Statistical Modelling of the Relationship between Learner Support Intervention and Matric Pass Rates in Letlhabile Area, North-West Province of South Africa

by

JOSEPH NKASHE MATSHEGO

Proposal submitted in fulfilment of the requirements for the degree

MASTER OF SCIENCE

in the

Department of Statistics and Operations Research

At the University of Limpopo

Medunsa Campus

Supervisor: Dr S.M. Seeletse

2012

ABSTRACT

This study used statistical methods to determine the relationship between matric pass rates and interventions in the schools in the Letlhabile area. Nine schools were identified in this area. Five years (2007 to 2011) were looked at. Regressions methods were tried in which various forms of relationships were compared. The methods were linear, curvilinear (quadratic to polynomial of 4^{th} power), exponential and power regressions were used in the tentative models investigated. The measures of bias and precision were used to compare the models. Multicollinearity was also investigated where it was possible. Time series analysis was used to illustrate the trend patterns of the pass rates in the various schools as well as the pattern of the numbers of interventions. In most of these schools the numbers of interventions increased over the five years and in only a few schools the number of interventions decreased over these years. A highlight of this study is that more interventions enhanced the matric pass rates. It was also evident that in the years in which the interventions decreased, the pass rates also decreased. The regression methods investigated were all showing to be applicable in the prediction of pass rates from the numbers of interventions. This was concluded from realising that the measures of bias, precision and quality all ratified them. The methods were compared in order to select the best one based on the measures. The linear regression in which the pass rates are regressed on the number of interventions came out as the leading model in terms of all the criteria used.

ACKNOWLEDGEMENTS

It would not have been possible for me to complete this study without some few peoples' assistance. While it is not possible to acknowledge everyone who was involved in one way or the other, a few persons are worth a mention here. First my most gratitude goes to my supervisor Dr Solly Matshonisa Seeletse who encouraged me to continue with my studies. I understand why he would say refine this paragraph. Thank You Sir. May God bless you and your family. Ke a leboga Ntate.

Special thanks to my younger brother, Tabane Matshego, who help me get started with this project, My Mother Naniki Matshego who always encouraged me to continue with my studies and my sister Debra Mmushi for the candid support she gave me during those difficult times. Special thanks to Mr Motshwane, Acting Head of the Department of Statistics and Operations Research at the Medunsa Campus of the University, who would say never give-up.

The special word of thanks goes to the Department of Education Northwest Province for giving me permission to conduct research in schools in Letlhabile Area Office, the schools principals for helping me with the data for this research and to the school governing bodies, management teams and, the wonderful staff members which cooperated during my studies.

In conclusion I would like to thank the Almighty God and my ancestors for having made this possible.

DECLARATION

This research project is my original work. Unless otherwise specifically stated, all the references cited have been consulted. The work of which the dissertation is a record has been done by me and has not been previously accepted for a higher degree or professional qualification at any other higher education institution

Signed

……

Matshego Joseph Nkashe

This dissertation has been submitted with my approval as University Supervisor and would certify that the requirement for the applicable Masters in Statistics' rules and regulation have been fulfilled

……

SM Seeletse, PhD, DBA

LIST OF ABBERVIATIONS

AR	=	autoregressive
ARCH	=	autoregressive conditional heteroskedasticity
ARIMA	=	autoregressive integrated moving average
ARMA	=	autoregressive moving average
ARFIMA	=	autoregressive fractionally integrated moving average
ARIMA	=	autoregressive integrated moving average
CFE	=	cumulative forecast error
I(*d*)	=	integrated of order *d*
MA	=	moving average
MAD	=	mean absolute deviation
MAE	=	mean absolute error
ME	=	mean error
MAPE	=	mean absolute percentage error
MSE	=	mean squared error
RMSE	=	root mean square error
SLR	=	simple linear regression
SSE	=	sum of squares for error
VIF	=	variance-inflation factor

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1: PROLOGUE

1.1 Introduction

The pass rates at the exit level of high school in South Africa hold the key to the entry of learners to a path to become tertiary education graduates, and in some cases to careers of the interest of learners. Over the years in the history of education in South Africa, this exit level, known as matric, has been a nightmare for learners with ambitions to follow their chosen career paths when they grow up. It was at matric where most students failed and could not pursue their education further. In some cases intervention programmes were introduced to help improve the matric pass rates. These interventions occurred as remedial programmes to help where the schools showed deficiencies. These programmes enhanced motivation of learners to study with more confidence and relaxation. Intervention, just like motivation, is believed to enhance school performance. In particular, experience shows that matric pass rates are higher in schools where learners and educators are more motivated. In South Africa unfortunately, schools are not equally resourced. Some are less resourced or equipped than others. It is, however, still necessary to find scientific explanations that would enhance performance of school learners even in schools that are not well resourced because history shows that poorly resourced schools of the past could still manage to produce national icons that are visible sometimes even more than those coming from fully resourced schools.

A pass in matric is vital, firstly as a gateway to work and careers, and secondly as a bridge to tertiary education. It is also used as access to higher learning. Applicants competing for space at higher education institutions are admitted in higher learning institutions of their choice ahead of others if they can demonstrate superior performance at matric. In South Africa, the most preferred contact higher learning institutions include the Universities of the Witwatersrand, Cape Town, Johannesburg, Pretoria, KwaZulu-Natal, Stellenbosch, and Rhodes. Those who lose to leading applicants at these universities would probably opt for the University of South Africa (Unisa), which is also a preferred university for its perceived high standard through distance learning and not through contact. There are other good universities in South Africa, but due to shorter history and less resources, they are least preferred, mainly because of perceptions. Nevertheless, matric certificate is the main requirement for admission in all of them. Without matric there is no admission into higher learning.

Thus, matric achievement is the ultimate goal at school education. For instance, in families without parents, oldest sibling children may need to use the limited family resources to earn matric, find a job and then help the younger siblings to go to school. People from poor families with no money to further their studies beyond matric could just want to reach and obtain a matric certificate. Others may just intend to obtain matric and find a job, instead of pursuing higher education. Thus, for these and other learners with ambitions to further their studies, matric certificate is enviable to have, both as a need and as a desire.

1.2 Background

Despite being desirable to have matric, for black communities who had a difficult education system, matric results have been a historical nightmare since the days of apartheid. Passes at matric were scarce for most schools due to lack of support and resources. The resources (such as libraries and books, good teachers and so on) that are necessary to improve pass rates and to equip learners were commonly lacking in these schools. These schools still remain less resourced compared to former white schools and private schools. As a result they do not have physical or tangible resources to enhance high performance at matric grade.

Obviously, the high passes were enhanced by consistent hard work and focus. These are often the outcomes of intervention or remedial programmes. Often learners experiencing remedial exercises tend to be stimulated to work and focus more on their studies. Intervention comes in different forms and frequencies. In the interest of improving matric results (and general performance in school) it is essential to find scientific relationships between the number of interventions and matric pass rates. Any model developed for this purpose should enhance pass rates, be theoretically sound and be pragmatic. This means that the model should be easy to use (i.e. a *black box* is not required) by educators. This study intends to develop a statistical model that associates/links intervention with achievement in the schools. Students may be discouraged to study if they know that they lack resources that are used elsewhere to enhance good performance. This study wants to find ways to offset such possibilities, and instead determine useful factors to ensure high matric pass rate.

1.3 Research Problem

There are no known models that represent a connection between the number of interventions and matric results in the rural areas of South Africa whose aim is to improve performance of learners in their matric studies.

1.4 Aim and Objectives

1.4.1 Aim

The aim of this study is to develop a statistical model that would signify a relationship between the number of interventions and performance at matric level for secondary and high schools in the Letlhabile area.

1.4.2 Objectives

The objectives were:

- To determine methods to present the numbers of interventions in a scientifically agreeable manner for use in enhancing matric pass rates.
- To develop a mathematical model that relates the numbers of interventions with matric performance

1.5 Hypothesis

This study hypothesised that high performance and achievement in schools are possible even without the required tools (finances, libraries, books and teachers, etc.) that make the most resourced schools succeed. It argues that intervention is one desirable approach to education to be applied in this study to ensure that the pass rates remain high.

1.6 Context

The area of focus was Letlhabile district. This area office was located in the township of Letlhabile in the North-West Province of South Africa. Letlhabile is about 40 km north-west of Pretoria, the capital city of South Africa, in the Gauteng Province. The four circuits of Letlhabile are Thuto Lesedi, Retlakgona, Toloane and Morula. Retlakgona is about 15 km on

the northern side of Letlhabile township. Thuto Lesedi is a township in Letlhabile. Toloane and Morula circuit are housed at Madidi Resource Centre which is about 35 km to the eastern side of Letlhabile. Letlhabile is a township, meaning that it is semi-developed, but the other areas around it are undeveloped rural areas. Intervention measures of matric learners initiated inside school campuses or that may come from outside the school is initiated by the school itself, usually by the educators dealing with careers or psychological services. It is understood that schools motivate their learners to work hard.

1.7 Methods

The study uses census of all the high schools presenting matric in the Letlhabile area. There are 19 of these. In the different circuits they are distributed as follows: five are in Morula circuit, five in the Retlakgona circuit, five in the Toloane circuit and four are in the Thuto-Lesedi circuit. Data consisted of matric pass rates in different years and the numbers of interventions that the learners in these schools were exposed to during their studies. There were comparisons of the various circuits as well as clustering of the schools according to similarities in the way they are motivated. Correlations where necessary, were also used in the establishment of the comparisons and relationships.

1.8 Benefits of the Study

The contributions of the study were anticipated at the theoretical and applied levels of statistics. These were as follows:

Contributions to theoretical statistics

- The study intends
    - To establish recording of number of intervention exercises for matric learners in relation to their pass rates.
    - To develop a model in which the number of interventions enhances improvement in matric results.
    - Since no statistical model exists for a relationship between performance and number of interventions at school, this served as a contribution to initiate that modelling.

Contributions to applied statistics

- Use of a real example can benefit the application of the scientific methods of educator selection in this study.
- The study intends to demonstrate use of numbers of interventions in enhancing improvement in matric results.
- A successful study is of benefit to schools that fall in the same category of lacking resources and needing a boost to enhance performance.

1.9 Studies on Pass Rates

Dealing with various relationships of pass rate regarding the Grade 12 learners is a positive initiative. According to Maharaj and Gokal (2006), a considerable amount of work has been published on the relationship between school leaving results and the success at various first-year university courses. Other scholars who have studied predictions of performance using other variables include Butcher and Muth (1985), Campbell and McCabe (1984), Golding and McNamarah (2005) as well as Kruck and Lending (2003). All these studies show that the pass rates at university are related to the level of performance at school level. However, no literature is available on the correlation between the number of interventions and performance at school level, which is an early entry for preparedness of the students for entry into higher education. Those authors made statements to motivate for a study on the relationship of matric passes and performance at university. Other related studies investigated the relationship between language results and university performance.

One attention-grabbing study by Rauchas *et al*. (2006) showed that high school first language results correlate better with university results than do the high school mathematics results. Other investigations investigated factors that cause student failure. Africa (2005) investigated reasons for and causes of failure of African students at the University of KwaZulu-Natal in March to April of the year 2005. That study did not find a single outstanding reason. Instead it exposed a mixture of several reasons as contributory factors for the abnormally high failure rate. The major reasons for this high failure rate were a combination of personal, financial, institutional, attitudinal, racial and academic reasons.

Campbell and McCabe (1984) studied the statistical relationship between a student's entrance characteristics and his/her success in the first year of a computer science major. That study found that students who majored in the sciences differed from those who left computer science for other degrees. These differences were related to the student's background in mathematics and science. Gender was not an achievement indicator, but it persistently appeared as a variable in their classification models.

