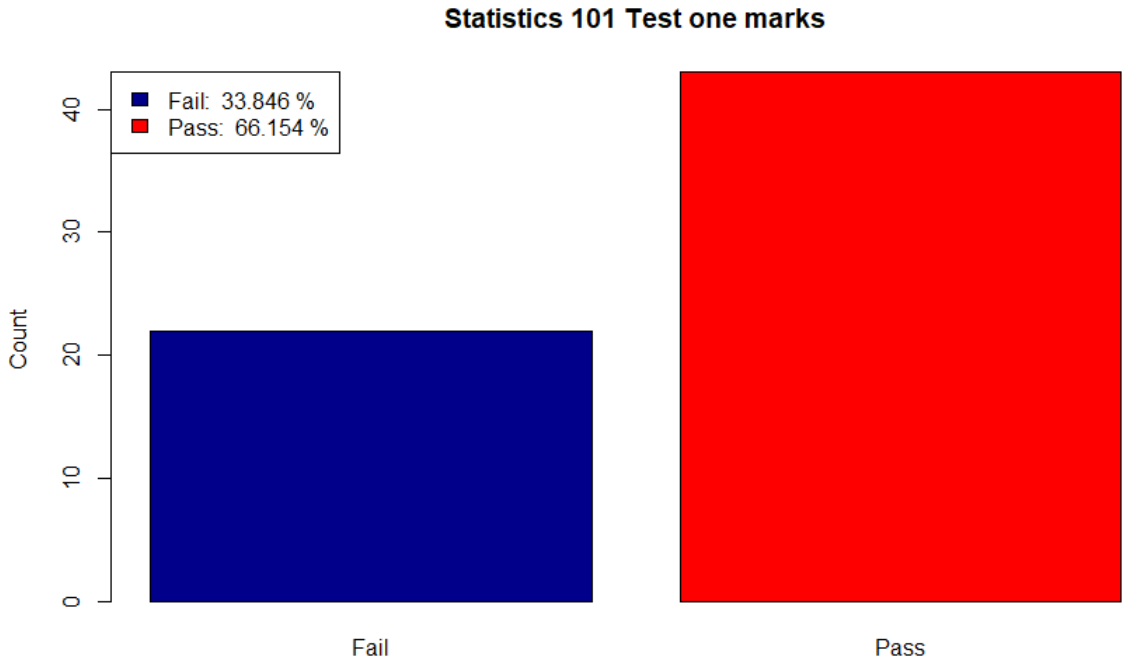

Figure 3.2: A bar plot summarising how many students failed and how many students passed Statistics 101.

3.2.2 NBT Test Results

The NBT test results (%) are numerical data represented in R as doubles. The summary statistics for these test is shown in table 3.2. The graphical summaries are shown in figure 3.3, 3.4 and 3.5.

	Minimum	Maximum	Mean	Median	Variance	Standard Deviation	n
AL Score	35.0000	81.0000	56.6923	57.000	195.9351	13.9977	65
QL Score	29.0000	96.0000	54.6308	52.0000	292.6740	17.1077	65
MAT Score	26.0000	95.0000	43.6308	40.0000	198.6115	14.0930	65

Table 3.2: Summary of the NBT test results.

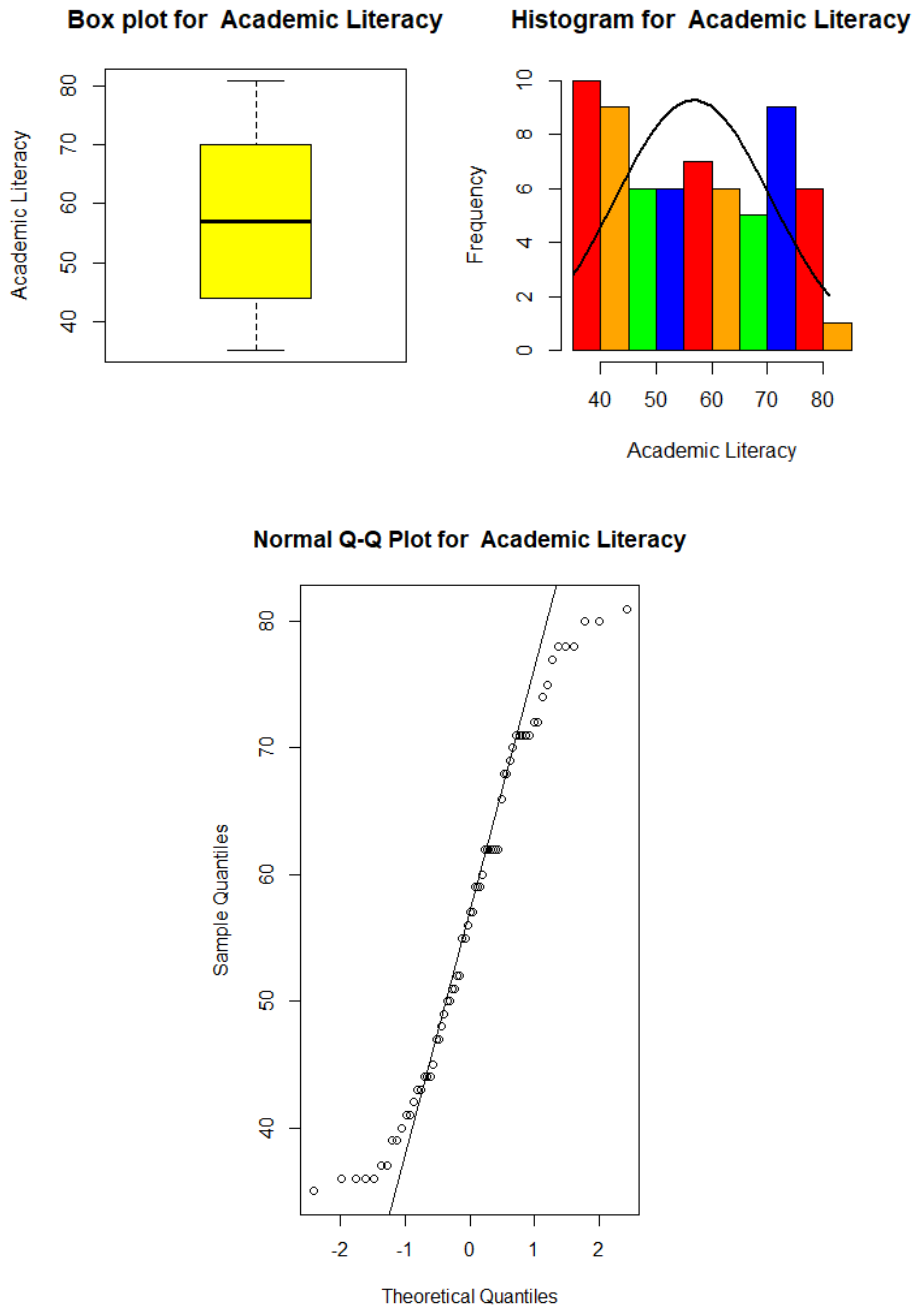

Figure 3.3: Box plot, histogram and Q-Q plot of Academic Literacy (AL).

The box plot of the AL test (figure 3.3) indicates that these data are symmetric.The Q-Q plot indicates that some of the data is close to the trend line and some of the data is far from the trend line, indicating a tailed distribution. The Shapiro-Wilk normality test indicates that these data provide sufficient evidence that the population from which this variable was selected is not normally distributed ($W = 0.94336$, p-value= 0.00502). This means that these data may be from a non-normal population.

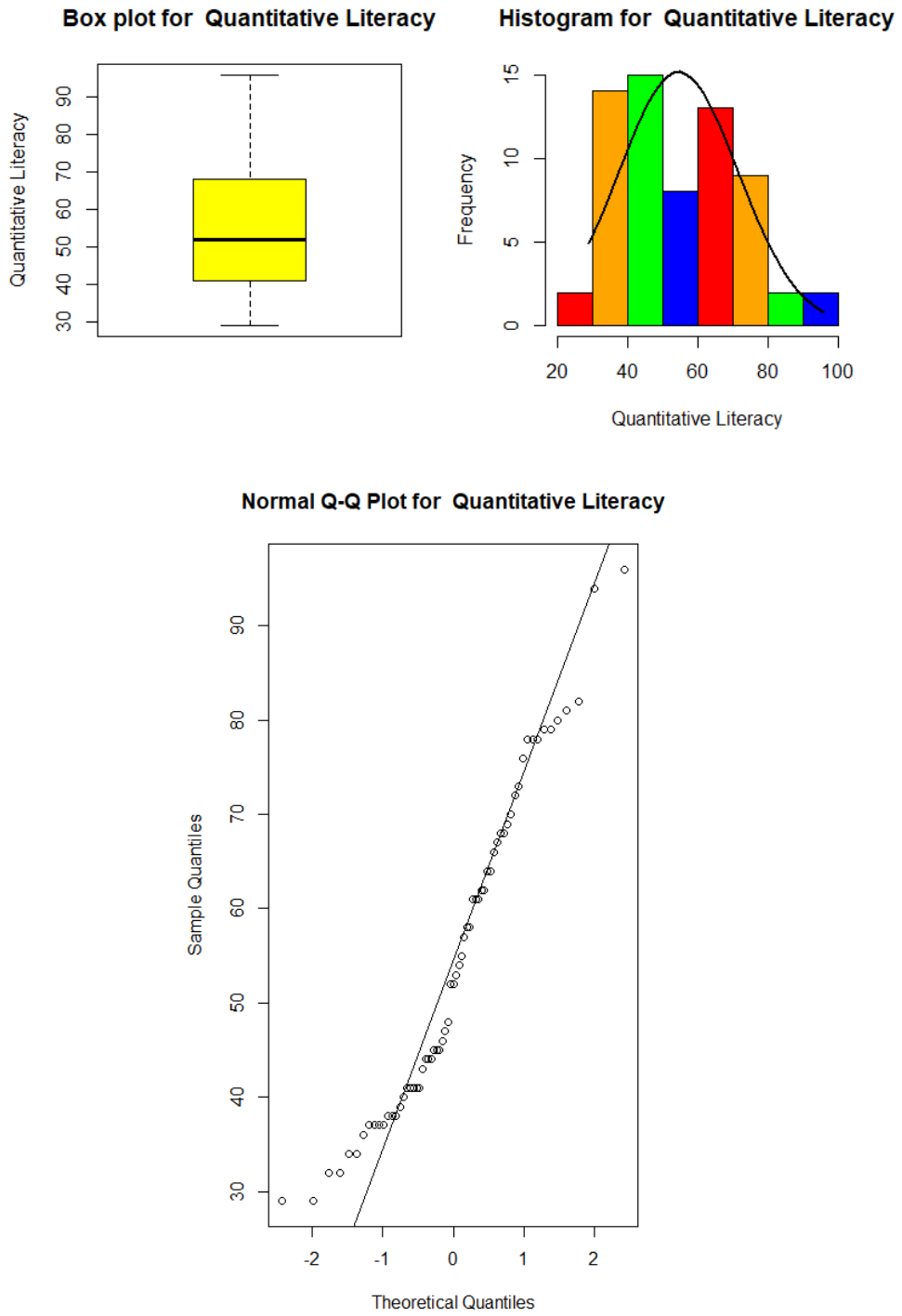

Figure 3.4: Box plot, histogram and Q-Q plot of Quantitative Literacy (QL).

The box plot and histogram of the QL test (figure 3.4) indicate that these data are right skew. Most points on the Q-Q plot are far from the trend line, suggesting that these data are not from a normally distributed population. The Shapiro-Wilk normality test indicated that these data provide sufficient evidence that the population from which this variable was selected is not normally distributed ($W = 0.94264$, p-value$= 0.004637$).

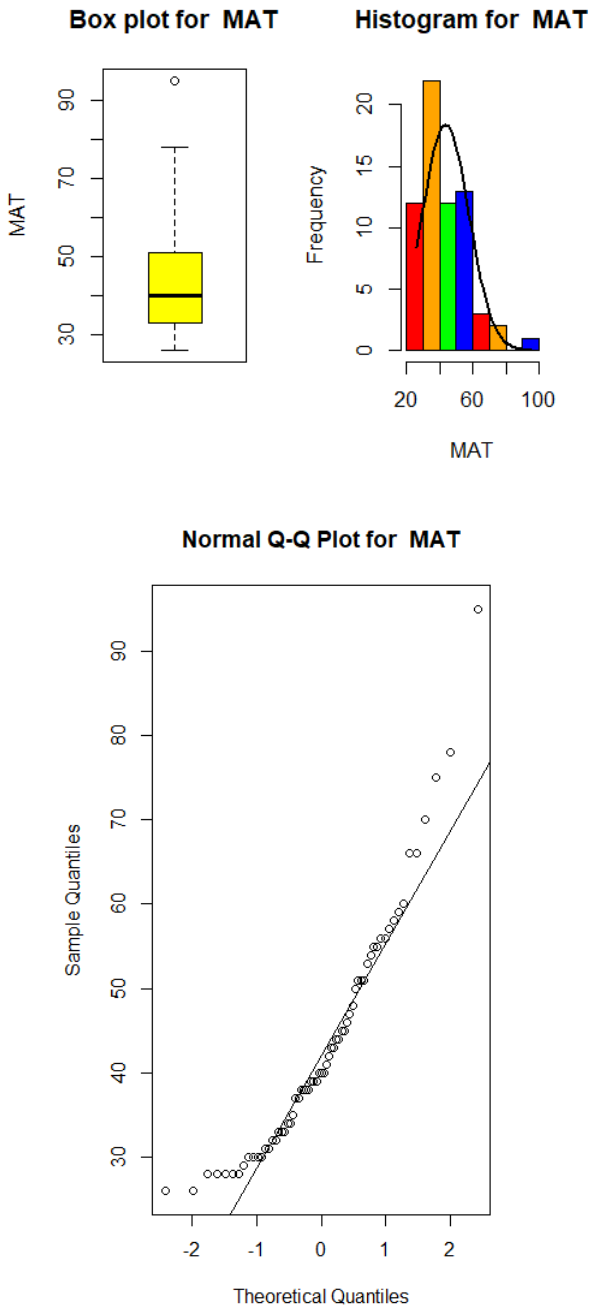

Figure 3.5: Box plot, Histogram and Q-Q plot of Mathematics (MAT).

The box plot and histogram of the MAT test (figure 3.5) indicate that these data are right skew. Most points on the Q-Q plot are far from the trend line, suggesting that these data are not from a normally distributed population. The Shapiro-Wilk normality test indicated that these data provide sufficient evidence that the population from which this variable was selected is not normally distributed ($W = 0.90555$, p-value$= 0.0001159$).