Kruck and Lending (2003) described a model to predict academic performance in an introductory college level information systems course. They hypothesised that academic performance is affected by gender. This hypothesis was not supported by any tests or data, and hence it was rejected.

International researchers also contributed to studies of relationships between school performance and tertiary studies. In Jamaica, Golding and McNamarah (2005) investigated the existence of relationships between students' personal attributes and other factors, and their performance in the School of Computing and Information Technology. This was undertaken at the University of Technology in Jamaica. They found that mathematics is a weak predictor of performance in Information Technology.

This study did not intend to focus on a single subject. Instead it aimed at focusing on the way overall school performance can benefit from intervention. It was therefore vital to determine interventions of significance to make a difference in matric performance.

1.10 Study Layout

Chapter 1 motivated the research by pointing at the problem being addressed, the study aim and objectives, the methods used, and the significance of the study.

Chapter 2 discussed various regression methods and formulae used in the study. It also presented necessary mathematical results used.

Chapter 3 provided time series methods as well as measures of bias.

Chapter 4 presented the data analyses using graphs, tables, statistical tests, and measures of bias.

Chapter 5 closed the study by providing the necessary conclusions, critiques, exposing limitations and strength of the study and reflecting on the extent of achievement of the study objectives.

CHAPTER 2: REGRESSION ANALYSIS

2.1 Overview of Regression Analysis

Regression analysis includes techniques for modeling and analysing several variables, when the focus is on the relationship between a variable of interest (which is the dependent variable) and one or more independent variables (Draper & Smith, 1998). It is beneficial to understand how the typical value of the dependent variable changes when any one of the independent variables is varied, while the other independent variables are held fixed. It also estimates the conditional expectation of the dependent variable given the independent variables; that is, the average value of the dependent variable when the independent variables are fixed. It may, even though rarely, focus on a quantile, or other location parameter of the conditional distribution of the dependent variable given the independent variables.

The methods used to perform regression analysis are collectively known as regression methods. As a result, saying regression analysis may be stated as regression methods without limiting the sense of regression. In all regression analyses cases, the estimation target is a function of the independent variables called the regression function. It is also of interest to characterise the variation of the dependent variable around the regression function, which can be described by a probability function. Several authors (Fox, 1997; Meade & Islam, 1995) point out that regression analysis is widely used for prediction and forecasting, where its use has substantial overlap with the field of machine learning. According to Chatfield (1993, regression analysis is also used to understand which among the independent variables are related to the dependent variable, and to explore the forms of these relationships. In restricted circumstances, regression analysis can be used to infer causal relationships between the independent and dependent variables. However this can lead to illusions or false relationships, so caution is advisable (Scott, 2012).

A large body of techniques for carrying out regression analysis has been developed (Kutner, Nachtsheim &. Neter, 2004). Familiar methods such as linear regression and ordinary least squares regression are parametric, in that the regression function is defined in terms of a finite number of unknown parameters that are estimated from the data. Nonparametric regression refers to techniques that allow the regression function to lie in a specified set of functions,

which may be infinite-dimensional (Hardle, 1990). The performance of regression analysis methods in practice depends on the form of the data generating process, and how it relates to the regression approach being used (Freedman, 2005). The true form of the data-generating process is generally not known. Hence, regression analysis often depends to some extent on making assumptions about this process. These assumptions are sometimes testable if a large amount of data is available. Regression models for prediction are often useful even when the assumptions are moderately violated, although they may not perform optimally (Cook & Weisberg, 1982). However, in many applications, especially with small effects or questions of causality based on observational data, regression methods give misleading results.

2.2 Regression Models

Kutner, Nachtsheim and Neter (2004) explain that regression models involve three fundamental variables. They are the unknown parameters, denoted as $\boldsymbol{\beta}$, which may represent a scalar or a vector; the independent variables, $\mathbf{X}$; and the dependent variable, Y. In various fields of application, different terminologies are used in place of dependent and independent variables. A regression model relates Y to a function of $\mathbf{X}$ and $\boldsymbol{\beta}$.

$$Y \approx f(\mathbf{X}, \boldsymbol{\beta}). \tag{2.1}$$

In statistical practice, the approximation is usually formalised using the relationship $E(Y \mid \mathbf{X}) = f(\mathbf{X}, \boldsymbol{\beta})$. In order to carry out regression analysis, the form of the function f must be specified. Sometimes the form of this function is based on knowledge about the relationship between Y and $\mathbf{X}$ that does not rely on the data. If no such knowledge is available, a flexible or convenient form for f is chosen. For technique's sake, assume that the vector of unknown parameters $\boldsymbol{\beta}$ is of length k. In order to perform a regression analysis the user must provide information about the dependent variable Y:

- If N data points of the form (Y, $\mathbf{X}$) are observed, where $N < k$, most classical approaches to regression analysis cannot be performed: since the system of equations defining the regression model is underdetermined, there is not enough data to recover $\boldsymbol{\beta}$.

- If exactly $N = k$ data points are observed, and the function f is linear, the equations $Y = f(\mathbf{X}, \boldsymbol{\beta})$ can be solved exactly rather than approximately. This reduces to solving a set of N equations with N unknowns (the elements of $\boldsymbol{\beta}$), which has a unique solution as long as the $\mathbf{X}$ are linearly independent. If f is nonlinear, a solution may not exist, or many solutions may exist.
- The most common situation is where $N > k$ data points are observed. In this case, there is enough information in the data to estimate a unique value for $\boldsymbol{\beta}$ that best fits the data in some sense, and the regression model when applied to the data can be viewed as an overdetermined system in $\boldsymbol{\beta}$.

In the last case, the regression analysis provides the tools for:

1. Finding a solution for unknown parameters $\boldsymbol{\beta}$ that will, for example, minimise the distance between the measured and predicted values of the dependent variable Y (also known as method of least squares).
2. Under certain statistical assumptions, the regression analysis uses the surplus of information to provide statistical information about the unknown parameters $\boldsymbol{\beta}$ and predicted values of the dependent variable Y.

2.2.1 Necessary number of independent measurements

Suppose that in a regression model which has k unknown parameters, an experimenter performs n measurements all at exactly the same value of independent variable vector $\mathbf{X}$ (which contains the independent variables X_1, X_2, and X_{3k}) where $k < n$. In such a case, regression analysis fails to give a unique set of estimated values for the three unknown parameters; the experimenter did not provide enough information. The best one can do is to estimate the average value and the standard deviation of the dependent variable Y. Similarly, measuring at two different values of $\mathbf{X}$ would give enough data for a regression with two unknowns, but not for three or more unknowns (Mogull, 2004). If the experimenter had performed measurements at k different values of the independent variable vector $\mathbf{X}$, then regression analysis would provide a unique set of estimates for the k unknown parameters in $\boldsymbol{\beta}$. In the case of general linear regression, the above statement is equivalent to the requirement that matrix $\mathbf{X}^{\mathrm{T}}\mathbf{X}$ is invertible.

2.2.2 Statistical assumptions

When the number of measurements, *N*, is larger than the number of unknown parameters, *k*, and the measurement errors ε_i are normally distributed, then the excess of information contained in (*N* - *k*) measurements is used to make statistical predictions about the unknown parameters (Galton, 1989). This excess of information is referred to as the degrees of freedom of the regression. Classical assumptions for regression analysis include:

- The sample is representative of the population for the inference prediction.
- The error is a random variable with a mean of zero conditional on the explanatory variables.
- The independent variables are measured with no error. (Note: If this is not so, modeling may be done instead using errors-in-variables model techniques).
- The predictors are linearly independent, i.e. it is not possible to express any predictor as a linear combination of the others.
- The errors are uncorrelated, that is, the variance-covariance matrix of the errors is diagonal and each non-zero element is the variance of the error.
- The variance of the error is constant across observations (homoscedasticity). (It should be noted that if this assumption is not valid, then weighted least squares or other methods might instead be used.)

Galton (1989) showed that there are sufficient conditions for the least-squares estimator to possess desirable properties, which imply that the parameter estimates will be unbiased, consistent and efficient in the class of linear unbiased estimators. However, actual data rarely satisfies the assumptions. That is, the method is used even though the assumptions are not true. Variation from the assumptions can sometimes be used as a measure of how far the model is from being useful. Many of these assumptions may be relaxed in more advanced treatments. Reports of statistical analyses usually include analyses of tests on the sample data and methodology for the fit and usefulness of the model. Assumptions include the geometrical support of the variables (Cressie, 1996).

Independent and dependent variables often refer to values measured at point locations. There may be spatial trends and spatial autocorrelation in the variables that violate statistical

assumptions of regression. Geographic weighted regression is one technique to deal with such data (Fotheringham, Brunsdon & Charlton, 2002). Also, variables may include values aggregated by areas. With aggregated data the Modified Areal Unit Problem can cause extreme variation in regression parameters (Fotheringham & Wong, 1991). When analysing data aggregated by political boundaries, postal codes or census areas results may be very different with a different choice of units.

2.3 Linear Regression

2.3.1 Simple linear regression

In linear regression, the model specification is that the dependent variable, y_i is a linear combination of the *parameters* (but need not be linear in the *independent variables*). For example, in simple linear regression (SLR) for modeling *n* data points there is one independent variable: x_i, and two parameters, β_0 and β_1; ε_i is some random variation assumed to be normally distributed with mean zero and variance 1, and the straight line equation is

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \qquad i = 1,\ 2,\ 3,\ \ldots,\ n. \tag{2.2}$$

In multiple linear regression there are several independent variables or functions of independent variables. Adding a term in x_i^2 to the preceding regression gives a parabola with equation

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \varepsilon_i, \qquad i = 1,\ 2,\ 3,\ \ldots,\ n. \tag{2.3}$$

This is still linear regression even though the expression on the right hand side is quadratic in the independent variable x_i, it is linear in the parameters β_0, β_1 and β_2. In both cases, ε_i is an error term and the subscript i indexes a particular observation. Given a random sample from the population, we estimate the population parameters and obtain the sample linear regression model:

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i . \qquad (2.4)$$

The residual, $e_i = y_i - \hat{y}_i$, is the difference between the value of the dependent variable predicted by the model, $\hat{y}_i$ and the true value of the dependent variable y_i. One method of estimation is ordinary least square. This method obtains parameter estimates that minimise the sum of squared residuals, SSE (Ravishankar & Dey, 2002), also sometimes denoted RSS and called the residual sum of squares:

$$SSE = \sum_{i=1}^{N} e_i^2 . \qquad (2.5)$$

Minimisation of this function results in a set of normal equations, a set of simultaneous linear equations in the parameters, which are solved to yield the parameter estimators, β_0 and β_1. The two parameters are the intercept and slope of the linear equation obtained after estimation. A graphical example takes the form in the following graph.

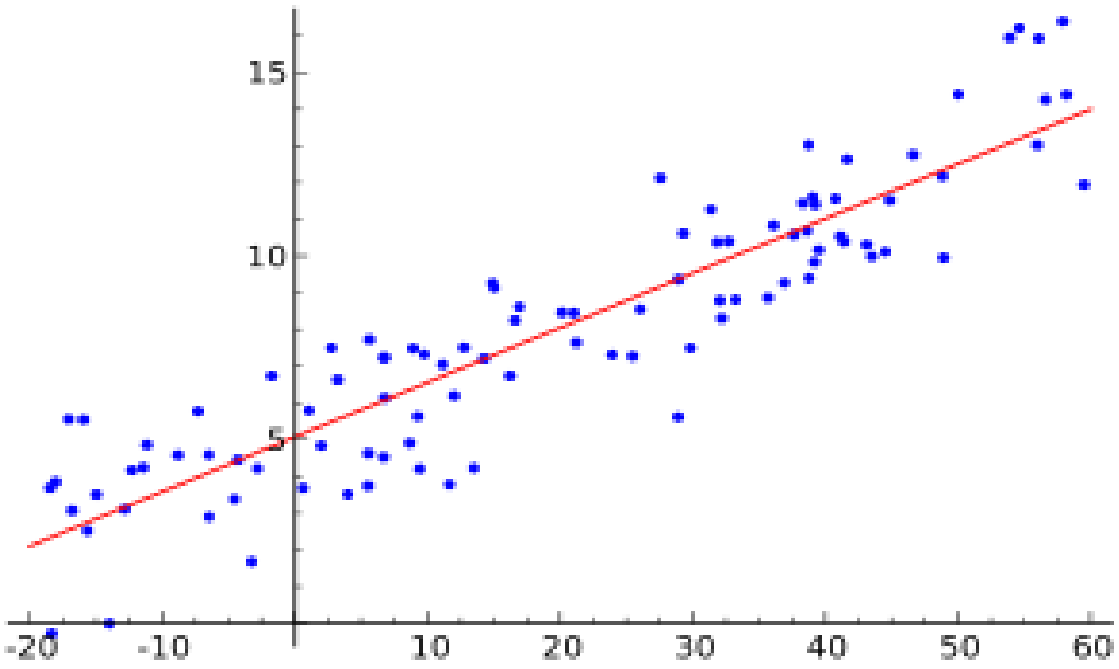


Illustration of linear regression on a data set.

In the case of simple regression, the formulas for the least squares estimates are

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{N}(x_i - \bar{x})^2} \tag{2.6}$$

and

$$\beta_0 = \bar{y} - \beta_1 \bar{x}. \tag{2.7}$$

where $\bar{x}$ is the mean (average) of the x values and $\bar{y}$ is the mean of the y values. Under the assumption that the population error term has a constant variance, the estimate of that variance is given by:

$$\hat{\sigma}_\varepsilon^2 = \frac{SSE}{N-2}. \tag{2.8}$$

This is called the mean square error (MSE) of the regression. The standard errors of the parameter estimates are given by

$$\sigma_{\beta_0} = \hat{\sigma}_\varepsilon \sqrt{\frac{1}{N} + \frac{\bar{x}}{\sum_{i=1}^{N}(x_i - \bar{x})^2}}. \tag{2.9}$$

$$\sigma_{\beta_1} = \hat{\sigma}_\varepsilon \sqrt{\frac{1}{\sum_{i=1}^{N}(x_i - \bar{x})^2}}. \tag{2.10}$$

Under the further assumption that the population error term is normally distributed, the researcher can use these estimated standard errors to create confidence intervals and conduct hypothesis tests about the population parameters.

2.3.2 General linear model

In the more general multiple regression model, there are *p* independent variables:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \ . \ . \ . + \beta_p x_{ip} + \varepsilon_i, \qquad i = 1,\ 2,\ 3,\ . \ . \ .,\ n. \quad (2.11)$$

where x_{ij} is the i^{th} observation on the j^{th} independent variable, and where β_0 is the regression intercept. The least squares parameter estimates are obtained from *p* normal equations. The residual can be written as

$$e_i = y_i - \hat{\beta}_1 x_{i1} - \ . \ . \ . - \hat{\beta}_p x_{ip}. \quad (2.12)$$

The normal equations are

$$\sum_{i=1}^{n} \sum_{k=1}^{p} X_{ij} X_{ik} \hat{\beta}_k = \sum_{i=1}^{n} X_{ij} y_i, \ j = 1, \ldots, p. \quad (2.13)$$

In matrix notation, the normal equations are written as

$$(\mathbf{X}^\top \mathbf{X}) \hat{\boldsymbol{\beta}} = \mathbf{X}^\top \mathbf{Y}, \quad (2.14)$$

where the *ij* element of *X* is x_{ij}, the *i* element of the column vector *Y* is y_i, and the *j* element of $\hat{\beta}$ is $\hat{\beta}_j$. Thus *X* is *n*×*p*, *Y* is *n*×1, and $\hat{\beta}$ is *p*×1. The solution is

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{Y}. \quad (2.15)$$

2.3.3 Regression diagnostics

Models need to be verified for quality use. As a result, once a tentative regression model has been constructed, it should be tested, adapted and then validated (Aldrich, 2005). It is always important to confirm the goodness-of-fit of the final model and the statistical significance of

the estimated parameters. Commonly used checks of goodness of fit include the R-squared, analyses of the pattern of residuals and hypothesis testing. Statistical significance can be checked by an F-test of the overall fit, followed by t-tests of individual parameters. Interpretations of these diagnostic tests rest heavily on the model assumptions.

Although examination of the residuals can be used to validate or invalidate a model, the results of a t-test or F-test are sometimes more difficult to interpret if the model's assumptions are violated (Fisher, 1922). For example, if the error term does not have a normal distribution, in small samples the estimated parameters will not follow normal distributions and this usually complicates inference. With relatively large samples, however, a central limit theorem can be invoked such that hypothesis testing may proceed using asymptotic approximations.

2.3.4 Regression with "limited dependent" variables

The phrase "limited dependent" is commonly used in econometric statistics for categorical and constrained variables (Ramcharan. 2006). The response variable may be non-continuous ("limited" to lie on some subset of the real line). For binary (zero or one) variables, if analysis proceeds with least-squares linear regression, the model is called the linear probability model. Nonlinear models for binary dependent variables include the probit and logit model. The multivariate probit model is a standard method of estimating a joint relationship between several binary dependent variables and some independent variables (Chiang, 2003). For categorical variables with more than two values there is the multinomial logit. For ordinal variables with more than two values, there are the ordered logit and ordered probit models. Censored regression models may be used when the dependent variable is only sometimes observed, and Heckman correction type models may be used when the sample is not randomly selected from the population of interest. An alternative to such procedures is linear regression based on polychoric correlation (or polyserial correlations) between the categorical variables. Such procedures differ in the assumptions made about the distribution of the variables in the population. If the variable is positive with low values and represents the repetition of the occurrence of an event, then count models like the Poisson regression or the negative binomial model may be used instead.

2.3.5 Interpolation and extrapolation

Regression models are used to predict a value of the *Y* variable given known values of the *X* variables (Chatfield, 1993). Prediction *within* the range of values in the dataset used for model-fitting is known informally as interpolation. Prediction *outside* this range of the data is known as extrapolation. Performing extrapolation relies strongly on the regression assumptions. The further the extrapolation goes outside the data, the more room there is for the model to fail due to differences between the assumptions and the sample data or the true values. It is generally advised that when performing extrapolation, one should accompany the estimated value of the dependent variable with a prediction interval that represents the uncertainty (Strutz, 2010). Such intervals tend to expand rapidly as the values of the independent variable(s) moved outside the range covered by the observed data. For such reasons and others, some tend to say that it might be unwise to undertake extrapolation (Chiang, 2003). However, this does not cover the full set of modeling errors that may be made: in particular, the assumption of a particular form for the relation between *Y* and *X*. A properly conducted regression analysis will include an assessment of how well the assumed form is matched by the observed data, but it can only do so within the range of values of the independent variables actually available (Yang-Jing, 2009). Thus, any extrapolation is particularly reliant on the assumptions being made about the structural form of the regression relationship.

A scientific perspective is that a linear-in-variables and linear-in-parameters relationship should not be chosen simply for computational convenience, but that all available knowledge should be deployed in constructing a regression model (Tofallis, 2009). If this knowledge includes the fact that the dependent variable cannot go outside a certain range of values, this can be used in selecting the model even in a case where the observed dataset has no values particularly near such bounds. The implications of this step of choosing an appropriate functional form for the regression can be great when extrapolation is considered. At a minimum, it can ensure that any extrapolation arising from a fitted model is "realistic" (or in accord with what is known).

2.4 Nonlinear Regression

Most of the linear regression models used in practice are estimations of nonlinear models mainly because linear regression has advanced in usage and relative simplicity. In reality, many situations are nonlinear and the relationships should have been modeled with nonlinear regression. When the model function is not linear in the parameters, the sum of squares must be minimised by an iterative procedure.

2.4.1 Power and sample size calculations

There are no generally agreed methods for relating the number of observations versus the number of independent variables in the model. One rule of thumb suggested by Good and Hardin (2009) is $N = m^n$, where N is the sample size, *n* is the number of independent variables and m is the number of observations needed to reach the desired precision if the model had only one independent variable. For example, a researcher is building a linear regression model using a dataset that contains 1000 patients (N). If he decides that five observations are needed to precisely define a straight line (m), then the maximum number of independent variables his model can support is 4, because

$$\frac{\log 1000}{\log 5} = 4.29$$

.

2.4.2 Other methods of estimation

The parameters of a regression model are usually estimated using the method of least squares, mainly because of long history of the method and its effectiveness. It is also not a difficult method to execute. However, there are other methods. Several authors (Good & Hardin, 2009; Lindley, 1987; Tofallis, 2009; among others) inform that other methods used in the estimation of parameters include:

- Bayesian methods, e.g. Bayesian linear regression.
- Percentage regression, for situations where reducing *percentage* errors is deemed more appropriate.
- Least absolute deviation, which is more robust in the presence of outliers, leading to quantile regression

- Nonparametric regression, requires a large number of observations and is computationally intensive
- Distance metric learning, which is learned by the search of a meaningful distance metric in a given input space.

2.5 Multicollinearity

2.5.1 Collinearity

Collinearity is a linear relationship between *two* explanatory variables. Two variables are perfectly collinear if there is an exact linear relationship between the two. For example, X_1 and X_2 are perfectly collinear if there exist parameters λ_0 and λ_1 such that, for all observations *i*, we have

$$X_{2i} = \lambda_0 + \lambda_1 X_{1i}. \tag{2.16}$$

2.5.2 Multicollinearity

Multicollinearity refers to a situation in which two or more explanatory variables in a multiple regression model are highly linearly related (Farrar & Glauber, 1967). We have perfect multicollinearity if, for example as in the equation above, the correlation between two independent variables is equal to 1 or -1. In practice, we rarely face perfect multicollinearity in a data set. More commonly, the issue of multicollinearity arises when there is a strong linear relationship among two or more independent variables. Mathematically, a set of variables is perfectly multicollinear if there exist one or more exact linear relationships among some of the variables. For example, we may have

$$\lambda_0 + \lambda_1 X_{1i} + \lambda_2 X_{2i} + \cdots + \lambda_k X_{ki} = 0 \tag{2.17}$$

holding for all observations *i*, where λ_j are constants and λ_{ji} is the i^{th} observation on the j^{th} explanatory variable. We can explore one issue caused by multicollinearity by examining the

process of attempting to obtain estimates for the parameters of the multiple regression equation

$$Y_i = \beta_0 + \beta_1 X_{1i} + \cdots + \beta_k X_{ki} + \varepsilon_i.$$

The ordinary least squares estimates involve inverting the matrix

$$X^T X$$

where

$$X = \begin{bmatrix} 1 & X_{11} & \cdots & X_{k1} \\ \vdots & \vdots & & \vdots \\ 1 & X_{1N} & \cdots & X_{kN} \end{bmatrix}.$$

An exact linear relationship (or perfect multicollinearity) exists among the independent variables if the rank of X (and therefore of X^TX) is less than k+1, and the matrix X^TX will not be invertible. Perfect multicollinearity is unlikely in actual application. An analyst is more likely to face a high degree of multicollinearity. For example, suppose that instead of the above equation holding, we have that equation in modified form with an error term υ_i:

$$\lambda_0 + \lambda_1 X_{1i} + \lambda_2 X_{2i} + \cdots + \lambda_k X_{ki} + v_i = 0.$$

In this case, there is no exact linear relationship among the variables, but the X_j variables are nearly perfectly multicollinear if the variance of v_i is small for some set of values for the λ's. In this case, the matrix X^TX has an inverse, but is ill-conditioned so that a given computer algorithm may or may not be able to compute an approximate inverse, and if it does so the resulting computed inverse may be highly sensitive to slight variations in the data (due to magnified effects of rounding error) and so may be very inaccurate.