3.2.3 NBT Subdomain Results

The NBT subdomain results (%) are numerical data represented in R as doubles. Box plots were used to describe each NBT subdomain and are shown in figure 3.6.

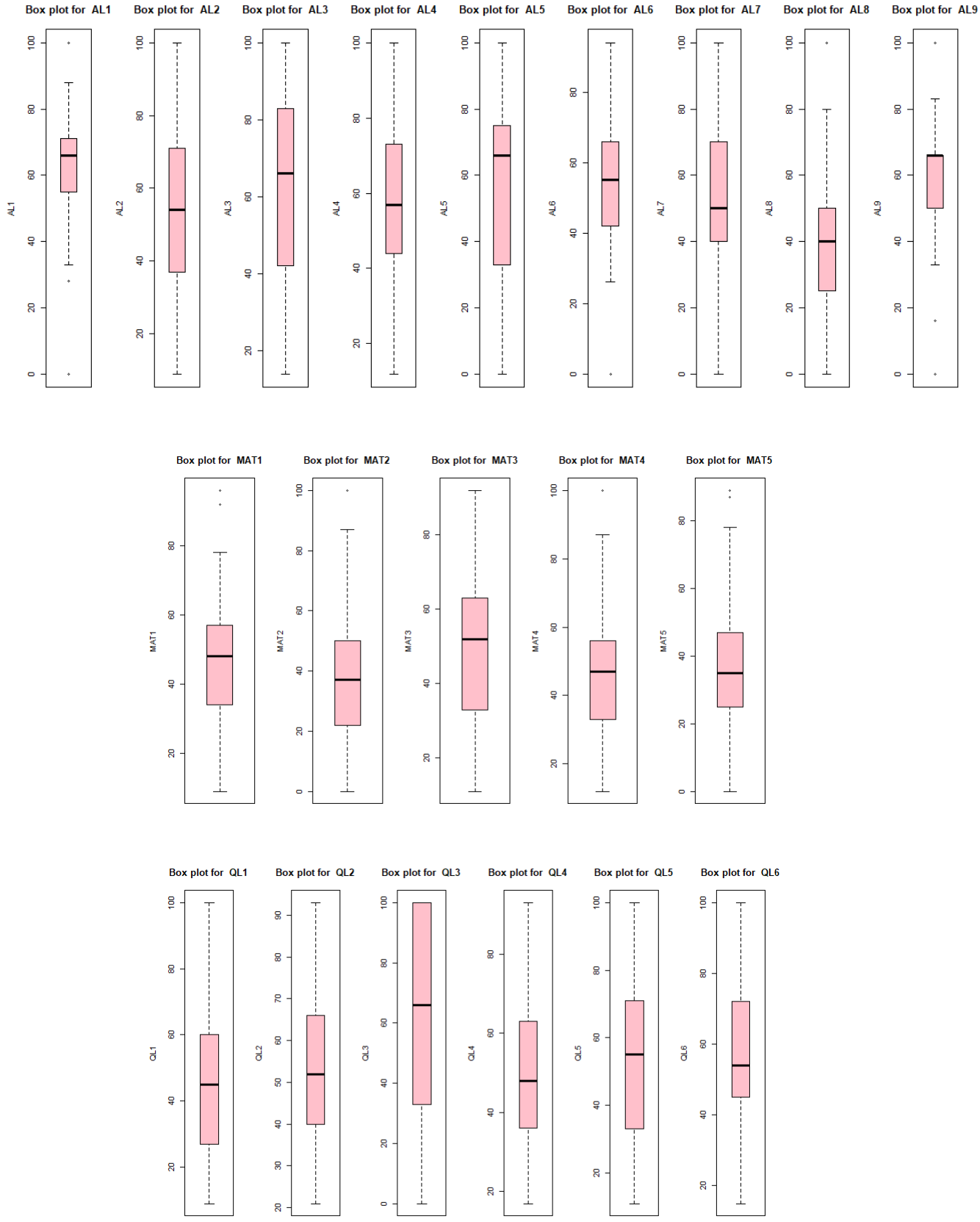

Figure 3.6: Box plots of the subdomains of the NBT tests.

3.3 Categorical Analysis:

3.3.1 Is Statistics 101 Test 1 (as Pass/Fail) Dependent on AL (in 3 categories)?

Contingency tables are a special type of frequency distribution tables which are used to summarize the relationship between several categorical variables (Faraway, Julian J, 2014). A contingency table (table 3.3) was constructed for the recorded variable PassFail_Factor, which separates the STA 101 marks into pass and fail categories and the recorded variable ALLevel which separates the AL marks into the 3 categories, namely Basic, Intermediate and Proficient. This is shown in table 3.3.

	ALLevel		
PassFail_Factor	Basic	Intermediate	Proficient
Fail	2	15	5
Pass	5	23	15

Table 3.3: Frequencies for PassFail_Factor and ALLevel factor.

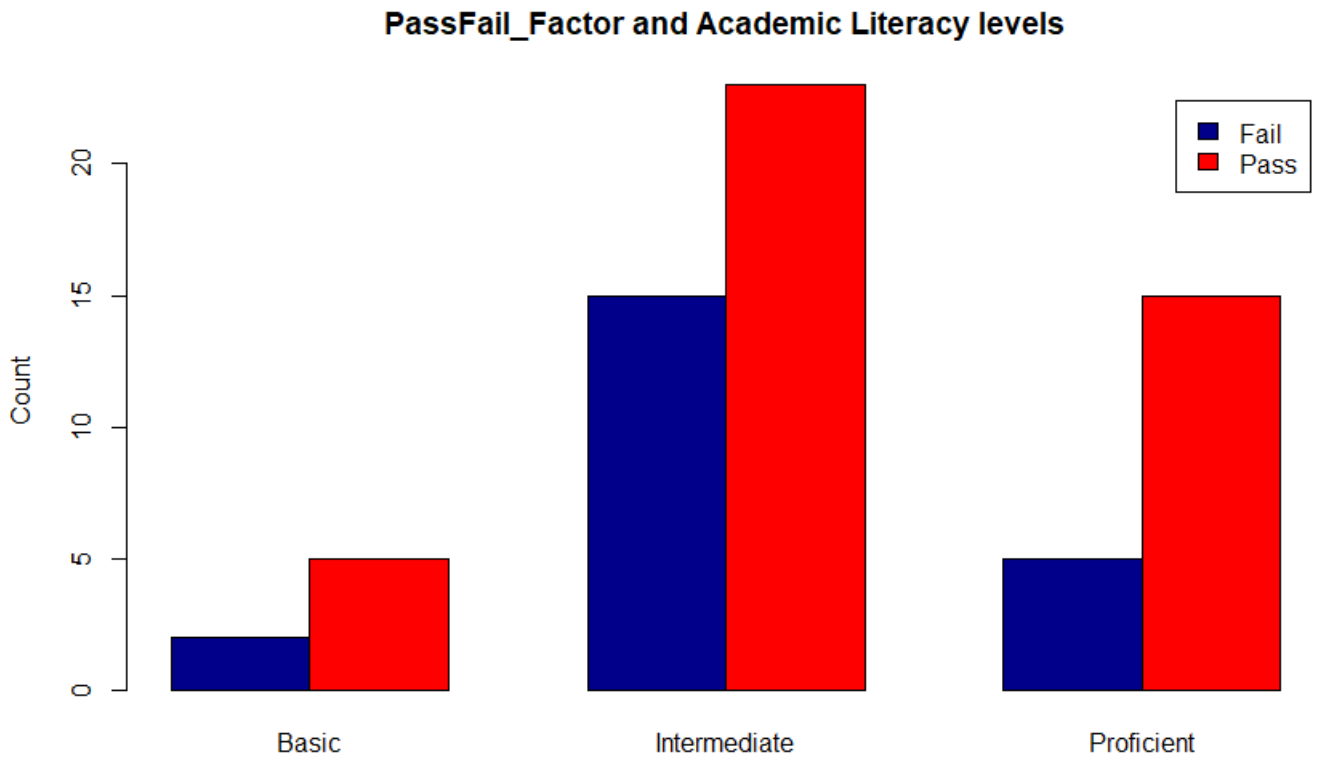

Figure 3.7: Frequencies for PassFail_Factor and ALLevel factor.

The Chi-squared test of independence (also known as Pearson's Chi-Squared) is a non-parametric statistic designed to test whether or not a relationship exists between categorical

variables (Bewick et al., 2003). The null hypothesis of the Chi-Squared test is that no relationship exists between the categorical variables in the population, that is they are independent. The alternative hypothesis states that the categorical variables are dependent. The Chi-Squared test tests for independence by measuring how well the observed distribution of data fits with the distribution that is expected if the variables are independent. Calculating the Chi-Square statistic and comparing it against a critical value from the Chi-Square distribution allows the researcher to assess whether the observed cell counts are significantly different from the expected cell counts (Bewick et al., 2003). As with any statistic, there are requirements for its appropriate use, which are called "assumptions" of the statistic.The assumptions are (McHugh, M. L. , 2013):

1. The data in the cells should be frequencies.

2. The categories of the variables are mutually exclusive.

3. Each subject may contribute data to one and only one cell in the χ^2.

4. The study groups must be independent.

5. The value of the cell expected values should be 5 or more in at least 80% of the cells, and no cell should have an expected of less than one.

The Chi-squared test was conducted for the PassFail_Factor variable and the ALLevel variable. The assumptions above were acceptable. The test indicates that these data provide insufficient evidence that the categorical STA 101 variable is dependent on the categorical AL variable ($\chi^2_{obs} = 1.3234$, df=2, p-value= 0.516).

3.3.2 Is Statistics 101 Test 1 (as Pass/Fail) dependent on QL (in 3 categories)?

A contingency table (table 3.4) was constructed for the recorded variable PassFail_Factor, which separates the STA 101 marks into pass and fail categories and the recorded variable QLLevel which separates the QL marks into the 3 categories, namely Basic, Intermediate and Proficient. This is shown in table 3.4.

	QLLevelBasic		
PassFail_Factor	Basic	Intermediate	Proficient
Fail	6	10	6
Pass	9	26	8

Table 3.4: Quantitative Literacy in the 3 categories.

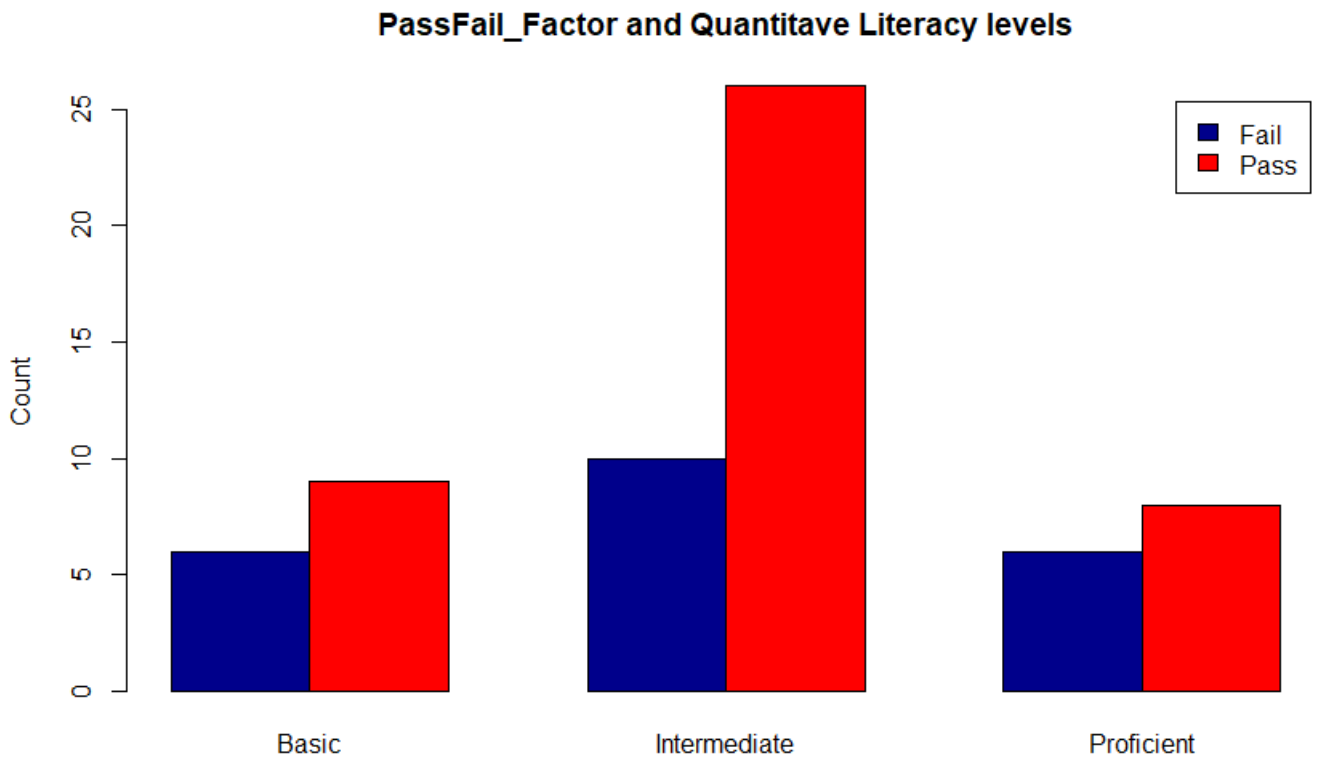

Figure 3.8: Frequencies for PassFail_Factor and QLLevelBasic.

The Chi-squared test assumptions were acceptable. The Pearson's Chi-Squared test indicates that these data provide insufficient evidence that the categorical STA 101 variable is dependent on the categorical QL variable ($\chi^2_{obs} = 1.3535$, df=2, p-value= 0.5083).

3.3.3 Is Statistics 101 Test 1 (as Pass/Fail) dependent on MAT (in 3 categories)?

A contingency table (table 3.5) was constructed for the recorded variable PassFail_Factor, which separates the STA 101 marks into pass and fail categories and the recorded variable MATLevel which separates the MAT marks into the 3 categories, namely Basic, Intermediate and Proficient. This is shown in table 3.5.

	MATLevelBasic		
PassFail_Factor	Basic	Intermediate	Proficient
Fail	8	13	1
Pass	13	27	3

Table 3.5: Mathematics in the 3 categories.

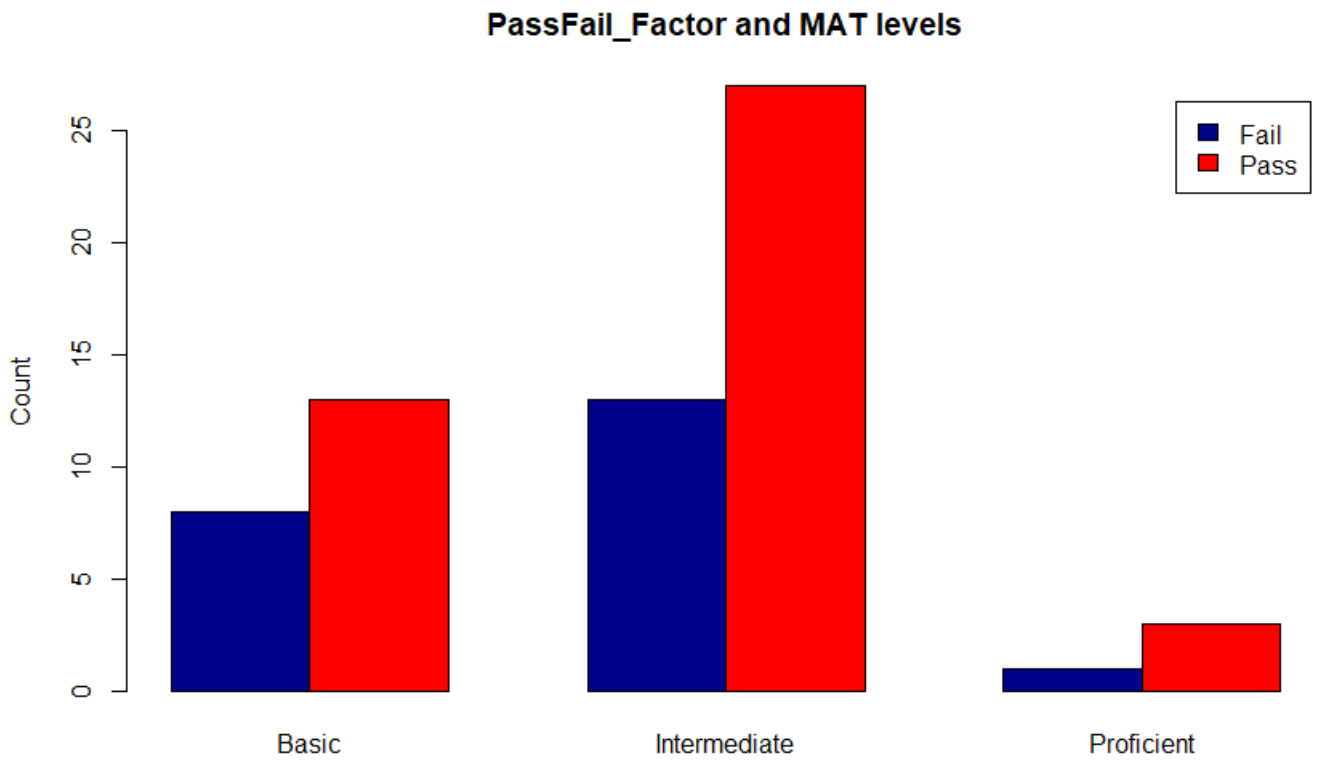

Figure 3.9: Frequencies for PassFail_Factor and MATLevelBasic.

The Chi-squared test assumptions were acceptable. The Pearson's Chi-Squared test indicates that these data provide insufficient evidence that the categorical STA 101 variable is dependent on the categorical MAT variable ($\chi^2_{obs} = 0.34151$, df=2, p-value= 0.843).

3.4 Linear Regression Modeling

A matrix scatter plot of the NBT results against the Statistics 101 test results is shown in figure 3.10 . Figures 3.13 shows the matrix scatter plots of the NBT subdomains.

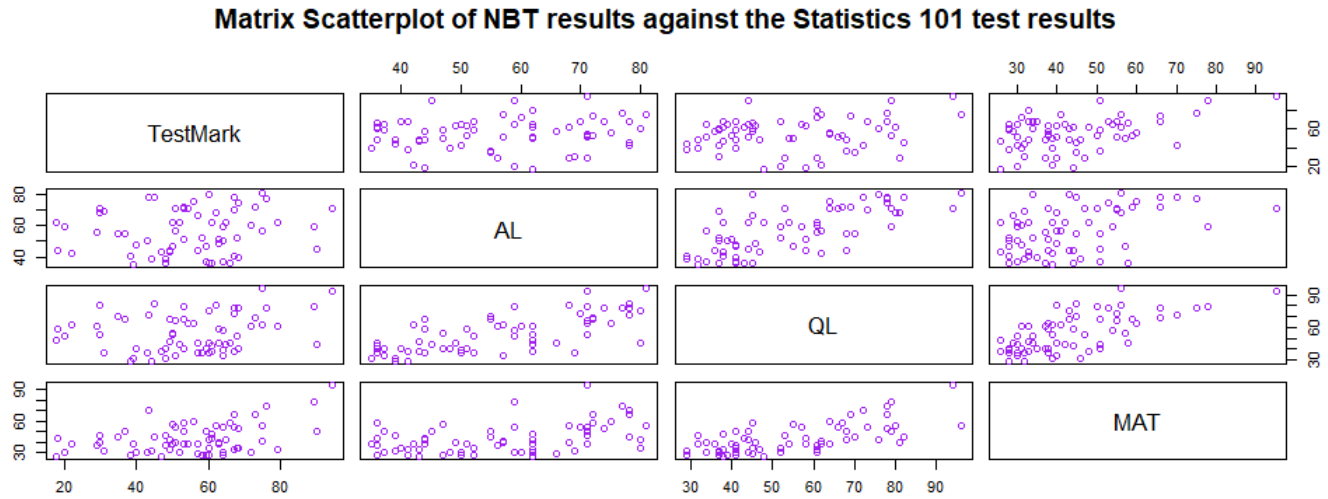

Figure 3.10: Scatter plot of NBT results vs Statistics 101 test results

As per literature reviewed, the scatter plot agrees with the conclusions Mahlobo, R. (2015) obtained. The scatter plot indicates that there is no linear relationship between the MAT scores and the STA 101 scores. Most of the literature reviewed states that NBTs can predict academic performance but these plots contradict the literature.

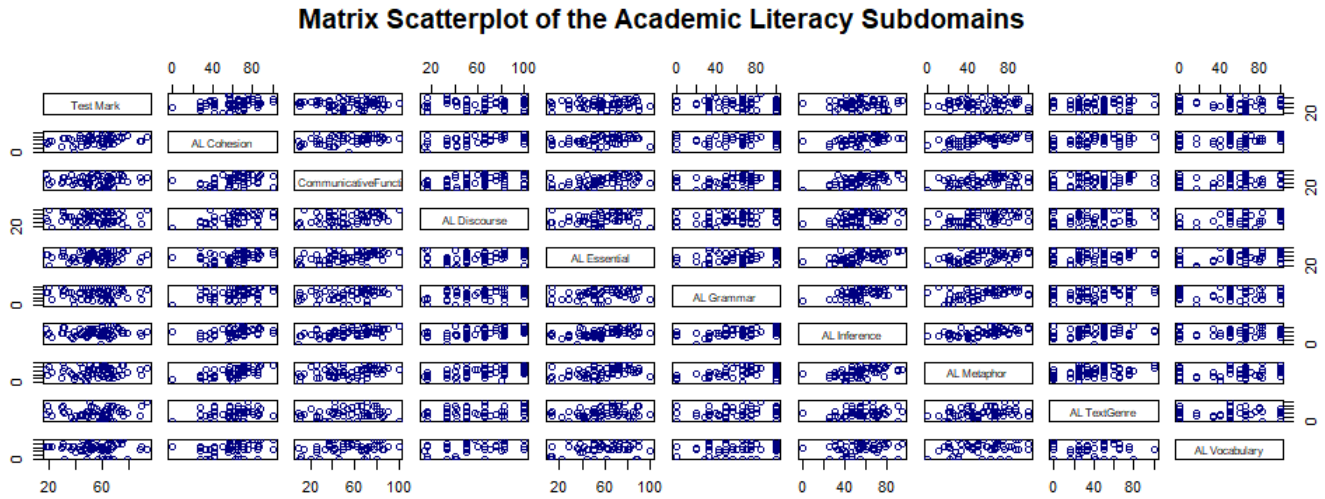

Figure 3.11: Scatter plot of the AL subdomains.

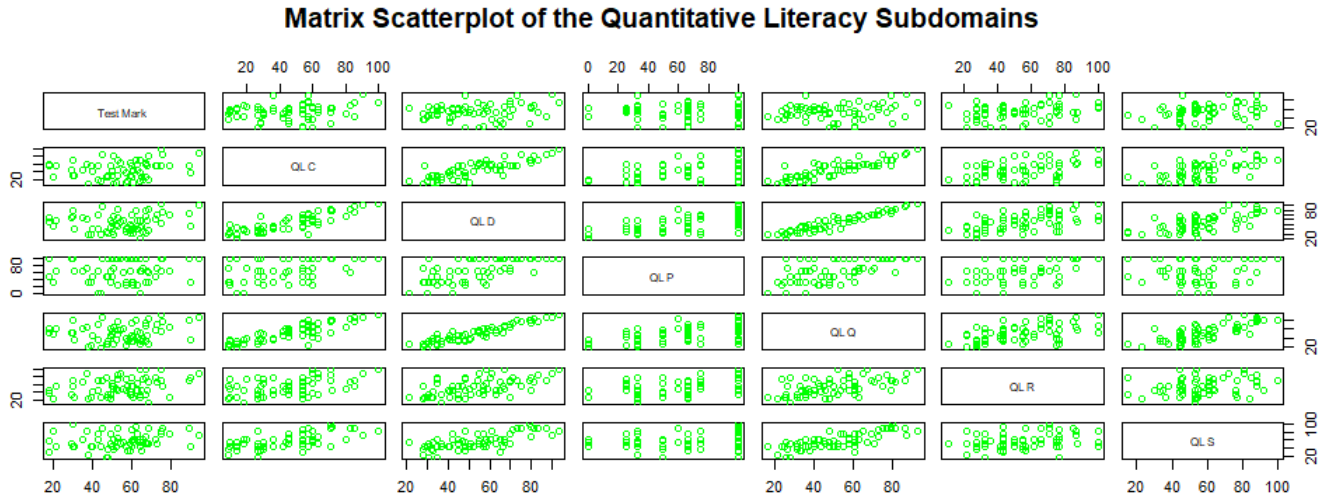

Figure 3.12: Scatter plot of the QL subdomains.

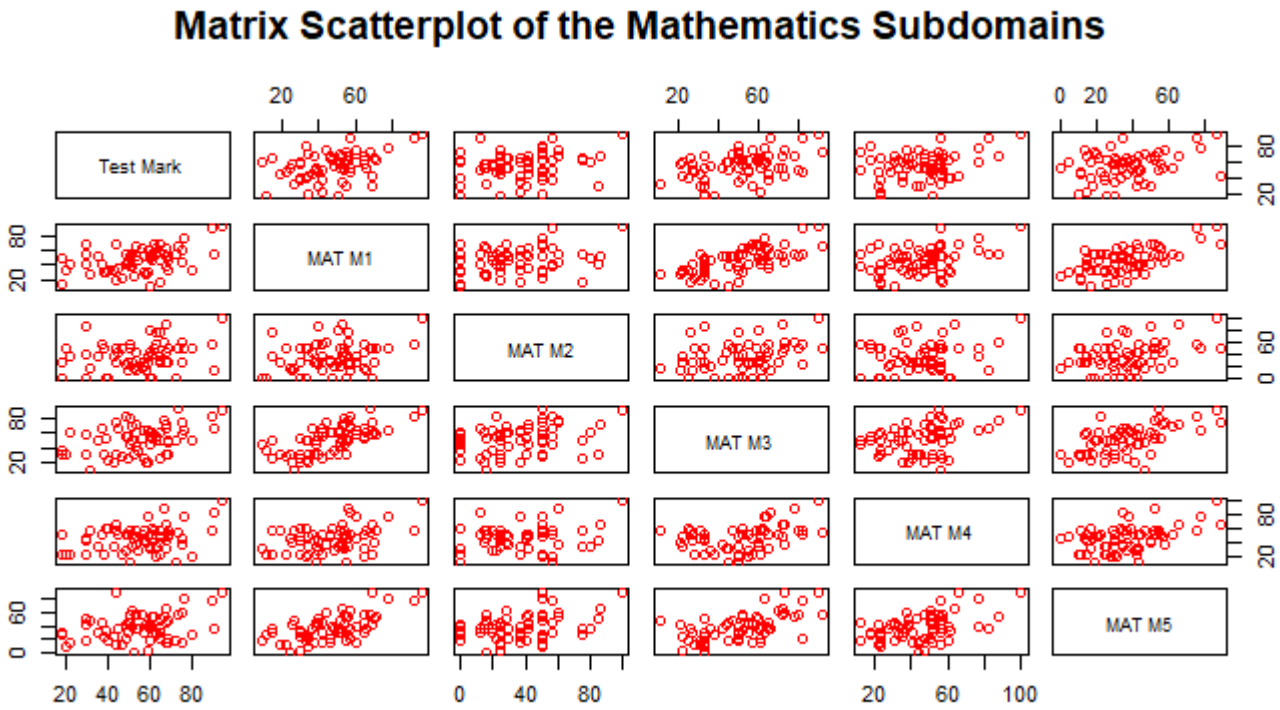

Figure 3.13: Scatter plot for the MAT subdomains.

3.4.1 STA 101 as a Linear Combination of AL, QL and MAT

The Statistics 101 results were modeled as a function of the AL, QL and MAT, denoted as model 1.