2.5.3 Detecting multicollinearity

There are indicators when multicollinearity may be present in a model. A common indicator is when large changes are observed in the estimated regression coefficients when a predictor variable is added or deleted. There may also be insignificant regression coefficients for the affected variables in the multiple regression, but a rejection of the joint hypothesis that those coefficients are all zero (using an F-test). Statistical tests can also be conducted to provide information about the existence of multicollinearity. O'Brien (2007) explains the use of the tolerance and variance-inflation factor (VIF) in detecting multicollinearity below:

Let be the coefficient of determination of regression of explanator *j* on all the other explanators. The tolerance measure is given by the formula

$$tolerance = 1 - R_j^2 \tag{2.18}$$

On the other hand The VIF is given by the formula

$$VIF = \frac{1}{tolerance} \tag{2.19}$$

The VIF can therefore be written as

$$VIF = \frac{1}{1 - R_j^2} \tag{2.19a}$$

Interpretation using tolerance and VIF

According to O'Brien (2007), a tolerance of less than 0.20 or 0.10 and/or a VIF of 5 or 10 and above indicates a multicollinearity problem.

2.5.4 Consequences of multicollinearity

A high degree of multicollinearity may lead to the matrix X^TX to be invertible, a computer algorithm may be unsuccessful in obtaining an approximate inverse, and if it does obtain one

it may be numerically inaccurate. However, an accurate X^TX matrix may still be obtained, but consequences of multicollinearity still arise. Multicollinearity may make the estimate of one variable's impact on the dependent variable while controlling for the others tends to be less precise than if predictors were uncorrelated with one another. There are also imprecise estimates of the effect of the independent changes in the individual variables. In some sense, the collinear variables contain the same information about the dependent variable. If nominally "different" measures actually quantify the same phenomenon then they are redundant. Alternatively, if the variables are accorded different names and perhaps employ different numeric measurement scales but are highly correlated with each other, then they suffer from redundancy.

A major hazard of redundancy is overfitting in regression analysis models. The best regression models are those in which the predictor variables each correlate highly with the dependent (outcome) variable but correlate at most only minimally with each other. Such a model is often called "low noise" and will be statistically robust (that is, it will predict reliably across numerous samples of variable sets drawn from the same statistical population).

Also, in multicollinearity, the standard errors of the affected coefficients tend to be large. In that case, the test of the hypothesis that the coefficient is equal to zero leads to a failure to reject the null hypothesis. However, if a simple linear regression of the explained variable on this explanatory variable is estimated, the coefficient will be found to be significant; specifically, the analyst will reject the hypothesis that the coefficient is zero. In the presence of multicollinearity, an analyst might falsely conclude that there is no linear relationship between an independent and a dependent variable.

Recap

Multicollinearity is a statistical phenomenon in which two or more predictor variables in a multiple regression model are highly correlated. In this situation the coefficient estimates may change erratically in response to small changes in the model or the data. It does not reduce the predictive power or reliability of the model as a whole, at least within the sample data themselves; it only affects calculations regarding individual predictors. That is, a multiple regression model with correlated predictors can indicate how well the entire bundle

of predictors predicts the outcome variable, but it may not give valid results about any individual predictor, or about which predictors are redundant with respect to others. A high degree of multicollinearity can also cause computer software packages to be unable to perform the matrix inversion that is required for computing the regression coefficients, or it may make the results of that inversion inaccurate (Van den Poel & Larivière, 2004).

2.6 Software

Modern statistical practice cannot be performed without a statistical package (Yang-Jing, 2009). This is true with regression analysis as well. All major statistical software packages perform least squares regression analysis and inference. Simple linear regression and multiple regression using least squares can be done in some spreadsheet applications and on some calculators. While many statistical software packages can perform various types of nonparametric and robust regression, these methods are less standardised; different software packages implement different methods and a method with a given name may be implemented differently in different packages. Specialised regression software has been developed for use in fields such as survey analysis and neuroimaging.

2.7 Implications for this Study

This study is about matric pass rates and the number of interventions embarked on in the high schools in the district of Letlhabile in the Brits area of the North West Province of South Africa. The dependent variable therefore, is the matric pass rates. The independent variable is the number of interventions or support systems administered on the matric learners. Since the number of interventions to which the learners were exposed is a univariate random variable, the envisaged relationship is univariate regression. Focus will be more on simple linear regression even though some analyses will be embarked on to explore (minimally) possibilities of other forms of relationships between the matric pass rates in the geographical area of interest and the number of interventions.

Also, regarding the relationships explored in the study, nonlinear relationships may be more suitable. This study opens up further exploration or future research regarding the suitability of nonlinear models. The other point is that the interventions may require a different

representation in order to produce a more relevant and effective model. The current study seeks effective answers regarding the most effective statistical modeling of matric pass rates and the interventions taking place in order to apply it in the schools in question to improve the matric pass rate and the overall throughput.

2.8 Conclusion

The chapter discussed regression methods and its relevant aspects for this study. Various formulae and mathematical formations were introduced to back up the discussions. Many of these formations feature in the formation of regression models in conjunction with forecast functions for time series analysis and forecast development. The next chapter presents time series analysis.

CHAPTER 3: TIME SERIES ANALYSIS

3.1 Overview of Time Series

The field of time series is a specialised statistical technique useful in every field that can collect data in a pattern required by time series methodologies. Areas that benefit regularly from time series include business, communication, econometrics, economics, education, finance, health, housing, information mathematics, technology, mining, management, operations research, population studies, signal processing, transportation, and weather bureau. Also, almost all modern fields that have planning in them use time series methods in one way or the other. Hamilton (1994) defines a time series as a sequence of data points, measured typically at successive time instants spaced at uniform time intervals.

In South Africa, examples of time series are the daily closing value of the in the Producer Price Index (PPI), the annual flow volume of the Orange River in the Free State of South Africa, and the water levels of the Vaal Dam in the Vanderbijlpark area of the Gauteng Province. Time series analysis comprises methods for analysing time series data in order to extract meaningful statistics and other characteristics of the data. Time series forecasting is the use of a model to predict future values based on previously observed values. Time series are very frequently plotted via line charts. One special feature of time series is that time series data have a natural temporal ordering. This makes time series analysis distinct from other common data analysis problems, in which there is no natural ordering of the observations (e.g. explaining people's wages by reference to their respective education levels, where the individuals' data could be entered in any order). This also makes time series a specialised subfield within the field of statistics and requiring separate training. It is also an applied subfield when used in forecasting. Use of time series can also help to backcast, which is to use existing time series to figure out the way the past was in relation to the phenomenon of interest to a study.

Time series analysis is also distinct from spatial data analysis where the observations typically relate to geographical locations (e.g. accounting for house prices by the location as well as the intrinsic characteristics of the houses). Since time series tends to occur as random variables, methods useful in stochastic modeling are also applicable in time series analysis. A

stochastic model for a time series would generally reflect the fact that observations close together in time are more closely related than observations further apart. In addition, time series models often make use of the natural one-way ordering of time so that values for a given period will be expressed as deriving in some way from past values, rather than from future values. Methods for time series analyses may be divided into two classes: frequency-domain methods and time-domain methods. The former include spectral analysis and recently wavelet analysis (including autocorrelation and cross-correlation analysis).

3.2 Time Series Analysis

Analysis in this study means in-depth investigation of the data available. In this study the matric pass rates in the high schools in the Letlhabile area are explored from a few years to the latest available ones to date. There are several types of data analysis available for time series which are appropriate for different purposes. Use of time series in this study helps to understand the changes in pass rates pattern in the past few years, mainly the trend. An increasing trend is an indication of improvement while a decreasing one shows that there is a decline. Constant trend informs that there is neither improvement nor decline in pass rates.

3.2.1 General exploration

The clearest way to examine a regular time series is with a line chart, also known as the time plot. This chart plots time in the horizontal axis and the corresponding time values on the vertical axis. The line chart is easy to plot using pencil and ruler, but for better results modern practice dictates that a spreadsheet program be used. The advantage of using spreadsheets is that once the time period is chosen, the calculation of change over the periods can be easily calculated and with great accuracy. The nature of the trend can be easily revealed as well as the type of seasonality. Other time series analysis techniques mentioned by various time series practitioners (Bloomfield, 1976; Nikolić, Muresan, Feng & Singer, 2012) include:

- Autocorrelation analysis to examine serial dependence
- Spectral analysis is used to examine cyclic behaviour which need not be related to seasonality. For example, sun spot activity varies over 11 year cycles. Other common examples include celestial phenomena, weather patterns, neural activity, commodity prices, and economic activity.

3.2.2 Description

A time series is easily described and comprehensible when decomposed or separated into its basic components. The decomposition of time series is a statistical method that deconstructs a time series into notional components (Shumway, 1988). One of the main objectives for decomposition is to estimate seasonal effects that can be used to create and present seasonally adjusted values. A seasonally adjusted value removes the seasonal effect from a value so that trends can be seen more clearly. For instance, in many regions of the South Africa unemployment tends to decrease in the summer due to increased employment in agricultural areas. Thus, a drop in the unemployment rate in October compared to September does not necessarily indicate that there is a trend toward lower unemployment in the country. To see whether there is a real trend, we should adjust for the fact that unemployment is always lower in October than in September.

Choosing between additive and multiplicative decompositions

- The additive model is useful when the seasonal variation is relatively constant over time.
- The multiplicative model is useful when the seasonal variation increases over time.

For practical purposes, the components of time series, according to source, may be:

1. **average:** the mean of the observations over time
2. **trend:** a gradual increase or decrease in the average over time
3. **seasonal influence:** predictable short-term cycling behaviour due to time of day, week, month, season, year, etc.
4. **cyclical movement:** unpredictable long-term cycling behaviour due to business cycle or product/service life cycle
5. **random error:** remaining variation that cannot be explained by the other four components

It is important to know the extent of accuracy of your forecasts. For notation purposes the actual values are denoted by A and the forecasts by F. Inspecting accuracy of forecasts is to determine the way the forecasts F is relative to the actual value A, as well as the meaning of $A - F$?

3.2.3 Time series forecasting

Forecasting is the complete formation of statistical models for stochastic simulation purposes, so as to generate alternative versions of the time series, representing what might happen over non-specific time-periods in the future (Box & Jenkins, 1976). By forecasting we shall mean a method for translating past experience into estimates of the future. Simple or fully formed statistical models are used to describe the likely outcome of the time series in the immediate future, given knowledge of the most recent outcomes (Shasha, 2004). This means that once a time series model is developed for forecasting purpose, it should be able to provide short term forecasts that give indication of the possible performance on medium (and maybe long term) forecasting.

Time series forecasting methods are based on analysis of historical data (time series: a set of observations measured at successive times or over successive periods) (Bloomfield, 1976). They make the assumption that past patterns in data can be used to forecast future data points.

1. Moving averages (simple moving average, weighted moving average): forecast is based on arithmetic average of a given number of past data points
2. Exponential smoothing (single exponential smoothing, double exponential smoothing): a type of weighted moving average that allows inclusion of trends, etc.
3. Mathematical models (trend lines, log-linear models, Fourier series, etc.): linear or non-linear models fitted to time-series data, usually by regression methods
4. Box-Jenkins methods: autocorrelation methods used to identify underlying time series and to fit the "best" model

It is vital to be able to measure the optimality of a time-series forecast since we cannot expect a time-series forecast to be perfect. There will always be prediction errors. Suppose that t time series observations $A_1, A_2, \ldots A_t$ were realised and the corresponding forecasts $F_1, F_2, \ldots F_t$ generated from a forecast model were . Defines the error terms:

$$e_i = A_i - F_i \quad \text{for } i = 1, 2, \ldots t \tag{3.1}$$

This is the difference between the actual time-series A_i and the forecast F_i. These error terms are useful in analysing and summarising the accuracy of the forecasts. Use of t instead of n is to be consistent with the notion of current time t so that the next time becomes $t+1$. Apart from the next period $t+1$, other future periods are $t+2$, $t+3$, and so on.

3.3 Measures of accuracy and bias

3.3.1 Cumulative error

The cumulative forecast error (CFE) is the sum of all prediction errors:

$$CFE = \sum_{i=1}^{t} e_i \tag{3.2}$$

3.3.2 Mean error

The mean error is the arithmetic average of all prediction errors:

$$ME = \frac{1}{n} \sum_{i=1}^{t} e_i \tag{3.3}$$

3.3.3 Mean squared error

The mean squared error (MSE) is the arithmetic mean of the sum of the squares of the prediction errors; this error measure is popular and corrects the 'cancelling out' effects of the previous two error measures:

$$MSE = \frac{1}{n} \sum_{i=1}^{t} e_i^2 \tag{3.4}$$

3.3.4 Root mean squared error

Another measure that is useful in almost precisely the same way as the MSE is the root mean squared error (RMSE). The RMSE is the square root of the MSE. Mathematically is written as follows:

$$RMSE = \sqrt{MSE} \tag{3.5}$$

3.3.5 Standard Deviation

The standard deviation is as the name implies the standard deviation of the prediction errors.

$$s_e = \sqrt{\frac{1}{n-1}\sum_{i=1}^{t}\left(e_i - \bar{e}\right)^2} \tag{3.6}$$

3.3.6 Mean absolute deviation

The mean absolute deviation (MAD) is another popular error measure that corrects the 'cancelling out' effects by averaging the absolute value of the errors:

$$MAD = \frac{1}{n}\sum_{i=1}^{t}\left|e_i\right| \tag{3.7}$$

3.3.7 Mean absolute percentage error

The mean absolute percentage error (MAPE) is a very popular measure that corrects the 'cancelling out' effects and also keeps into account the different scales at which this measure can be computed and thus can be used to compare different predictions:

$$MAPE = \frac{100}{n}\sum_{i=1}^{t}\frac{\left|e_i\right|}{A_i} \tag{3.8}$$

3.4 Models

In real-life situations models for time series data can have many forms and represent different stochastic processes. When modeling variations in the level of a process, three broad classes of practical importance are the autoregressive (AR) models, the *integrated* (I) models, and the moving average (MA) models. Gershenfeld (1999) point out that these three classes depend linearly on previous data points. Combinations of these ideas produce autoregressive moving average (ARMA) and autoregressive integrated moving average (ARIMA) models. The autoregressive fractionally integrated moving average (ARFIMA) model generalises the former three. Extensions of these classes deal with vector-valued data are available under the heading of multivariate time-series models and sometimes the preceding acronyms are extended by including an initial "V" for "vector". An additional set of extensions of these models is available for use where the observed time-series is driven by some "forcing" time-series (which may not have a causal effect on the observed series): the distinction from the multivariate case is that the forcing series may be deterministic or under the experimenter's control. For these models, the acronyms are extended with a final "X" for "exogenous".

Non-linear dependence of the level of a series on previous data points is of interest, partly because of the possibility of producing a chaotic time series. However, more importantly, empirical investigations can indicate the advantage of using predictions derived from non-linear models, over those from linear models, as for example in nonlinear autoregressive exogenous models (Gershenfeld, 2000). Among other types of non-linear time series models, there are models to represent the changes of variance along time (heteroskedasticity). These models represent autoregressive conditional heteroskedasticity (ARCH) and the collection comprises a wide variety of representation (GARCH, TARCH, EGARCH, FIGARCH, CGARCH, etc.). Here changes in variability are related to, or predicted by, recent past values of the observed series. This is in contrast to other possible representations of locally varying variability, where the variability might be modeled as being driven by a separate time-varying process, as in a doubly stochastic model. In recent work on model-free analyses, wavelet transform based methods (for example locally stationary wavelets and wavelet decomposed neural networks) have gained favour among forecasting practitioners (Durbin & Koopman

2001; Nikolić *et al.*, 2012). Multiscale (often referred to as multiresolution) techniques decompose a given time series, attempting to illustrate time dependence at multiple scales.

3.4.1 Notation

A number of different notations are in use for time-series analysis. A common notation specifying a time series X that is indexed by the natural numbers is written

$$X = \{X_1, X_2, ...\}.$$

Another common notation is

$$Y = \{Y_t: t \in T\},$$

where T is the index set.

3.4.2 Conditions

There are two sets of conditions under which much of the times series theory is constructed, namely; stationary process and ergodic process (Allen, 2010; Gardiner, 2004). A stationary process is a stochastic process which whose joint probability distribution does not change when shifted in time or space. Consequently, parameters such as the mean and variance, if they exist, also do not change over time or position. On the other hand, a stochastic process is said to be an ergodic process if its statistical properties (such as its mean and variance) can be deduced from a single, sufficiently long sample (realisation) of the process. Stationarity is important in time series analysis due to its easing of analysis of nonstationary processes. The practice of analysing nonstationary processes is based on the fundamentals developed for stationary processes. As a result the ideas of stationarity must be expanded to consider two important ideas: strict stationarity and second-order stationarity. Both models and applications can be developed under each of these conditions, although the models in the latter case might be considered as only partly specified. In addition, time-series analysis can be applied where the series are seasonally stationary or non-stationary. According to Boashash (2003), situations where the amplitudes of frequency components change with time

can be dealt with in time-frequency analysis which makes use of a time-frequency representation of a time-series or signal.

3.4.3 ARIMA models

The ARIMA models are made of a combination of the autoregressive (AR) and moving average (MA) models, and this model is also integrated (I). The general orders of these models are p for AR, d for I and q for MA. Their notations are individually written as AR(p), I(d) and MA(q) and then the general ARIMA model is written with as ARIMA(p, d, q). The general representation of an AR(*p*), is

$$Y_t = \alpha_0 + \alpha_1 Y_{t-1} + \alpha_2 Y_{t-2} + \cdots + \alpha_p Y_{t-p} + \varepsilon_t \qquad (3.9)$$

where the term ε_t is the source of randomness called the white noise process with the following characteristics:

- $E[\varepsilon_t] = 0\,,$
- $E[\varepsilon_t^2] = \sigma^2\,,$
- $E[\varepsilon_t \varepsilon_s] = 0 \quad \text{for all } t \neq s\,.$

When the above assumptions are satisfied, the time series process is specified up to second-order moments and, subject to conditions on the coefficients, may be second-order stationary. A noise that has a normal distribution is called normal or Gaussian white noise (Brillinger, 1975; Box & Jenkins, 1976). In this case, the AR process may be strictly stationary, again subject to conditions on the coefficients.

The moving-average (MA) model is a common approach for modeling univariate time series models. The notation MA(*q*) refers to the moving average model of order *q*, and its general representation is of the form:

$$X_t = \mu + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \cdots + \theta_q \varepsilon_{t-q}$$

where μ is the mean of the series, the θ_1, ..., θ_q are the parameters of the model and the ε_t, ε_{t-1},... are white noise error terms. The value of q is called the order of the MA model. That is, a moving-average model is conceptually a linear regression of the current value of the series against previous (unobserved) white noise error terms or random shocks. The random shocks at each point are assumed to come from the same distribution, typically a normal distribution, with location at zero and constant scale. The distinction in this model is that these random shocks are propagated to future values of the time series.

A time series is said to be integrated when firstly it is nonstationary, but after a specific number of differencing, it becomes stationary (Granger, 1981). There are some nonstationary time series that are not integrated because they never become stationary even if they mat be differenced infinitely. Due to the depth of the concept especially when it is treated in a multivariate setting, it is presented separately.

3.5 Integration

3.5.1 Order of integration

The order of integration, denoted $I(d)$, is a summary statistic for a stochastic (including a time series) process that reports the minimum number of differences required to obtain a stationary series (Engle & Granger, 1987). This means that an I(d) process requires to be differenced d times to reach stationarity. One important integration value is $d = 0$, which may be confusing. It is obvious that a stationary process requires no (or $d = 0$) number of times to be differenced to reach stationarity.

Integration of order zero

A time series is integrated of order 0 if it admits a moving average representation with

$$\sum_{k=1}^{\infty} \left| \gamma_k^2 \right| < \infty, \tag{3.10}$$

which means that the autocovariance is decaying to 0 sufficiently fast. This is a necessary, but not sufficient condition for a stationary process (Granger & Newbold, 1974). Therefore, all stationary processes are I(0), but not all I(0) processes are stationary.

Integration of order *d*

A time series is integrated of order *d* if

$$(1-L)^d X_t$$

is integrated of order 0, where L is the lag operator and $(1-L)$ is the first difference, that is:

$$(1-L)X_t = X_t - X_{t-1} = \Delta X_t \ .$$

In other words, a process is integrated to order *d* if taking repeated differences d times yields a stationary process.

Constructing an integrated series

An $I(d)$ process can be constructed by summing an $I(d-1)$ process:

- Suppose X_t is $I(d\text{-}1)$
- Now construct a series $Z_t = \sum_{k=0}^{t} X_k$
- Show that Z is $I(d)$ by observing its first-differences are $I(d\text{-}1)$:

$$\Delta Z_t = \sum_{k=0}^{t} X_k - \sum_{k=0}^{t-1} X_k = X_k \sim I(d-1). \qquad (3.11)$$

3.5.2 Cointegration

The concept of integration extends to more than one variable and when the individually integrated time series are stationary when viewed relative to each other. The concept is known as cointegration. Therefore, cointegration is a statistical property of time series variables. In formal terms, two or more time series are cointegrated if they share a common

stochastic drift (Gregory & Hansen, 1996). The mathematical property of cointegrated time series is thought-provoking. In mathematical terms, if two or more series are individually integrated (in the time series sense) but some linear combination of them has a lower order of integration, then the series are said to be cointegrated. A common example is where the individual series are first-order integrated (I(1**)**) but some (cointegrating) vector of coefficients exists to form a stationary linear combination of them.

Before the 1980s many economists used linear regressions on (de-trended) non-stationary time series data, which Nobel laureate Clive Granger and others showed to be a dangerous approach that could produce spurious correlation (Granger, 1981). His 1987 paper with Nobel laureate Robert Engle formalised the cointegrating vector approach, and coined the term (Engle & Grnger, 1987). The possible presence of cointegration must be taken into account when choosing a technique to test hypotheses concerning the relationship between two variables having unit roots (i.e. integrated of at least order one) (Granger & Newbold, 1974). The traditional method for testing hypotheses concerning the relationship between non-stationary variables was to run ordinary least squares (OLS) regressions on data which had initially been differenced. This method is correct in large samples, but cointegration provides more powerful tools when the data sets are of limited length, which is the case with most economic time-series.