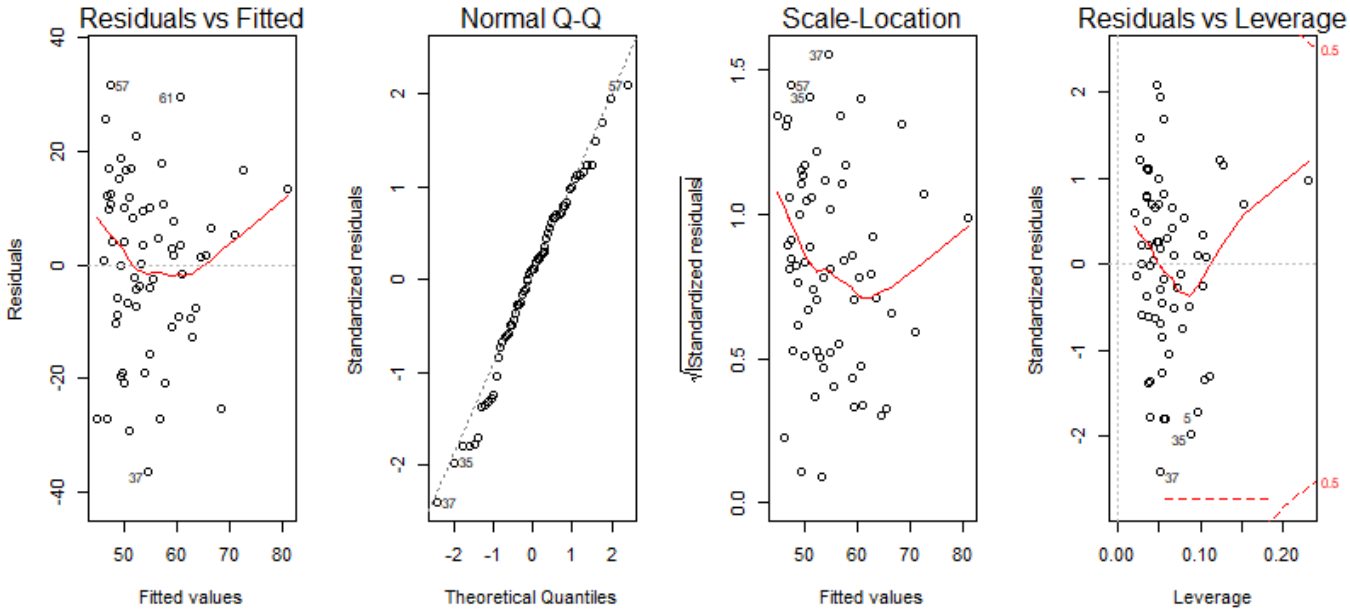

Figure 3.14: Diagnostics plots for model 1.

According to Faraway, Julian J (2014) estimation and inference from a regression model depends on the assumptions mentioned in chapter 2. Residuals plots are a useful graphical tool for identifying non-linearity and non-constant variance. The residual plot in figure 3.14 is slightly u-shaped which provides some indication of non-linearity (James, G. and Witten, D. and Hastie, T. and Tibshirani, R., 2013). The residual plot shows that there is constant variance. The Q-Q plot suggests that the residuals are normally distributed. The Shapiro-Wilk normality test indicates that these data provide insufficient evidence that the residuals are not normally distributed ($W = 0.98289$, p-value= 0.5072), meaning that the normality of the residuals assumption is acceptable. To investigate the independence of the error assumption the Durbin-Watson test for independence was used (Faraway, Julian J, 2014). The Durbin-Watson Test is a measure of autocorrelation in residuals from a regression analysis. The null hypothesis states that there is no autocorrelation. The alternative hypothesis for this test is that there is autocorrelation. The Durbin-Watson independence test indicates that these data provide insufficient evidence that the residuals are auto correlated or independent distributed ($DW = 2.006$, $p - value = 0.4993$). This means that the correlation of the error terms is acceptable. The extreme values do not seem to be that influential (last two plots in figure 3.14). Therefore the usual observation assumption is acceptable. Checking for multicollinearity is important. The correlation plot, figure 3.15 shows that AL and QL are significantly correlated ($R = 0.70$). A better way to assess multicollinearity is to compute the

the variance inflation factor (VIF) (James, G. and Witten, D. and Hastie, T. and Tibshirani, R., 2013). A VIF value of close to 1, indicates the absence of collinearity. A VIF value of 5 or more indicates the presence of collinearity. The computed VIF for the NBT test scores (AL, QL and MAT), are $AL = 1.985018$, $QL = 2.930374$ and $MAT = 1.844301$. The VIF values are close to 1 and less than 5 which means that there is no evidence of multicollinearity.

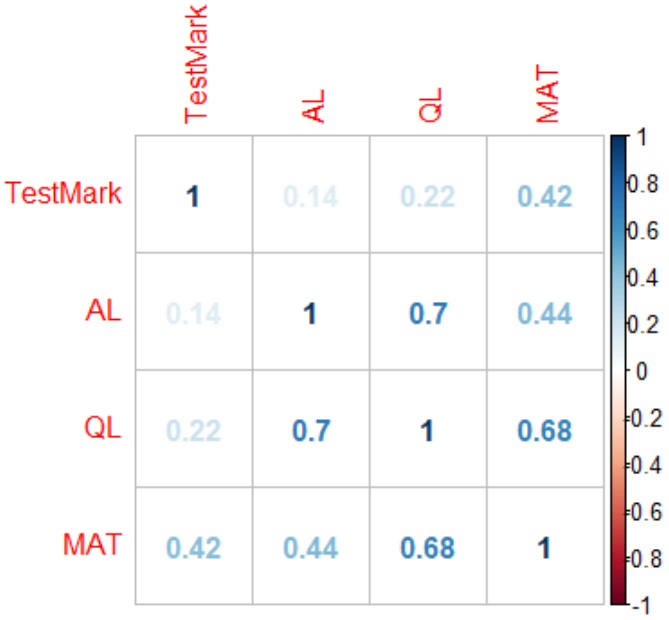

Figure 3.15: Correlation plot for the NBT test scores.

The adjusted R^2 value for this linear model is 0.1467 implying that approximately 85% of the variation in the Statistics 101 test marks are not explained by variation in the MAT and QL scores (Faraway, Julian J, 2014). This is very low and an indication that there are variables not included in this model that make a contribution to the Statistics 101 marks. This model is however significant, there is significant linear relationship between a Statistics 101 test score and the AL ,MAT and QL test scores ($F_{obs} = 4.667$, df= $3, 61$, p-value=0.0053). The intercept term makes a significant contribution to this model ($t_{obs} = 8.5871$, p-value= 0.00123) as does the MAT score ($t_{obs} = 3.220$, p-value= 0.002056). However the QL score does not make a significant contribution to this model ($t_{obs} = -0.019$, p-value $= 0.545972$) and also the AL score does not make a significant contribution to this model ($t_{obs} = -0.607$, p-value $=$ 0.984731).

3.4.2 STA 101 as a Linear Combination of the AL Subdomains

The Statistics 101 results were modeled as a linear combination of the AL subdomains, denoted as model 2. The output results of the model are shown in R code in chapter 6 4. The assumptions were assessed via the relevant diagnostic plots (figure 3.16)

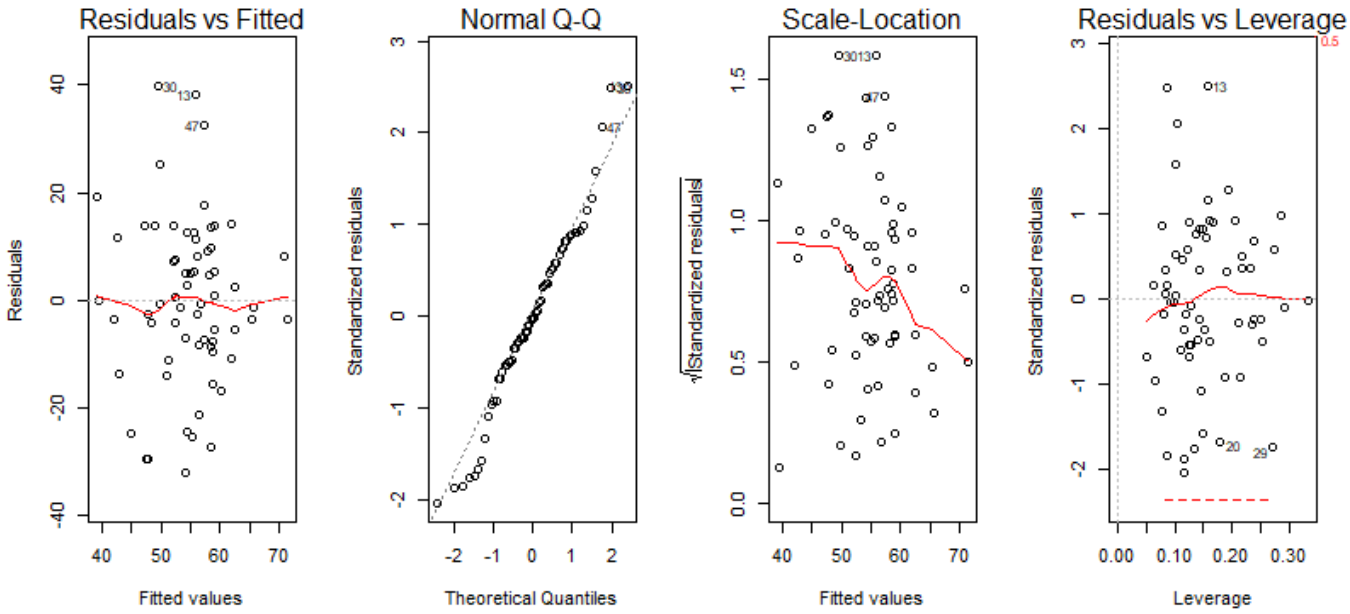

Figure 3.16: Diagnostic plots for 2.

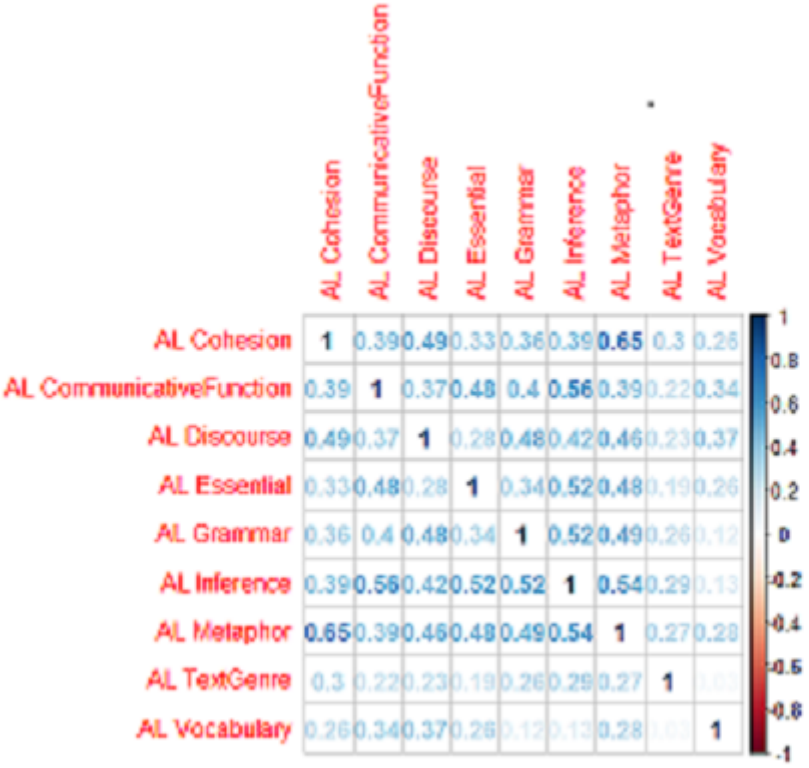

Figure 3.17: AL correlation matrix

The residual plot in figure 3.16 is slightly W-shaped. The residual plot indicates that there is constant variance. The Q-Q plot suggest the residuals are normally distributed. The Shapiro-Wilk test indicated that these data provide insufficient evidence that the residuals are not normally distributed ($W = 0.97708$, p-value= 0.269). The Durbin-Watson test indicates that these data provide insufficient evidence that the residuals are autocorrelated ($DW = 1.9266$, $p - value = 0.3403$). The extreme values do not seem to be that influential (last two plots in figure 3.16). The correlation plot figure 3.17 suggest that none of the AL subdomains are correlated. The VIF for the NBT AL subdomain test scores, the values are $AL1 = 1.986268$, $AL2 = 1.793228$, $AL3 = 1.739336$, $AL4 = 1.592966$, $AL5 = 1.684007$, $AL6 = 2.139636$, $AL7 = 2.372173$, $AL8 = 1.161998$ and $AL9 = 1.342320$. The VIF values are close to 1 and less than 5 which suggests that multicollinearity is not an issue in this model.

The output of the regression analysis (see Appendix 4) indicates that the adjusted R^2 value is 0.1148 implying that approximately 88% of the variation in the Statistics 101 test marks is not explained by variation in the NBT AL subdomains test score. This is very low and an indication that there are variables not included in this model that make a contribution to the Statistics 101 marks. There is no significant linear relationship between a Statistics 101 test score and the NBT AL subdomains test scores ($F_{obs} = 1.083$, df= $9, 55$, p-value=0.3904). The intercept term makes a significant contribution to the model ($t_{obs} = 4.500$, p-value $<$ 0.0001) as does the AL Cohesion test score ($t_{obs} = 2.753$, p-value= 0.008), however the other AL subdomains do not make a significant contribution to this model.

3.4.3 STA 101 as a Linear Combination of the QL Subdomains

The Statistics 101 results were modeled as a linear combination of the QL subdomains, denoted as model 3. The output results of the model are shown in R code in chapter 6. When the assumptions were analysed, there was presence of multicollinearity. The VIF values for the QL subdomains were: QL D had a VIF value of 10.317917 and QL Q had VIF value of 13.027306. This was solved by dropping the problematic variables and fitting the model again (James, G. and Witten, D. and Hastie, T. and Tibshirani, R., 2013). The regression diagnostics is shown in figure 3.18.

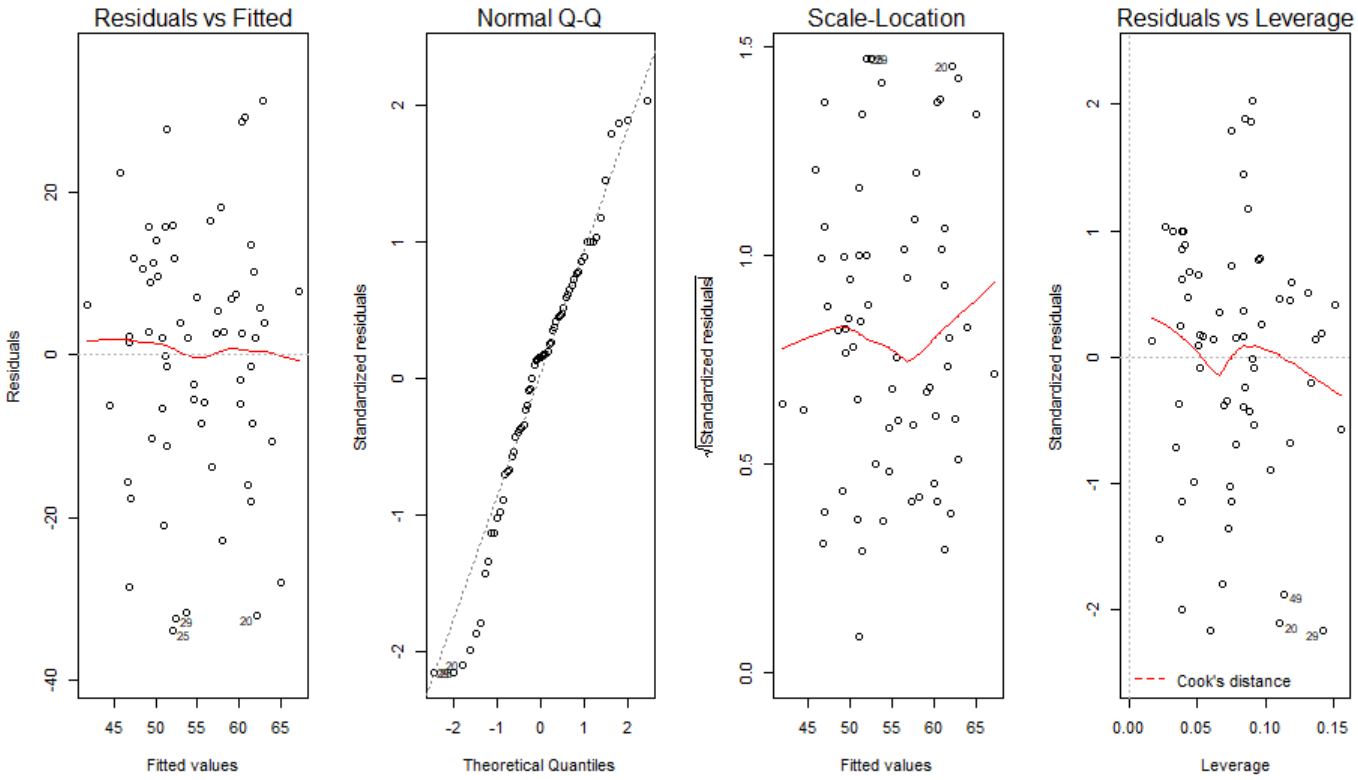

Figure 3.18: Diagnostic plots for model 3.

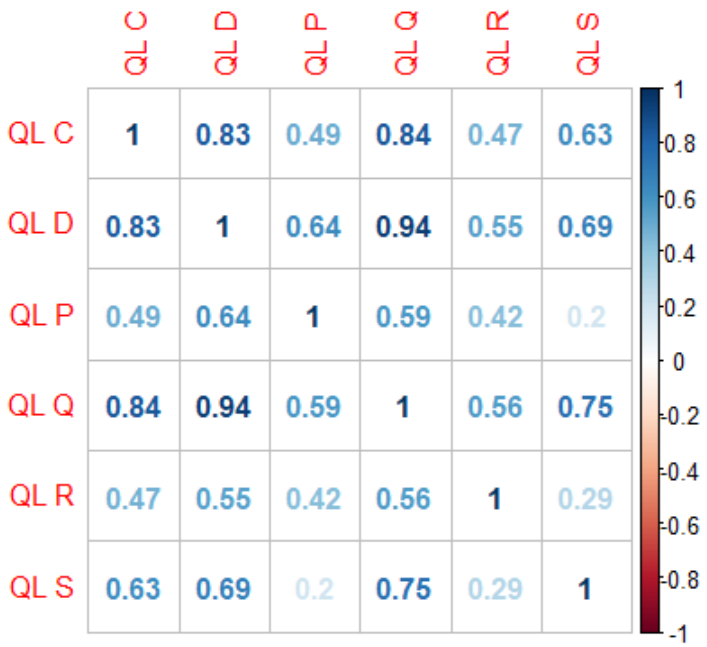

Figure 3.19: QL correlation matrix

The residual plot in figure 3.18 looks reasonable, the linearity assumption is valid. The residual plot shows that there is constant variance. The Q-Q plot suggest that the residuals

are normally distributed. The Shapiro-Wilk test indicated that these data provide insufficient evidence that the residuals are not normally distributed ($W = 0.97164$, p-value= 0.1409). The Durbin-Watson test indicates that these data provide insufficient evidence that the residuals are autocorrelated ($DW = 1.929$, p-value = 0.3719). The extreme values do not seem to be that influential (figure 3.18). The correlation plot (figure 3.19) show none of the QL subdomains are correlated. The VIF for the four NBT QL subdomain test scores are: $QL1 = 2.289135$, $QL3 = 1.445600$, $QL5 = 1.372804$ and $QL6 = 1.695369$. The VIF values are close to 1 and less than 5 which which suggest these variables are not multicollinear.

The output of this regression model (see Appendix 4) indicates that the adjusted R^2 value is 0.06693 implying that approximately 93% of the variation in the Statistics 101 test marks is not explained by variation in the NBT QL subdomains QL1, QL3, QL5 and QL6 test scores . This is very low and an indication that there are variables not included in this model that make a contribution to the Statistics 101 marks. There is no significant linear relationship between a Statistics 101 test score and the NBT QL subdomains test scores ($F_{obs} = 2.148$, df= $4, 60$, p-value=0.0859). The intercept term makes a significant contribution to the model ($t_{obs} = 4.960$, p-value < 0.0001) as does the QL5 test score ($t_{obs} = 2.156$, p-value= 0.0351), however the other QL subdomains do not make a significant contribution to this model.

3.4.4 STA 101 as Linear Combination of the MAT subdomains

The Statistics 101 results were modeled as a function of the MAT subdomains, denoted as model 4. The output results of the model are shown in R code in appendix. The diagnostics plots are shown in figure 3.18.

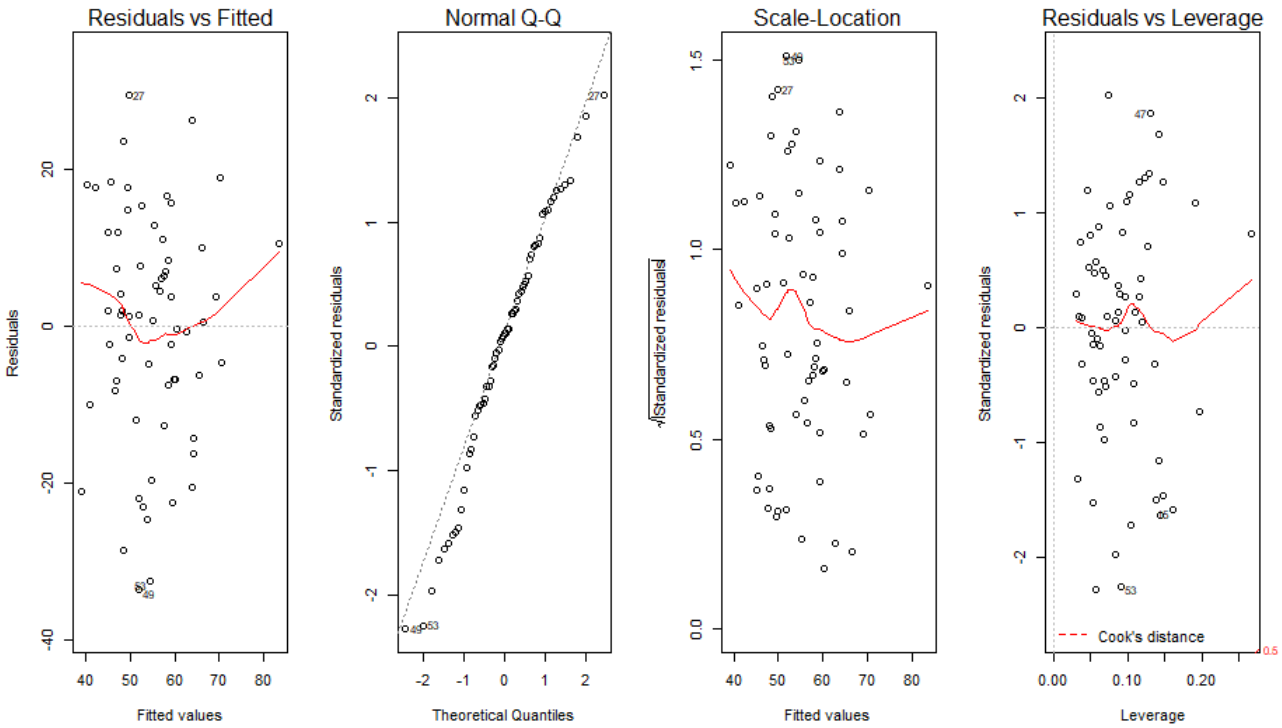

Figure 3.20: Diagnostic plots for model 4.

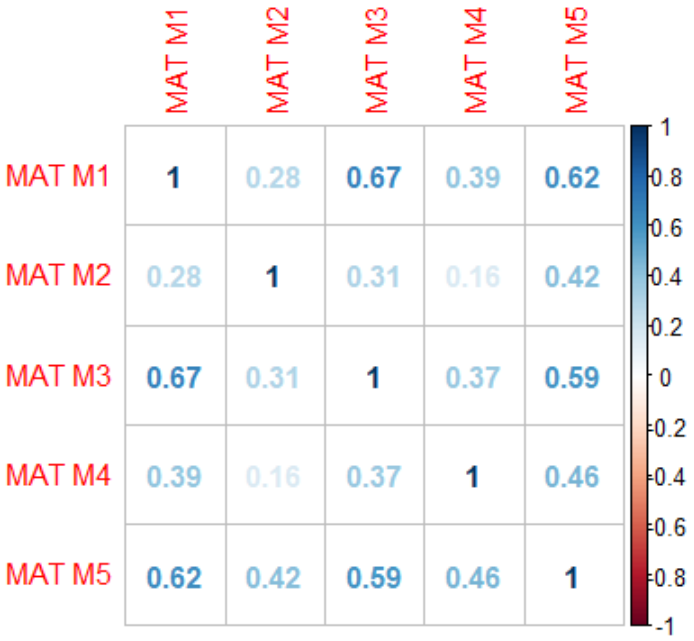

Figure 3.21: MAT correlation matrix

The residual plot in (figure 3.20) is also slightly U-shaped indicating some non-linearity. The residual plot shows that there is constant variance. The Q-Q plot suggest these residuals are normally distributed. The Shapiro-Wilk test indicated that these data provide insufficient

evidence that the residuals are not normally distributed ($W = 0.97874$, p-value= 0.3254). The Durbin-Watson test indicates that these data provide insufficient evidence that the residuals are autocorrelated ($DW = 1.8506$, p-value = 0.2476). The extreme values do not seem to be that influential (figure 3.20). The correlation plot (figure 3.21) suggests that none of the MAT subdomains are correlated. The VIF for the five NBT MAT subdomain test scores are $MAT1 = 2.146608$, $MAT2 = 1,229641$, $MAT3 = 2.014347$, $MAT4 = 1.314495$ and $MAT5 = 2.133283$. The VIF values are close to 1 and less than 5 which means these variables are not multicollinear.

The output of the regression model indicates that the adjusted R^2 value is 0.1819 implying that approximately 81% of the variation in the Statistics 101 test marks is not explained by variation in the NBT MAT subdomains test scores. This is very low and an indication that there are variables not included in this model that make a contribution to the Statistics 101 marks. There is a significant linear relationship between a Statistics 101 test score and the NBT MAT subdomains test scores ($F_{obs} = 3.846$, df= $5, 59$, p-value=0.004395). The intercept term makes a significant contribution to the model ($t_{obs} = 6.71006$, p-value < 0.0001), however the other MAT subdomains do not make a significant contribution to this model. This suggests that the MAT subdomains do not affect the dependent variable, but it does not mean that the dependent variable does not depend on independent variables at all.A stepwise regression analysis was used to find the best model based on all the MAT subdomains. Stepwise regression is used when there are a large number of independent variables that might have an effect on the response (Faraway, Julian J, 2014). The step function in R finds the best model by dropping insignificant variables one at a time. Backward, forward and both regression were used. Backward step regression was computed and this resulted in a better model, denoted as model 5. Model 5 consisted of two MAT subdomain variables, namely MAT3 and MAT4.

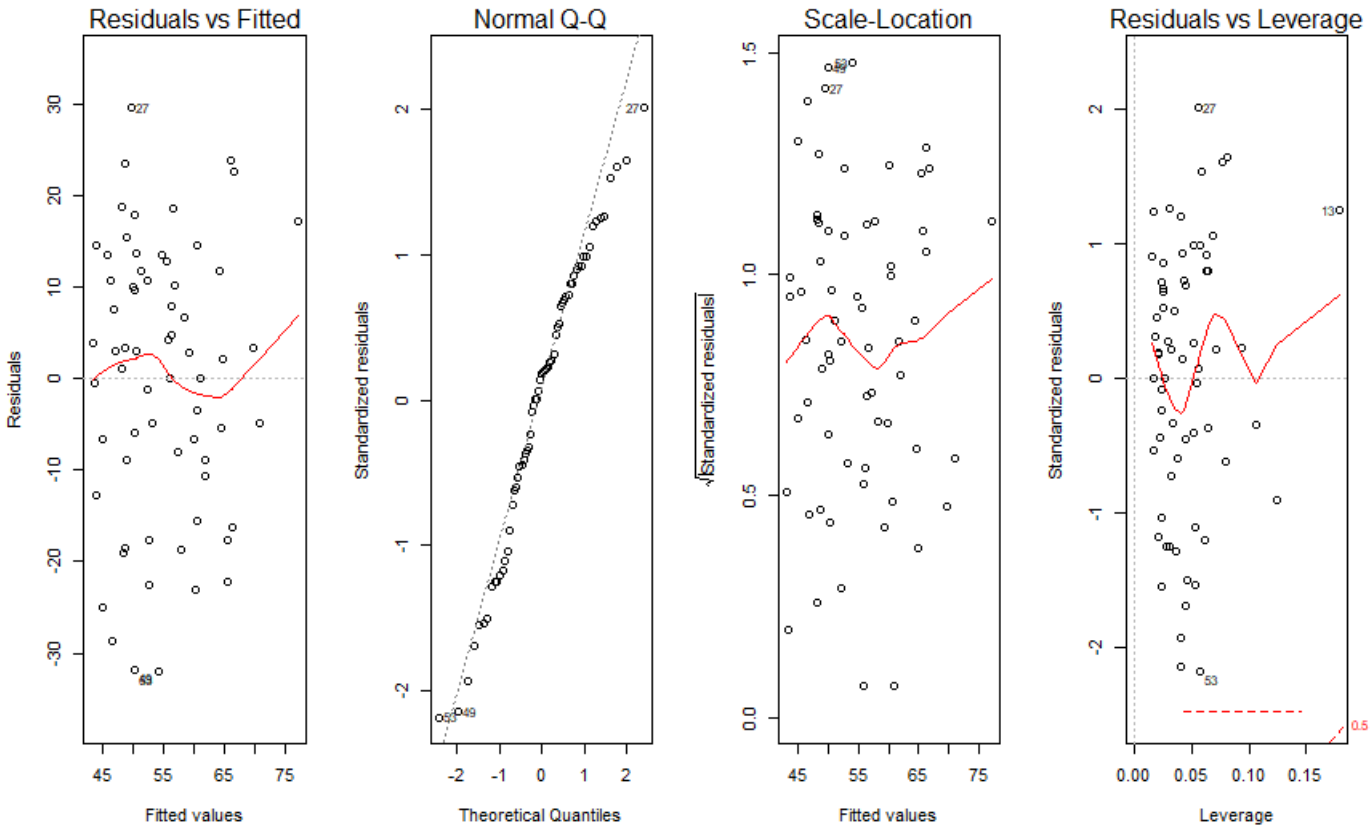

Figure 3.22: Diagnostic plots for model 5.