3.6 Implications for the Study

The matric passes in the study are explored over a period of five years from 2007 to 2011, which then become time series data. Due to several schools being studied, this becomes multivariate time series. Forecasting will be explored to anticipate what the pass rates could be if the conditions of learner support are maintained in the future years. Regarding forecast accuracy, the errors should be minimised for more reliable forecasts. As a result, an accurate forecast model should provide smaller measures of accuracy to reflect low errors. In addition, the matric pass rates will be tested for stationarity and nonstationarity, and the implications for their conditions will be explored in education exercises. The meaning of integration and cointegration issues for the pass rates is not comprehensible. Hence the theoretical study did not extend to presenting the various tests of cointegration.

3.7 Conclusion

The chapter presented time series analysis and measures of bias and precision. Curvilinear polynomials (linear, quadratic and higher power polynomials), exponential, logarithmic and power functions were discussed for use in the tentative model for predicting pass rates from the numbers of interventions. The tests of hypotheses and various measures for determining the most suitable models are also discussed. The next chapter presents the findings.

CHAPTER 4: FINDINGS

4.1 Introduction

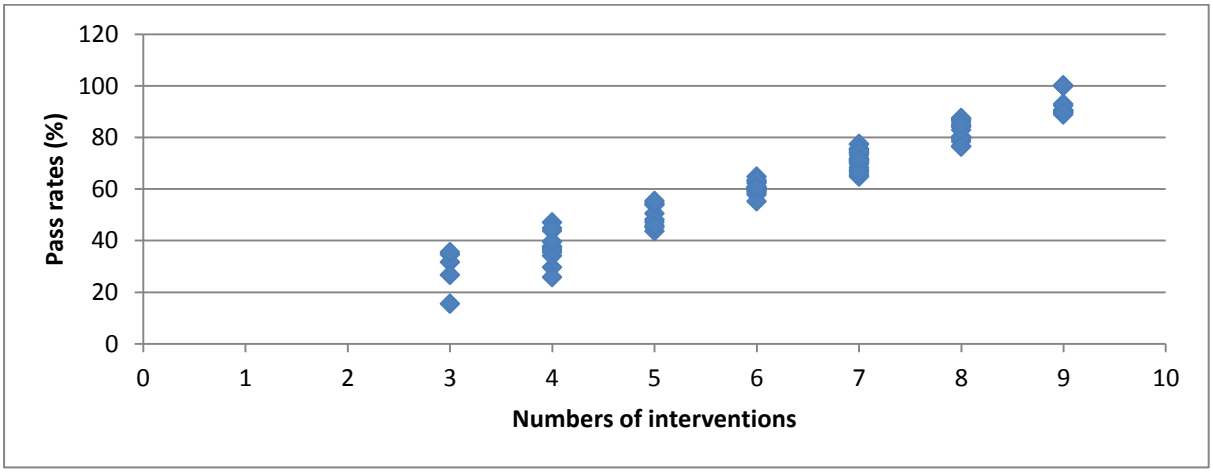
Multiple times series data on the 19 high schools of the Letlhabile area for the five years of this study (2007 to 2011) appear in Annexure A. They are given as pass rates (in percentages) for the schools and the numbers of interventions that took place in each school during the five years in question. Time series data were required to examine the trend patterns over these years for the different schools at individual and comparative level. Various mathematical models were established to determine their suitability in forecasting the matric pass rates in this area, based on the number of interventions that took place in a school. However, these models were also still going to be compared using regression methods and the appropriate statistical tests and measures.

These same data were also presented for the regression relationships in which a bivariate setting was presented. This mode led to $n = 95$ pass rates (y) that appear together with the corresponding numbers of interventions (x) (see Annexure B). Various forms that are evaluated against one another are the exponential relationship, curvilinear forms (linear and polynomial regressions), logarithmic, and the power relationship. The analyses are also given in the various graphical displays. Measures of bias and precision are also calculated for the various measures in order to fortify the comparisons made among the models developed. For this purpose, a table displaying the various measures is used. A discussion explaining the methods follows each table. One version of the comparisons of the models is based on these measures. The coefficients of determination values are estimated by the values of the respective R-squared. These coefficients are needed to provide the measures of the strengths of the relationships. In order to ensure that the final choice of the selected relationship between pass rates and the numbers of intervention, appropriate tests are also conducted. The statistical package used in the data analysis is STATA.

The various data organisations and statistical analyses follow in the presentations of the next sections.

4.2 Graphical Display of Passes against Numbers of Interventions

Figure 4.1: Scatter plot of pass rates vs. numbers of interventions



The data show a pattern where relationship of a linear or slowly increasing (or lax) curve can be said to exist between the pass rates and the numbers of interventions.

4.3 Mathematical Equations

4.3.1 Exponential equation

The exponential equation takes the form

$$Y = ae^{bx} + \varepsilon \tag{4.1}$$

where the parameters a and b are to be estimated. Their formulae are:

$$b = \frac{n\sum_{i=1}^{n} x_i \ln y_i - \sum x_i \sum \ln y_i}{n\sum x_i^2 - \left(\sum x_i^2\right)^2} \tag{4.2}$$

and

$$a = \overline{\ln y} - b\bar{x} \tag{4.3}$$

where $\overline{\ln(x)}$ is the mean of the $\ln(x_i)$.

Using the data, the values of the parameters are calculated as a = 17.777 and b = 0.1932. Then the estimated equation becomes

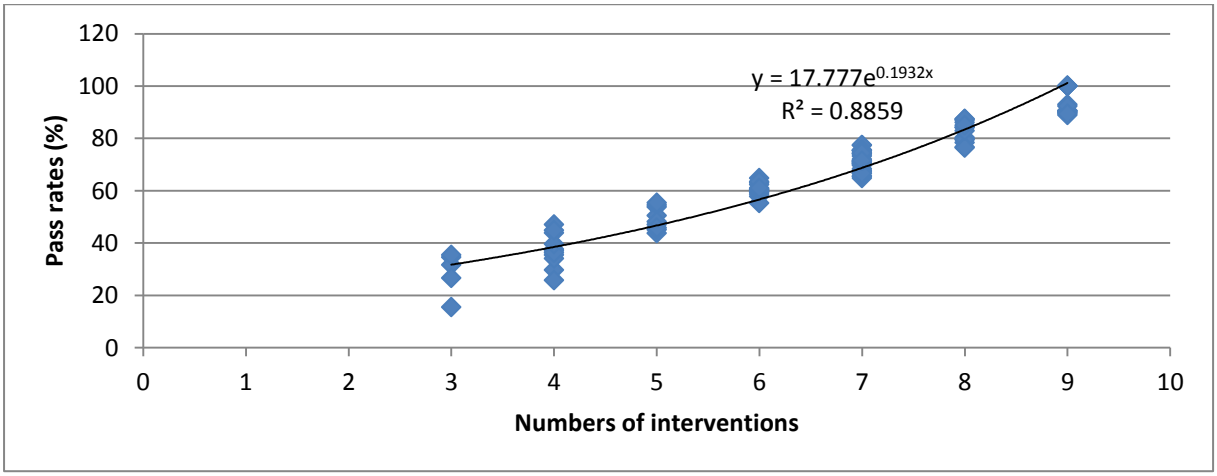
$$y_{est} = 17.777e^{0.1932x}. \quad (4.4)$$

The coefficient of determination, given by the R-squared, is

$R^2 = 0.8859$.

The exponential curve on the graph looks as follows:

Figure 4.2: Exponential regression equation



4.3.2 Linear equation

The linear equation takes the form

$$Y = a + bx + \varepsilon \quad (4.5)$$

where, again, the parameters a and b are to be estimated. The formulae for these two parameters are:

$$b = \frac{n\sum xy - \sum x \sum y}{n\sum x^2 - \left(\sum x\right)^2} \quad (4.6)$$

and

$$a = \bar{y} - b\bar{x} \quad (4.7)$$

Using the data, the estimated equation is

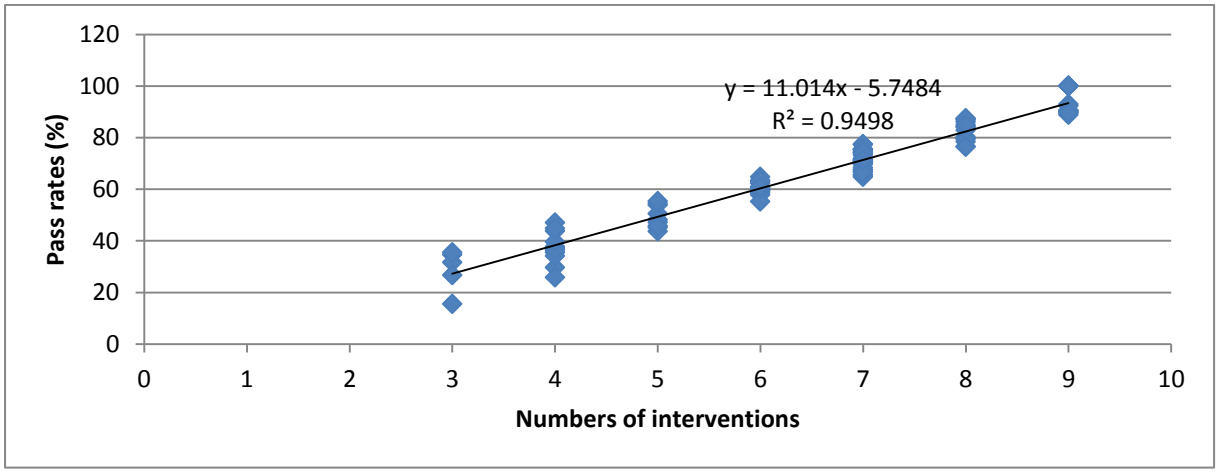
$$y_{est} = 11.014x - 5.7484. \quad (4.8)$$

The coefficient of determination, given by the R-squared, is

$$R^2 = 0.9498.$$

The linear graph on the graph looks as follows:

Figure 4.3: Linear regression equation



4.3.3 Logarithmic equation

The logarithmic equation takes the form

$$Y = a\ln(x) + b + \varepsilon \quad (4.9)$$

where, again, the parameters a and b are to be estimated. The formulae for these two parameters are:

$$b = \frac{n\sum \ln(x_i)y_i - \sum \ln(x_i)\sum y_i}{n\sum (\ln(x_i))^2 - \left(\sum \ln(x_i)\right)^2} \qquad (4.10)$$

and

$$a = \bar{y} - b\overline{\ln(x)} \qquad (4.11)$$

Using the data, the estimated equation is

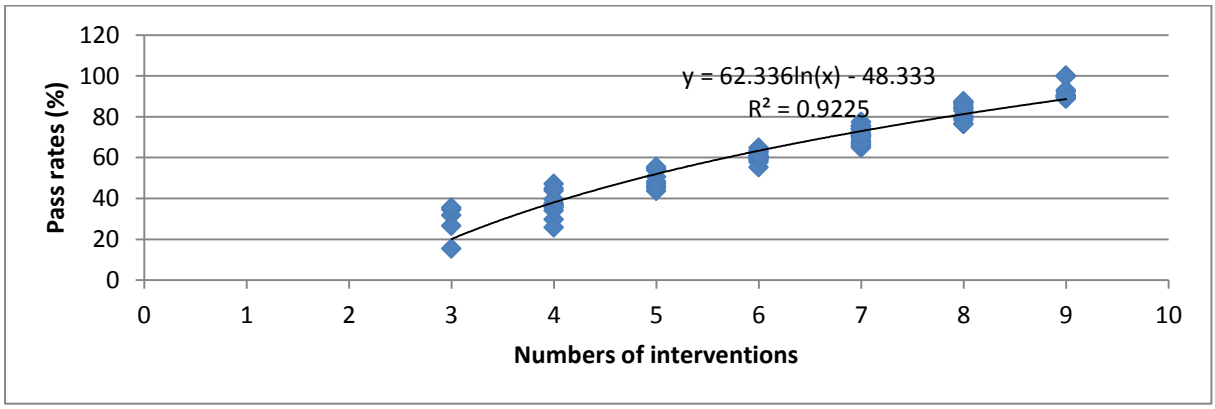
$$y_{est} = 62.3361\ln(x) - 48.333. \qquad (4.11)$$

The coefficient of determination, given by the R-squared, is

$$R^2 = 0.9225.$$

The logarithmic graph on the graph looks as follows:

Figure 4.4: Logarithmic regression equation



4.3.4 Curvilinear regression

The curvilinear regression (Bless & Kathuria, 1993) of the variable y on x is given by the formula

$$Y = a_0 + a_1 x + a_2 x^2 + a_3 x^3 + \ldots$$

The above form is used to present the quadratic (second power) regression below, as well as the subsequent powers of the regressions. The value of R-squared for each equation is used if it is necessary to proceed beyond to a higher power/exponent.