The residual plot for model 5 in (figure 3.22) is slightly U-shaped indicating some non-linearity. The residual plot shows that there is constant variance. The Q-Q plot suggest these residuals are normally distributed. The Shapiro-Wilk test indicated that these data provide insufficient evidence that the residuals are not normally distributed ($W = 0.97384$, p-value= 0.1835). The Durbin-Watson test indicates that these data provide insufficient evidence that the residuals are autocorrelated ($DW = 1.8513$, p-value $= 0.2642$). The extreme values do not seem to be that influential (figure 3.20). The correlation plot (figure 3.22) suggests that none of the MAT subdomains are correlated. The VIF for the two NBT MAT subdomain test scores are $MAT3 = 1.15992$, and $MAT4 = 1.15992$. The VIF values are close to 1 and less than 5 which means these variables are not multicollinear.

The output of the regression model indicates that the adjusted R^2 value is 0.1864 implying that approximately 81% of the variation in the Statistics 101 test marks is not explained by variation in the two NBT MAT subdomains test scores. This is very low and an indication that there are variables not included in this model that make a contribution to the Statistics 101 marks. There is a significant linear relationship between a Statistics 101 test score and the two NBT MAT subdomains test scores ($F_{obs} = 8.333$, df= $2, 62$, p-value=0.0006233). The intercept term makes a significant contribution to the model ($t_{obs} = 4.771$, p-value $<$ 0.0001) as does the MAT3 test score ($t_{obs} = 2.928$, p-value= 0.00477). The MAT4 does not make a significant contribution.

3.4.5 STA 101 as Linear Combination of the NBT subdomains

The Statistics 101 test results were modeled as a linear combination of the NBT subdomains test scores, denoted as model 6. The correlated QL subdomains (QL2 and QL4) were removed when the model was fitted.

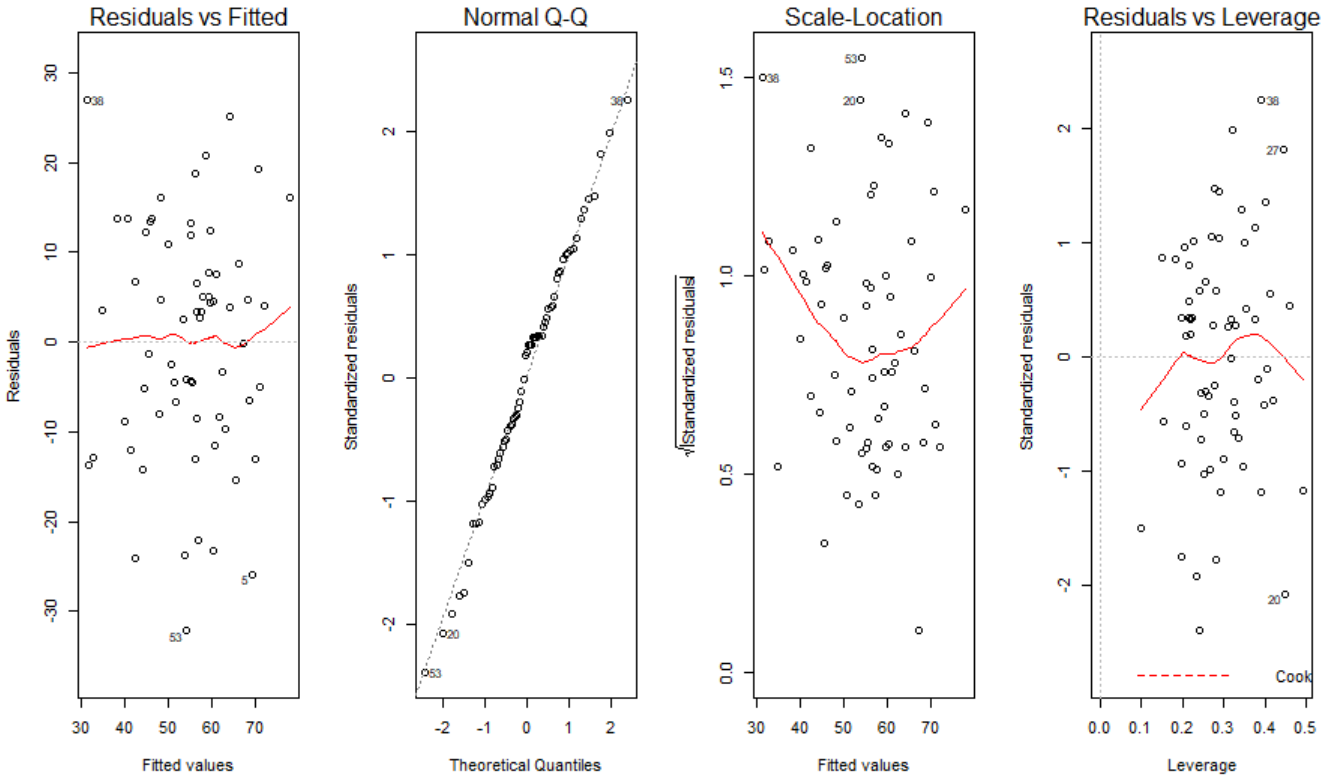

Figure 3.23: Diagnostic plots for model 5

The linear model assumptions were found to be acceptable. The output for model 6 indicates that the adjusted R^2 value is 0.1561 implying that approximately 84% of the variation in the Statistics 101 test marks is not explained by variation in the NBT subdomains test scores. There is no significant linear relationship between a Statistics 101 test score and all the NBT subdomains test scores ($F_{obs} = 1.658$, df= $18, 46$, p-value=0.08466). However the AL1 score makes a significant contribution to the model ($t_{obs} = 2.534$, p-value < 0.01), while the other NBT subdomains scores and the intercept term do not contribute to this model. Backward stepwise regression and both stepwise regression produced the same model, this model was better than the one produced by forward stepwise regression. This model was denoted as model 7. The variables selected are AL1, AL7, MAT3 and MAT4.

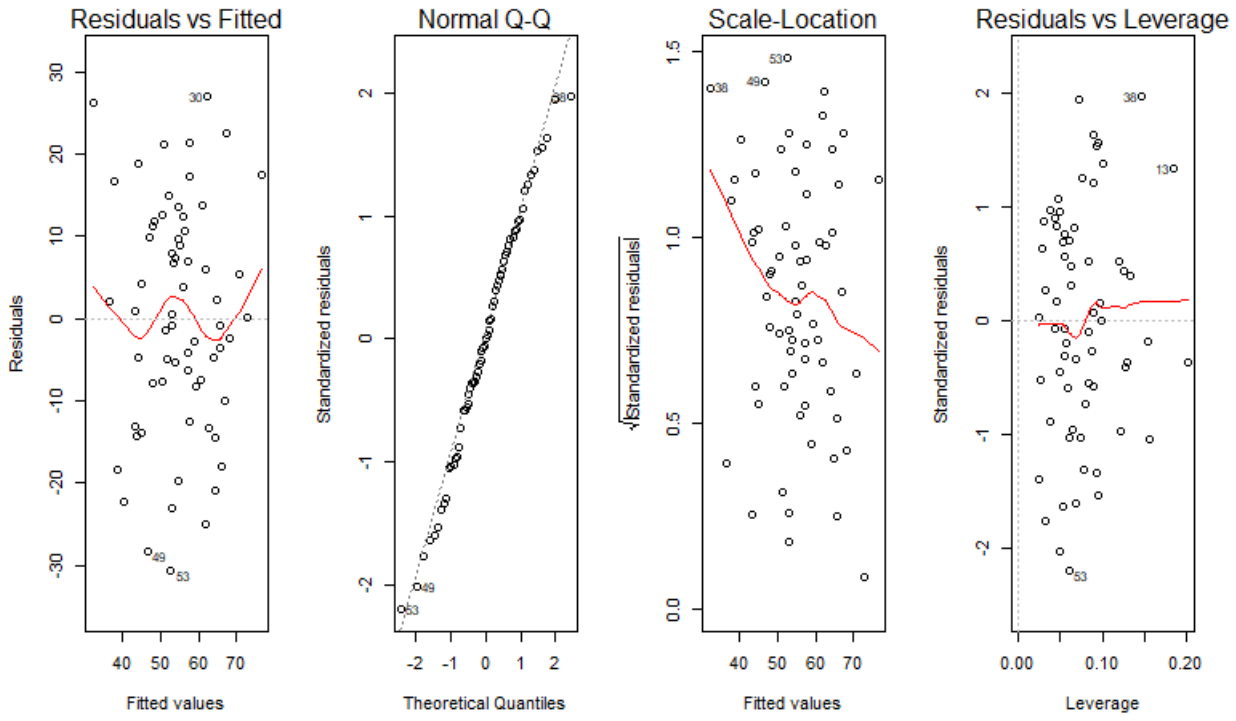

Figure 3.24: Diagnostics plot for model 6.

The residual plot (figure 3.24) is W-shaped indicating some non-linearity. The residual plot suggests that there is constant variance. The Q-Q plot suggest these residuals are normally distributed. The Shapiro-Wilk test indicates that these data provide insufficient evidence that the residuals are not normally distributed ($W = 0.984552$, p-value= 0.5926). The Durbin-Watson test indicates that these data provide insufficient evidence that the residuals are autocorrelated ($DW = 1.7782$, p-value $= 0.1516$). The extreme values do not seem to be that influential (figure 3.24). The VIF values of $AL1 = 1.76089$, $AL7 = 1.75612$, $MAT3 = 1.24200$ and $MAT4 = 1.16524$ were close to 1 and less than 5 which means that these variables are not multicollinear.

The output for model 7 indicates that the adjusted R^2 value is 0.2596 implying that approximately 74% of the variation in the Statistics 101 test marks is not explained by variation in these four NBT subdomains test scores. This is low and an indication that there are variables not included in this model that make a contribution to the Statistics 101 marks. There is a significant linear relationship between a Statistics 101 test score and these four NBT subdomains test scores ($F_{obs} = 6.609$, df= $4, 60$, p-value=0.001784). The intercept term makes a significant contribution to the model ($t_{obs} = 2.582$, p-value $= 0.01$) as does AL1 ($t_{obs} = 2.835$, p-value $= 0.00623$), AL7 ($t_{obs} = -2.013$, p-value $= 0.04859$) and MAT3 ($t_{obs} = 2.695$, p-value $= 0.00912$), however MAT4 does not make a significant contribution to this model.

3.5 Summary: Linear Regression Models

The following table summaries the regression models as discussed above.

Model	Independent Variables	Is the model significant?	Adjusted R^2value
1	AL, QL, MAT	Yes	0.1467
2	AL subdomains	No	0.01148
3	QL1, QL3, QL5, QL6	No	0.06693
4	MAT subdomains	Yes	0.1819
5	MAT3 and MAT4	Yes	0.1864
6	all the NBT subdomains	No	0.1929
7	AL1, AL7, MAT3 and MAT4	Yes	0.2596

Table 3.6: Linear regression models summaries.

3.6 Neural Network for regression

A Feed Forward Neural Network (FFNN) also know as a multilayer perceptron (MLP), was constructed to determine if the NBT test results can be used to predict the performance of a Statistics 101 student. FFNNs are artificial neural networks that have connections that are not recurrent (Burton, 2019). A MATLAB script was written, to load and organise the data, construct the neural network, train it and save it. The data was divided into three sets using the function dividerand(). This function separates the data into the training set, testing set and validation set, this is a set used to find the best model for the given problem (Burton, 2019). The dividerand function ensures that there is a higher percentage of training data (Burton, 2019). The default setting for dividerand is 60% training, 20% testing and 20% validation respectively. MATLAB was used to build the MLP. The sample size of 65 was already organised with features or variables as rows. Three layers were used with 10, 10 and 1 nodes respectively. The last layer is automatically determined by the targets. Once the MLP was trained, plots of the performance, training state and regression either at the end or during training were produced (figures 3.25 and 3.27). The regression plot shows the correlation coefficient, R, between the activation and target for training, validation and test sets. The assessment plots are shown in figures 3.26and 3.28.

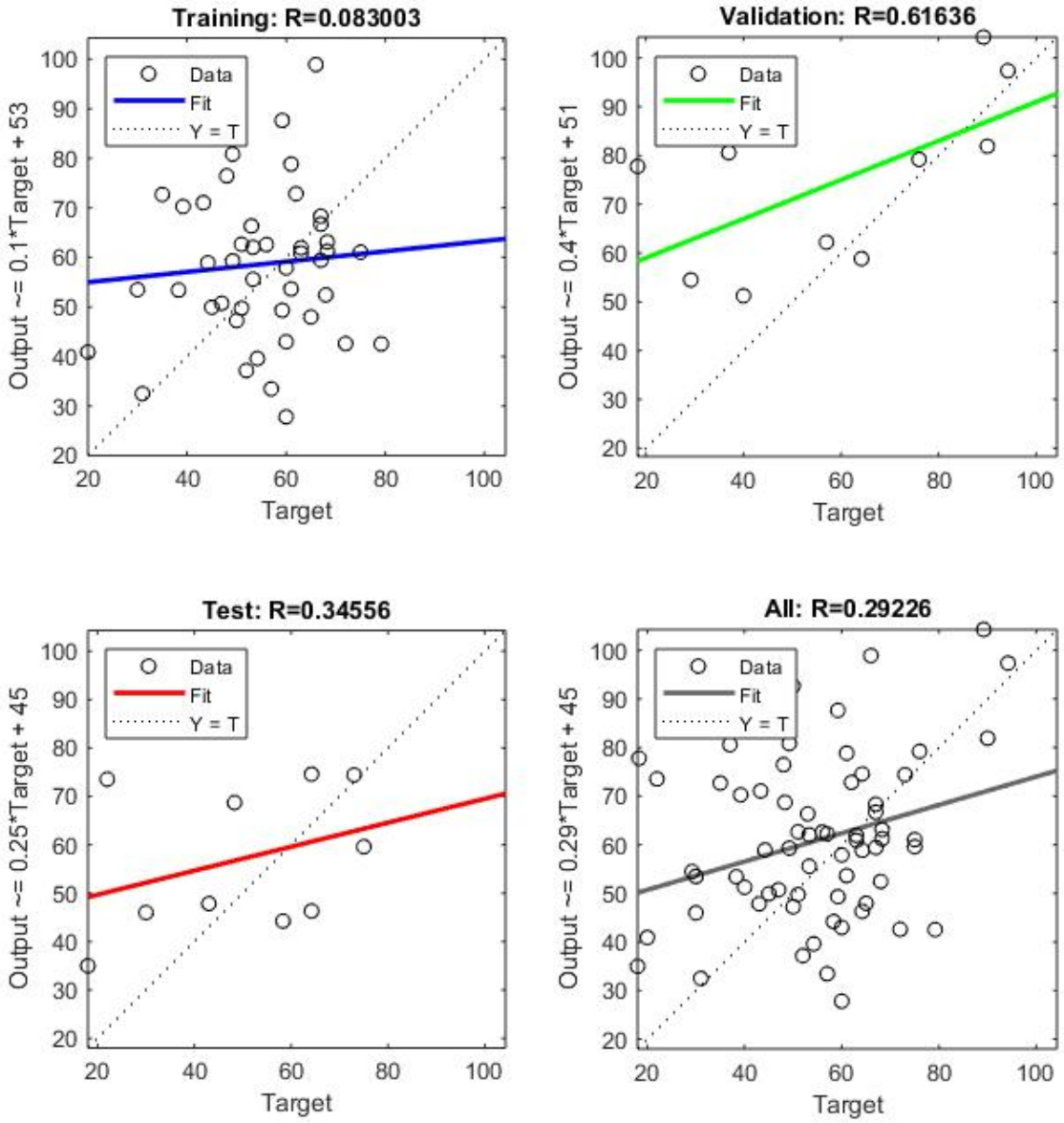

Figure 3.25: MLP output for the NBT test scores and Statistics test scores.

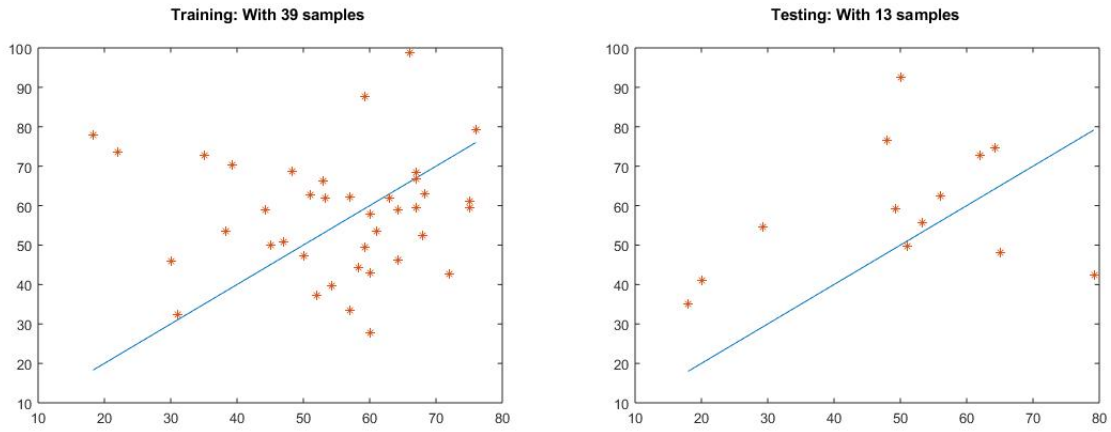

Figure 3.26: Assessment of the training set and the testing set fit.

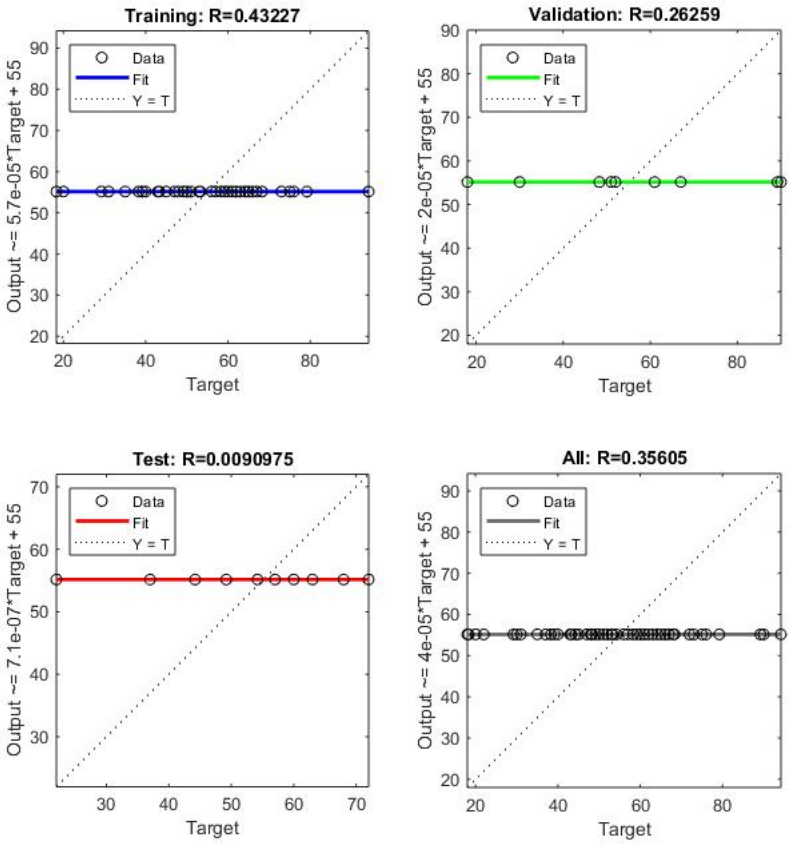

Figure 3.27: MLP output for the NBT subdomain test scores and Statistics test scores.

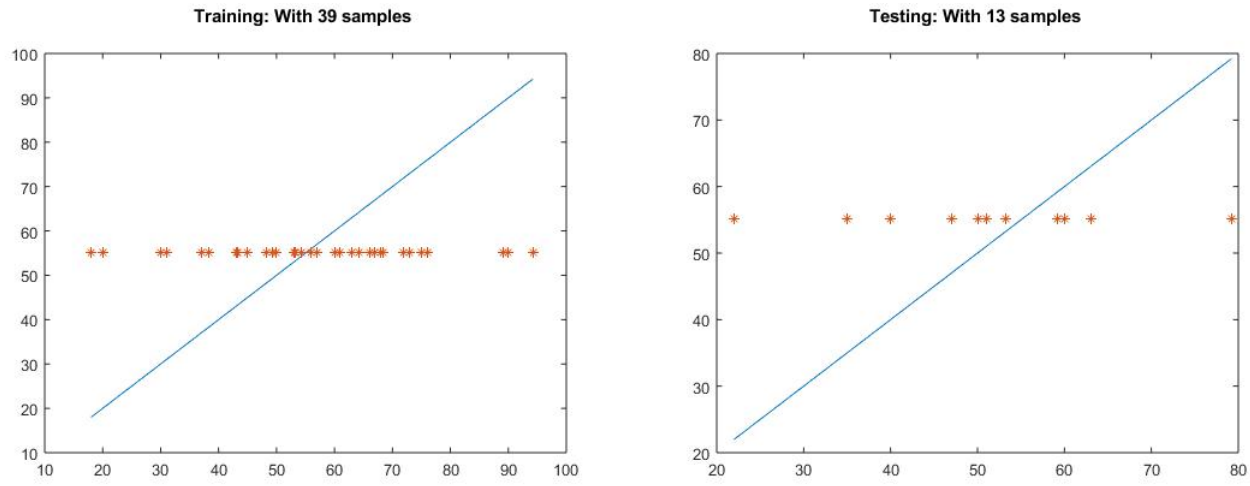

Figure 3.28: Assessment for the training set and the testing set fit.

The correlation coefficient, R, for all the data sets is low (figures 3.25 and 3.27). For the NBT test scores, namely AL, QL, MAT and the Statistics 101 test one score, the correlation coefficient is 0.29226, which is very low. There is a positive weak relationship between the NBT test scores and the Statistics 101 test one marks. For the NBT subdomain test scores and the Statistics 101 test one the correlation coefficient is 0.35605, which is also low. There is a positive weak relationship between the NBT subdomains test scores and the Statistics 101 test one marks. The second neural network with the NBT subdomains has a better R value but predicts the same value for every student. The prediction accuracy of the MLP for the Statistics 101 test one scores and the NBT scores was 0.0063, which is very low. The prediction of the STA 101 scores was overestimated for 13 students in the test data (see appendix 4). MLP for the Statistics 101 test one scores and the NBT subdomain scores was 0.004, which is also low. The prediction of the STA 101 scores was the same for 13 students in the test data (see appendix 4). The MLP results were used independently of the linear regression results because different training sample sizes were used.

Chapter 4

Conclusion

This study assessed if the NBT scores, namely Academic Literacy, Quantitative Literacy and Mathematics, can be used to predict academic performance in Statistics 101. The categorical analysis of the categorized NBT test scores and STA 101 test results indicated that there were no significant relationships between the STA 101 results and the AL results, the STA 101 results and the QL results and the STA 101 results and MAT results. Based on the literature reviewed this result was not expected.

The various regression models are summarised in table 3.6 on page 41. From this table, models 1, 4, 5 and 7 indicate that the Statistics 101 results can be predicted based on the NBT test results. However the low adjusted R^2 statistics clearly indicate that there are a number of variables not included in these models; a high percentage of the variation in Statistics 101 test 1 marks is not being explained by the various NBT test domain and or subdomain scores. This was not expected. Other important factors, including perhaps instruction in a non-home language or lecture attendance or the student being the first family member to be exposed to tertiary studies, need to be considered and possibly included in a linear regression model prior to predicting academic performance in Statistics 101 test 1.

The results from the MLP indicate that the model has very low prediction accuracy. This analysis showed it is not suggested that a student's Statistics 101 test 1 academic performance be predicted based on a MLP using the NBT test domains and subdomains as feature variables. This result is similar to that of Mahlobo, R. (2015) who found that the NBT mathematics and the NSC mathematics scores are not good predictors of first-year mathematics performance.

To conclude the NBT scores can be used to predict academic performance in Statistics 101 test 1, however the predictions are poor and inaccurate. It is suggested that future research be conducted to investigate if these predictions can be improved by including additional factors, for example lecture attendance, lecture engagement, tutorial attendance etc.

References

A. A. Wadee and A. Cliff (2016). Pre-admission tests of learning potential as predictors of academic success of first-year medical students. *South African Journal of Higher Education*, 30(2):264–278.

Baxter, J. (2019). Multivariate Statistical Analysis Course Notes: An Introduction to Clustering. Department of Statistics, Rhodes University.

Bewick, V., Cheek, L., and Ball, J. (2003). Statistics review 8: Qualitative data–tests of association. *Critical Care*, 8(1):46.

Burton, M. (2019). *Neural Networks Course Notes*. Department of Mathematics, Rhodes University.

CETAP (2018). The National Benchmark Tests National 2018 Report, Centre for Educational Testing for Access and Placement. `https://www.nbt.ac.za/`, Accessed March 2019.

Cliff, A. (2015). The National Benchmark Test in Academic Literacy: How might it be used to support teaching in higher education? *Language Matters*, 46(1):3–21.

CUT (2015). National Benchmark Test, Central University of Technolo, Free State. `https://www.cut.ac.za/nbt`, Accessed March 2019.

Faraway, J. J. (2016). *Extending the Linear Model with R: Generalized linear, mixed effects and nonparametric regression models*. Chapman and Hall/CRC.

Faraway, Julian J (2014). *Linear Models with R*. CRC Press.

Gerritsen, L. (2017). *Predicting student performance with Neural Networks*. PhD thesis, Tilburg University, The Netherlands. `https://0-scholar.google.co.za.wam.seals.ac.za/scholar?hl=en&as_sdt=0%2C5&q=Predicting+student+performance+with+Neural+Networks%2C+Gerritsen+&btnG=`, Accessed July 2019.

Ghasemi, A. and Zahediasl, S. (2012). Normality Tests for Statistical Analysis: a Guide for Non-Statisticians. *International Journal of Endocrinology and Metabolism*, 10(2):486.

James, G. and Witten, D. and Hastie, T. and Tibshirani, R. (2013). *An Introduction to Statistical Learning*. Springer.

Kumar, S. A. and Vijayalakshmi, M. N. (2011). Efficiency of Decision Trees in predicting students' academic performance. *International Journal of Computing*.

Mahlobo, R. (2015). National Benchmark Test as a Benchmark Tool. `http://uir.unisa.ac.za/bitstream/handle/10500/22446/Radley%20Mahlobo.pdf?sequence=1`, Accessed July 2019.

McCormick, R. E. and Tinsley, M. (1987). Athletics versus academics? Evidence from SAT scores. *Journal of Political Economy*, 95(5):1103–1116.

McHugh, M. L. (2013). The Chi-square test of independence. *Biochemia Medica: Biochemia Medica*, 23(2):143–149.

Mendes, M. and Pala, A. (2003). Type I Error Rate and Power of three Normality Tests. *Pakistan Journal of Information and Technology*, 2(2):135–139.

Moodley, P. and Singh, R. J. (2015). Addressing Student Dropout Rates at South African Universities. `http://hdl.handle.net/10321/1648`, Accessed July 2019.

Neethling, L. (2015). The determinants of academic outcomes: A competing risks approach. In *Conference of the Economic Society of South Africa, Cape Town: University of Cape Town*.

Osmanbegovic, E and Suljic, M (2012). Data mining approach for predicting student performance. *Economic Review: Journal of Economics and Business*, 10(1):3–12.

Prince, R. (2017). The relationship between school-leaving examinations and university entrance assessments: The case of the South African system. *Journal of Education*, (70):133–160.

Scott, I. and Yeld, N. and Hendry, J. (2007). A Case for improving teaching and learning in South African higher education. *Higher Education Monitor*, 6(2):1–8.

Spaull, N. (2013). South Africa's Education Crisis: The quality of Education in South Africa 1994-2011. *Johannesburg: Centre for Development and Enterprise*, pages 1–65.

Surampudi, S. (2015). *Oracle Data Mining Concepts*. `https://docs.oracle.com/database/121/DMCON/title.htm`, Accessed July 2019.

Uyanık, G. K. and Güler, N. (2013). A study on multiple linear regression analysis. *Procedia-Social and Behavioral Sciences*, 106:234–240.

Van Schalkwyk, S. and Bitzer, Eli and Van der Walt, C. (2010). Acquiring Academic Literacy: A Case of first-year Extended Degree Programme Students. *Southern African Linguistics and Applied Language Studies*, 27(2):189–201.