In order to find solutions for the coefficients of the polynomial we might use the least squares estimation. The solution is based on the linear equation. Rich and Schmidt (2004) explain the simple regression derivation that in the measured quantity y (depended variable) is a linear function of x (independent variable), i.e. $y = a_0 + a_1 x$, the most probable values of a_0 (intercept) and a_1 (slope) can be estimated from a set of n pairs of experimental data (x_1, y_1), (x_2, y_2)…, (x_n, y_n), where y-values are contaminated with a normally distributed - zero mean random error (e.g. noise, experimental uncertainty). This estimation is known as least-squares linear regression. Higham (2002) states that least-squares linear regression is only a partial case of least-squares polynomial regression analysis. By implementing this analysis, it is easy to fit any polynomial of *m* degree

$$Y = a_0 + a_1 x + a_2 x^2 + a_3 x^3 + \ldots + a_m x^m$$

to experimental data (x_1, y_1), (x_2, y_2)…, (x_n, y_n), (provided that $n \geq m+1$) so that the sum of squared residuals S is minimized:

$$S = \sum_{i=1}^{n} \left(Y_i - \hat{Y}_1 \right)^2$$

$$= \sum_{i=1}^{n} Y_i - \left(a_0 + a_1 x_i + a_2 x_i^2 + a_3 x_i^3 + \ldots + a_m x_i^m \right)^2$$

By obtaining the partial derivatives of S with respect to a_0, a_1, .., a_m and equating these derivatives to zero, the following system of m-equations and m-unknowns (a_0, a_1, .., a_m) is defined:

$$s_0 a_0 + s_1 a_1 + s_2 a_2 + \ldots + s_m a_m = t_0$$

$$s_0 a_0 + s_1 a_1 + s_2 a_2 + \ldots + s_m a_m = t_0$$

. . . . . .

$$s_0 a_0 + s_1 a_1 + s_2 a_2 + \ldots + s_m a_m = t_0$$

where:

$$s_k = \sum_{i=1}^{n} x_i^k, \qquad t_k = \sum_{i=1}^{n} Y_i x_i^k,$$

Also, always, $s_0 = n$. This system is known a system of normal equations. The set of coefficients: a_0, a_1, …, a_m is the unique solution of this system. For m = 1, the familiar expressions used in linear least-square fit are obtained:

$$a_0 = \frac{\sum_{i=1}^{n} x_i^2 \sum_{i=1}^{n} y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i x_i}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2}$$

and

$$a_1 = \frac{n\sum_{i=1}^{n} y_i x_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2}$$

Higham (2002) counsels that accuracy and stability of numerical algorithms in contemporary practice depend on generating values using a computer. Therefore, in the formulae of the forthcoming quadratic and higher power curvilinear functions the equations of formulae for the equations of the coefficients are not given. The coefficient values are generated from the spreadsheet of statistical packages.

4.3.4.1 Quadratic equation

The quadratic equation takes the form

$$Y = ax^2 + bx + c + \varepsilon \qquad (4.12)$$

The parameters a, b and c are to be estimated. Using the data, .these values obtained for these coefficients are a = 0.0503, b = 10.391 and c = -3.9718. The estimated equation is

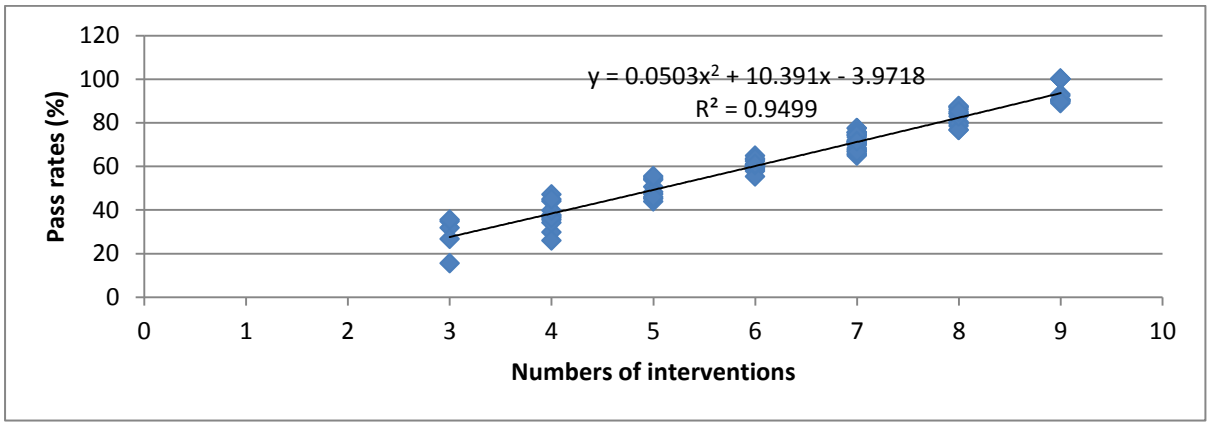
$$y_{est} = 0.0503x^2 + 10.391x - 3.9718. \quad (4.13)$$

The coefficient of determination, given by the R-squared, is

$$R^2 = 0.9499.$$

The quadratic form of the relationship appears on the following graph:

Figure 4.5: Quadratic regression equation



4.3.4.2 Polynomial of third power relationship

The 3rd power polynomial relationship takes the form

$$Y = ax^3 + bx^2 + cx + d + \varepsilon \quad (4.14)$$

The parameters a, b, c and d are to be estimated. Using the data, estimates of the coefficient values are a = -0.0477, b = 0.9382, c = 5.1807 and d = 5.5083. The corresponding curvilinear equation then becomes

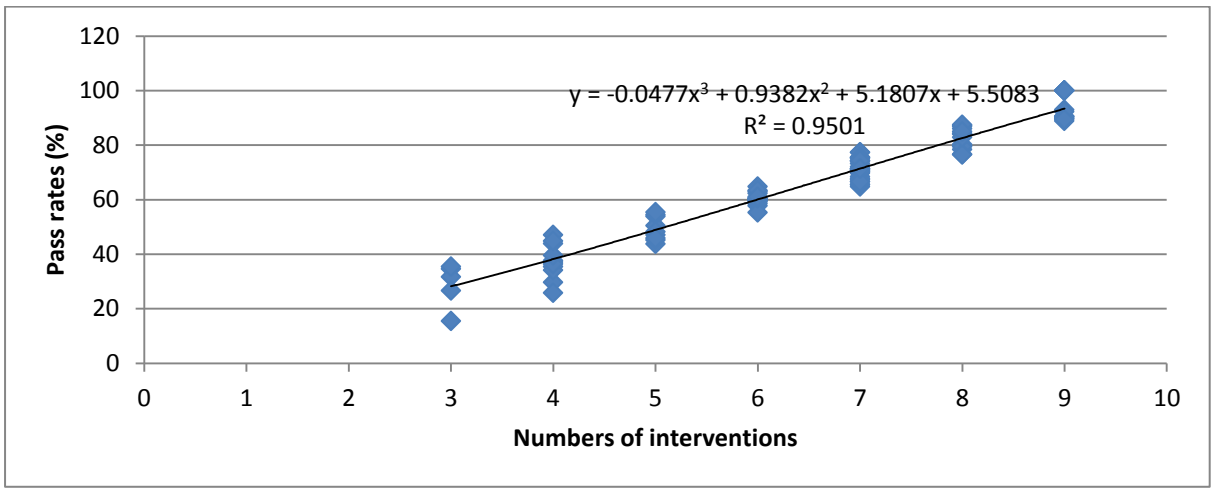
$$y_{est} = -0.0477x^3 + 0.9382x^2 + 5.1807x + 5.5083. \quad (4.15)$$

The coefficient of determination, given by the R-squared, is

$R^2 = 0.9501$.

The third power polynomial form of the relationship appears on the following graph:

Figure 4.6: Power 3 polynomial regression equation



4.3.4.3 Polynomial of fourth power relationship

The 4th power polynomial relationship takes the form

$$Y = ax^4 + bx^3 + cx^2 + dx + e + \varepsilon \qquad (4.16)$$

The parameters a, b, c, d and e are to be estimated. Using the data, the estimates of these coefficients are a = 0.0341, b = –0.8738, c = 8.1369, d = –21.403 and e = 40.412. The equation then becomes

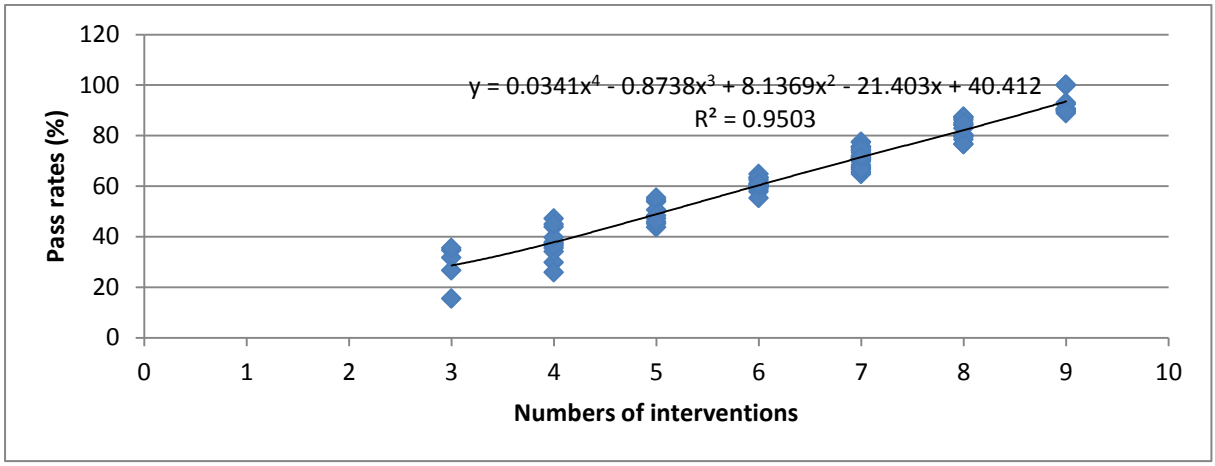
$$y_{est} = 0.0341x^4 - 0.8738x^3 + 8.1369x^2 - 21.403x + 40.412. \qquad (4.17)$$

The coefficient of determination, given by the R-squared, is

$R^2 = 0.9503$.