Wilson-Strydom, M. (2012). Using the NBTs to inform institutional understandings of 'under-preparedness': Implications for admissions criteria. *South African Journal of Higher Education*, 26(1):136–151.

Wurf, G. and Croft-Piggin, L. (2015). Predicting the academic achievement of first-year, pre-service teachers: the role of engagement, motivation, atar, and emotional intelligence. *Asia-Pacific Journal of Teacher Education*, 43(1):75–91.

Yeld, N. (2007). Critical questions? Some responses to issues raised in relation to the National Benchmark Tests project. *South African Journal of Higher Education*, 21(5):610–616.

Yoo, W. and Mayberry, R. and Bae, S. and Singh, K. and He, Q. P. and Lillard Jr, J. W. (2014). A Study of Effects of Multicollinearity in the Multivariable analysis. *International Journal of Applied Science and Technology*, 4(5):9.

R Code

Model 1 :

```
Call:
lm(formula = testmark ~ ALscore + QLscore + MATscore)

Residuals:
     Min       1Q   Median       3Q      Max
-36.381   -9.248    1.581   10.677   31.549
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 35.231056   8.587052   4.103 0.000123 ***
ALscore     -0.003746   0.194917  -0.019 0.984731
QLscore     -0.117659   0.193772  -0.607 0.545972
MATscore     0.600878   0.186611   3.220 0.002056 **

Residual standard error: 15.49 on 61 degrees of freedom
Multiple R-squared:  0.1867,    Adjusted R-squared:  0.1467
F-statistic: 4.667 on 3 and 61 DF,  p-value: 0.005322
```

Model 2:

```
Call:
lm(formula = NBTSdata$TestMark ~ NBTSdata$AL1 + NBTSdata$AL2 +
    NBTSdata$AL3 + NBTSdata$AL4+ NBTSdata$AL5 +
    NBTSdata$AL6 + NBTSdata$AL7 + NBTSdata$AL8
    + NBTSdata$AL9)

Residuals:
     Min       1Q   Median       3Q      Max
-32.154   -8.351   -0.682    9.854   39.790
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  40.75383    9.05617   4.500 3.56e-05 ***
NBTSdata$AL1  0.42570    0.15466   2.753    0.008 **
```

```
NBTSdata$AL2 -0.05531    0.12362  -0.447    0.656
NBTSdata$AL3 -0.02278    0.11048  -0.206    0.837
NBTSdata$AL4  0.01719    0.12613   0.136    0.892
NBTSdata$AL5 -0.08885    0.09500  -0.935    0.354
NBTSdata$AL6  0.09767    0.17614   0.555    0.581
NBTSdata$AL7 -0.13563    0.13557  -1.000    0.321
NBTSdata$AL8  0.02403    0.08916   0.270    0.789
NBTSdata$AL9 -0.03985    0.08443  -0.472    0.639


Residual standard error: 16.67 on 55 degrees of freedom
Multiple R-squared:  0.1505, Adjusted R-squared:  0.01148
F-statistic: 1.083 on 9 and 55 DF,  p-value: 0.3904
```

Model 3:

```
Call:
lm(formula = NBTSdata$TestMark ~ NBTSdata$QL1 + NBTSdata$QL3
                    + NBTSdata$QL5 + NBTSdata$QL6)
Residuals:
     Min       1Q   Median       3Q      Max
-34.009  -8.503   2.300  10.201  31.343
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  38.54083    7.77047   4.960 6.12e-06 ***
NBTSdata$QL1  0.02195    0.13709   0.160   0.8733
NBTSdata$QL3 -0.06851    0.07899  -0.867   0.3892
NBTSdata$QL5  0.23292    0.10806   2.156   0.0351 *
NBTSdata$QL6  0.12399    0.13859   0.895   0.3745


Residual standard error: 16.2 on 60 degrees of freedom
Multiple R-squared:  0.1252, Adjusted R-squared:  0.06693
F-statistic: 2.148 on 4 and 60 DF,  p-value: 0.0859
```

Model 4:

```
Call:
lm(formula = NBTSdata$TestMark ~ NBTSdata$MAT1 + NBTSdata$MAT2
            + NBTSdata$MAT3 + NBTSdata$MAT4 + NBTSdata$MAT5)


Residuals:
     Min       1Q   Median       3Q      Max
-33.571  -7.464   1.484  10.625  29.473
```

```
Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)    26.88213    6.71006   4.006 0.000175 ***
NBTSdata$MAT1   0.18132    0.15517   1.169 0.247279
NBTSdata$MAT2   0.10440    0.09134   1.143 0.257695
NBTSdata$MAT3   0.22468    0.14583   1.541 0.128725
NBTSdata$MAT4   0.16697    0.12143   1.375 0.174337
NBTSdata$MAT5  -0.09279    0.14657  -0.633 0.529153


Residual standard error: 15.17 on 59 degrees of freedom
Multiple R-squared:  0.2458,    Adjusted R-squared:  0.1819
F-statistic: 3.846 on 5 and 59 DF,  p-value: 0.004395
```

Model 5:

```
Call:
lm(formula = NBTSdata$TestMark ~ NBTSdata$MAT3 + NBTSdata$MAT4)

Residuals:
     Min       1Q   Median       3Q      Max
-32.058  -8.973   2.751  11.744  29.580
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   30.2833     6.3480   4.771 1.16e-05 ***
NBTSdata$MAT3  0.3231     0.1104   2.928  0.00477 **
NBTSdata$MAT4  0.1769     0.1138   1.555  0.12505

Residual standard error: 15.13 on 62 degrees of freedom
Multiple R-squared:  0.2119, Adjusted R-squared:  0.1864
F-statistic: 8.333 on 2 and 62 DF,  p-value: 0.0006233
```

Model 6:

```
Call: lm(formula = NBTSdata$TestMark ~ NBTSdata$AL1+NBTSdata$AL2
                 + NBTSdata$AL3 + NBTSdata$AL4 + NBTSdata$AL5
                 + NBTSdata$AL6 + NBTSdata$AL7 + NBTSdata$AL8
                 + NBTSdata$AL9 + NBTSdata$QL1 + NBTSdata$QL3
                 + NBTSdata$QL5 + NBTSdata$QL6 + NBTSdata$MAT1
                 + NBTSdata$MAT2 + NBTSdata$MAT3
                 + NBTSdata$MAT4 + NBTSdata$MAT5)

Residuals:
```

```
      Min       1Q   Median       3Q      Max
-32.153   -8.426    2.751    8.755   26.979
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  20.318345  11.383157   1.785   0.0809 .
NBTSdata$AL1  0.378320   0.149280   2.534   0.0147 *
NBTSdata$AL2 -0.118861   0.127996  -0.929   0.3579
NBTSdata$AL3 -0.069199   0.112395  -0.616   0.5411
NBTSdata$AL4  0.006376   0.121634   0.052   0.9584
NBTSdata$AL5 -0.060709   0.095832  -0.633   0.5295
NBTSdata$AL6  0.117970   0.175878   0.671   0.5057
NBTSdata$AL7 -0.220407   0.151273  -1.457   0.1519
NBTSdata$AL8 -0.045262   0.094724  -0.478   0.6350
NBTSdata$AL9  0.014759   0.085107   0.173   0.8631
NBTSdata$QL1  0.048861   0.145564   0.336   0.7386
NBTSdata$QL3 -0.071030   0.085530  -0.830   0.4106
NBTSdata$QL5  0.184170   0.126000   1.462   0.1506
NBTSdata$QL6  0.021379   0.163622   0.131   0.8966
NBTSdata$MAT1 0.171853   0.185865   0.925   0.3600
NBTSdata$MAT2 0.069725   0.103884   0.671   0.5055
NBTSdata$MAT3 0.096454   0.166212   0.580   0.5645
NBTSdata$MAT4 0.166909   0.140846   1.185   0.2421
NBTSdata$MAT5 -0.006832  0.188952  -0.036   0.9713

Residual standard error: 15.41 on 46 degrees of freedom
Multiple R-squared:  0.3934, Adjusted R-squared:  0.1561
F-statistic: 1.658 on 18 and 46 DF,  p-value: 0.08466
```

Model 7:

```
Call:
lm(formula = TestMark ~ AL1 + AL7 + MAT3 + MAT4)

Residuals:
      Min       1Q   Median       3Q      Max
-30.7534  -8.2286   0.1063  10.6095  27.0183

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 20.2750     7.8532   2.582  0.01229 *
AL1          0.3573     0.1260   2.835  0.00623 **
```

AL7	−0.2032	0.1010	−2.013	0.04859 *
MAT3	0.2936	0.1089	2.695	0.00912 **
MAT4	0.1898	0.1088	1.745	0.08615 .

Residual standard error: 14.43 on 60 degrees of freedom
Multiple R–squared: 0.3058, Adjusted R–squared: 0.2596
F–statistic: 6.609 on 4 and 60 DF, p–value: 0.0001784

Matlab Code

.1 mathlabNN.m

```
clear
close all
clc
%data importing  data = xlsread('NBTStatistics.xlsx','FactorStatsT');
P=data(:,[ 2:21]); T=data(:,[1]);
p=P';t=T';
%network building %neurons in layers 1, 2,3
s1=input('number of neurons=');
s2=input('number of neurons=');
s3=input('number of neurons=');
% s1=3; s2=1
%create the net net=feedforwardnet([s1, s2,s3]);
%training function
net.trainFcn='trainbr';
%Batch training with weight and bias learning rules Bayesian regularization
%maxit net.trainParam.epochs=200;
%set the number of epochs that the error on the validation set increases
net.trainParam.max_fail=20;
%set ratio using:
[ptrain,pval,ptest,trainInd,valInd,testInd] = dividerand(p,0.6,0.2,0.2);
[ttrain,tval,ttest] = divideind(t,trainInd,valInd,testInd);
%initiate
net=init(net);
%assessing the degree of fit %%%%%%%%%%%%%%%%%%%%%%%%%%%%%
%train
[net,netstruct]=train(net,p,t);
%name the net and structure
net.userdata='NBT'; NBTnet=net;
NBTstruct=netstruct;
```

```
%batch sizes q1=size(ptrain,2);
q2=size(ptest,2);
%simulate atrain=sim(NBTnet,ptrain);
%train
atest=sim(NBTnet,ptest);
%test
a=sim(NBTnet,p);
%all
%train
r2=rsq(ttrain,atrain);
[R,PV]=corrcoef(ttrain,atrain);
%figures
%training figure
plot(ttrain,ttrain,ttrain,atrain,'*')
title(sprintf('Training: With %g samples \n',q1))
disp('train')
disp('activation target')
M=[atrain ;ttrain];
fprintf('%4.1f\t\t\t%4.1f\n',M)
%test:
r2test=rsq(ttest,atest);
[R,PV]=corrcoef(ttest,atest);
fprintf('Testing:\n\n')
fprintf(' corr coeff: %g\n p value: %g\n r2: %g\n',R(1,2),PV(1,2),r2test)
figure
plot(ttest,ttest,ttest,atest,'*')
title(sprintf('Testing: With %g samples \n',q2))
disp('test')
disp('activation target')
M=[atest ;ttest];
fprintf('%4.1f\t\t\t%4.1f\n',M)
%all r2all=rsq(t,a);
[R,PV]=corrcoef(t,a);
fprintf('All:\n\n')
fprintf(' corr coeff: %g\n p value: %g\n r2: %g\n',R(1,2),PV(1,2),r2all)
figure
plot(t,t,t,a,'*')
title(sprintf('All: With %g samples \n',q1+q2))
disp('all')
disp('activation target')
```

```
M=[a ;t];
fprintf('%4.1f\t\t\t%4.1f\n',M)
plotperform(netstruct)
[R,pv]=corrcoef(ttest,atest)
save NBTnet.mat
```

.2 MatlabNN2.m

```
clear
close all
clc
%data importing
data = xlsread('NBTStatistics.xlsx','Sheet3');
P=data(:,[ 2:4]); T=data(:,[1]);
p=P';t=T';
%network building
%neurons in layers 1, 2 ,3
s1=input('number of neurons=');
s2=input('number of neurons=');
s3=input('number of neurons=');
%create the net
net=feedforwardnet([s1, s2,s3]);
%training function net.trainFcn='trainbr';
%Batch training with weight and bias learning rules Bayesian regularization
%maxit net.trainParam.epochs=200;
%set the number of epochs that the error on the validation set increases
net.trainParam.max_fail=20;
%set ratio using:
[ptrain,pval,ptest,trainInd,valInd,testInd] = dividerand(p,0.6,0.2,0.2);
[ttrain,tval,ttest] = divideind(t,trainInd,valInd,testInd);
%initiate net=init(net);
%assessing the degree of fit %%%%%%%%%%%%%%%%%%%%%%%%%%%%%%% %train
[net,netstruct]=train(net,p,t);
%name the net and structure
net.userdata='nbnet1'; nbnet12net=net;
nbnet11struct=netstruct;
%batch sizes
q1=size(ptrain,2);
q2=size(ptest,2);
%simulate
```

```
atrain=sim(nbnet12net,ptrain);
%train
atest=sim(nbnet12net,ptest);
%test
a=sim(nbnet12net,p);
%all
%train
r2=rsq(ttrain,atrain);
[R,PV]=corrcoef(ttrain,atrain);
%figures
%training figure plot(ttrain,ttrain,ttrain,atrain,'*')
title(sprintf('Training: With %g samples \n',q1))
disp('train')
disp('activation target')
M=[atrain ;ttrain];
fprintf('%4.1f\t\t\t%4.1f\n',M)
%test: r2test=rsq(ttest,atest);
[R,PV]=corrcoef(ttest,atest);
fprintf('Testing:\n\n')
fprintf(' corr coeff: %g\n p value: %g\n r2: %g\n',R(1,2),PV(1,2),r2test)
figure
plot(ttest,ttest,ttest,atest,'*')
title(sprintf('Testing: With %g samples \n',q2))
disp('test')
disp('activation target')
M=[atest ;ttest];
fprintf('%4.1f\t\t\t%4.1f\n',M)
%all
r2all=rsq(t,a);
[R,PV]=corrcoef(t,a);
fprintf('All:\n\n')
fprintf(' corr coeff: %g\n p value: %g\n r2: %g\n',R(1,2),PV(1,2),r2all)
figure
plot(t,t,t,a,'*')
title(sprintf('All: With %g samples \n',q1+q2))
disp('all')
disp('activation target')
M=[a ;t];
fprintf('%4.1f\t\t\t%4.1f\n',M)
plotperform(netstruct)
```

```
[R,pv]=corrcoef(ttest,atest)
save nbnet12net.mat
```