The fourth power polynomial form of the relationship appears on the following graph:

Figure 4.7: Power 4 polynomial regression equation



The higher exponents cannot be considered beyond this point because the R-square values of the last two have not shown any significant increase.

4.3.5 Power relationship

The power relationship takes the form

$$Y = ax^b + \varepsilon \qquad (4.18)$$

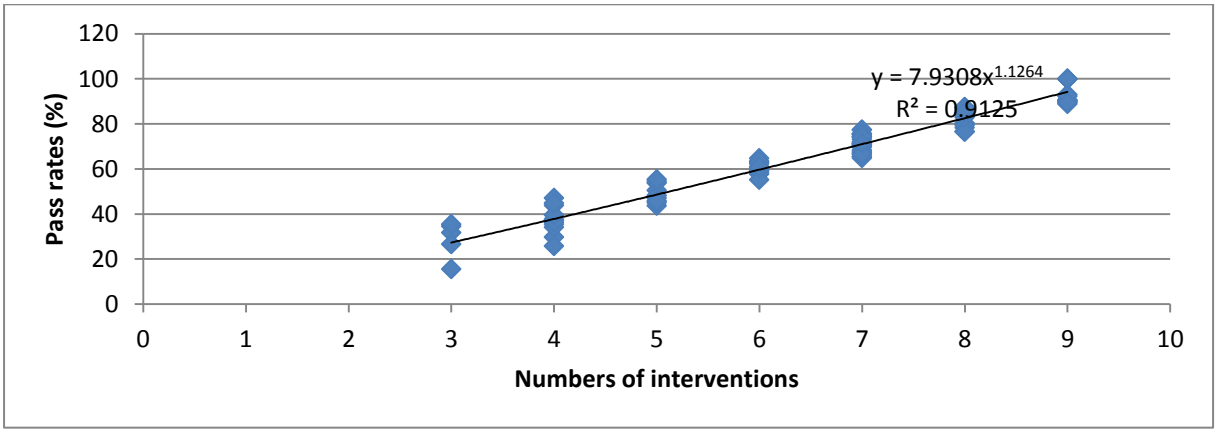
The parameters a and b are to be estimated. Using the data, the estimates are a = 7.9308 and b = 1.1264. The power relationship appears as

$$y_{est} = 7.\ 9308x^{1.1264}. \qquad (4.19)$$

The coefficient of determination, given by the R-squared, is

$$R^2 = 0.9125.$$

Figure 4.8: Power regression equation



Observation

The regression equations are all optimistic in using the number of interventions to boost the matric pass rates in the schools of Letlhabile because all their R-squared values are high (close to 1). The correlations are also all positive by observing the positive slopes of their line charts. This means that the number of interventions improve the pass rates. At the moment the different patterns have to be compared.

4.4 Preliminary Comparisons of Polynomials

4.4.1 Coefficients of linear equation and polynomials

The coefficients a = 0.8225 and b = 0.0862 of the linear equation are small (below 1). When the power of the independent variables is increased (to quadratic form and larger powers) the coefficients become even much smaller. These coefficients are tested in the next section to determine if they should be included. What is also not impressive is the increase in the value of R-squared. When the higher power from linear to quadratic and other higher power curvilinear equations are considered, the increase in the value of R-squared is slight. It is insignificant.

4.4.2 Multicollinearity

The polynomials are extensions of the linear equation. However, they are expected to have some multicolinearity in variables since the added variables are powers of the initial variable.

In comparing the linear equation with the quadratic equation the value of R-squared rises by 0.0002 (from 0.9498 to 0.95) from linear to quadratic. This is an insignificant increase since it is an almost 0% increase. To estimate the tolerance, then $R_{x^2} = 0.0002$. Then

$$\begin{aligned} tolerance &= 1 - R_{x^2} \\ &= 0.9998 \end{aligned}.$$

On the other hand

$$\begin{aligned} VIF &= \frac{1}{tolerance} \\ &= 1 - R_{x^2} \\ &\approx 1 \end{aligned}.$$

There is no multicollinearity problem since tolerance is neither less than 0.1 nor 0.2. on the other hand the VIF is less than both 5 and 10. The increases in R-squared values from the quadratic equation to the third and fourth polynomials are also slight increases, almost insignificant. Even though there is no problem of multicollinearity, polynomials of higher than power 1 do not lead to improvement in the relationship between the matrix pass rates and the number of interventions. Thus the higher power exponential curvilinear equations are eliminated and only the linear equation will be used to contest the inclusion into viable prediction methods. As a result the methods being compared for use in using the numbers of observations to predict pass rates follow in Table 4.1 below.

Table 4.1: Models in the contest

Model	***Estimated equation***
Exponential model	$y_{est} = 17.777e^{0.1932x}$
Linear model	$y_{est} = 11.014x - 5.7484$
Logarithmic model	$y_{est} = 62.3361\ln(x) - 48.333$
Power model	$y_{est} = 7.9308x^{1.1264}$

4.5 Bias and Precision, Goodness-of-fit, Statistical Tests of Coefficient Values

4.5.1 Measuring error

It is important to know the extent of accuracy of the estimated values and the reliability of the models developed. For notation purposes the actual values are denoted by A and the estimates by E. This is the bases to understand how far the estimates would be relative to the actual values. This is estimated by the difference $A - E$. The measures of importance to this study were discussed in Chapter 3. They are the cumulative forecast error, mean error, mean square error, root mean square error, standard deviation, mean absolute deviation, and mean absolute percentage error. Some of the measures are used in developing others as they do not need to be included when the main ones are used. Hence, the measures proposed in this section are the mean error (ME), mean square error (MSE), mean absolute deviation (MAD), standard deviation of errors (s_e) and mean absolute percentage error (MAPE).

4.5.2 Bias and precision

The measures of bias and precision are based on residuals, where each residual is based on the difference between the actual (i.e. observed) value and the corresponding estimate. Then, logic leads to that a smaller residual is more desirable because the estimate would be close to the actual value. Thus, the measures of bias and precision are lower for better model. The calculations for the various measures appear in Annexure C and the measures are presented in Table 4.2 below.

Table 4.2: Measures of bias and precision

	Exponential	**Linear**	**Logarithmic**	**Power**
ME	0.122799	-0.00219	0.989645	-19.7444
s_e	194.6688	33.91778	41.19732	75.65476
MAD	4.504603	3.461642	4.162263	19.91056
MAPE	492087.4	10.95757	43510.31	29.95186
MSE	31.78981	18.45349	29.50305	484.533
R-squared	0.8859	0.9498	0.9225	0.9125

Mean error

The linear regression and power functions overestimate the pass rates due to their negative mean deviation values. On the other hand, the exponential and the logarithmic functions, due to their positive mean deviation values, overestimate them. The linear regression does not seem to provide much deviant underestimates due to a very low value of its mean deviation. On the other hand the mean deviation from the power function is much higher in absolute value. The logarithmic regression's deviation from the actual would be much lower than the power function, but still higher in absolute value from the linear regression.

Standard deviation of errors

The standard errors of the exponential and power functions are too high compared to those of linear and logarithmic functions. This makes these function less desirable in using them to predict the pass rates using the numbers of observations.

Mean absolute deviation

The mean absolute deviation of the power function is much higher than for the other three functions. Thus, the power function becomes again unsuitable for use in this context.

Mean absolute percentage error

The mean absolute deviations of the exponential and logarithmic functions are much out of range of the other two. The values they produce indicate a high level of risking wrong predictions when estimating pass rates using numbers of interventions.

Mean square error

The means square error of the power function is too high. Others are much lower.

R-squared

The R-squared value of the linear function is the largest of the four.

Observation

The bias and precision measures (ME, s_e, MAD, MAPE, and MSE) of the exponential, logarithmic and power functions are all higher than those of the linear function. Also each of these methods has at least one measure that makes them undesirable to use in the predictions

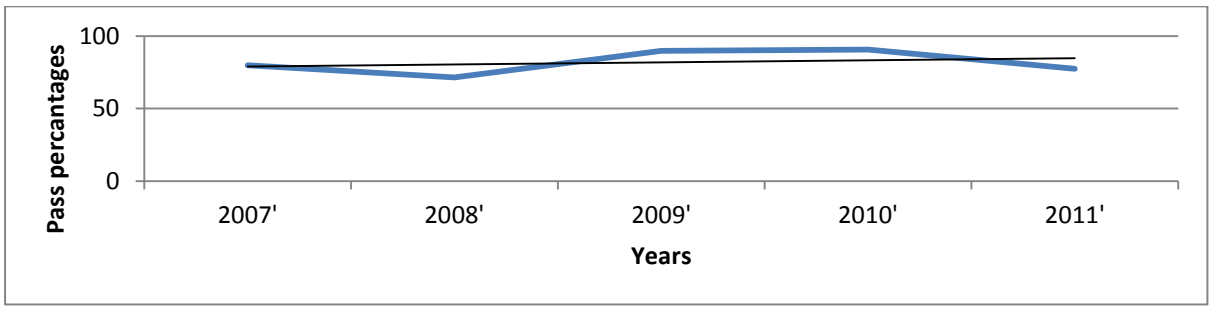
of the pass rates based on the numbers of interventions. The power function is the worst method as it has no measure making it attractive. The exponential function has the standard deviation and the MAPE that are 'way too high'. The logarithmic function has a MAPE that is too high, but less than that of the exponential function. In contrast the linear function has the least values in all these measures, and has no value that makes it undesirable. The linear function is the best method in terms of all the criteria. To make it even stronger, its R-squared value is the highest of all the others.

4.6 Time Series Line Charts

Line charts (or time plots) are used to present the examination mark percentages (displayed as pass percentages). The purpose was to examine the trend patterns of each school's results, as well as determine the growth pattern of the pass rates of each school. Each line chart also has an accompanying trend line for this purpose.

4.6.1 School number 1

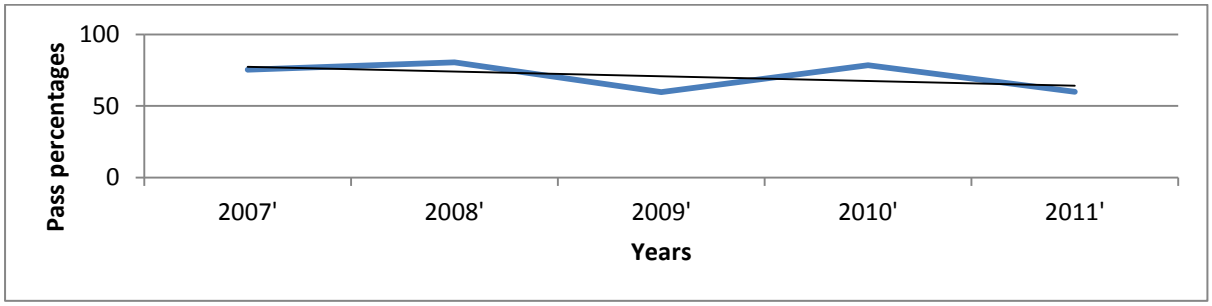
Figure 4.9: Line chart of school 1 matric pass rates



The pass rates for this school are high, all ranging over 70% to as far as just above 90%. The pass rate for the second year of the study showed a decline, but the next year's rate picked up again. It was the year in which the number of interventions also dropped. The last year also went slightly down, which was the year in which the new examination system was implemented for all South African schools. The trend line shows a slight upward slope, which is an indication of s slow increase in the pass rates for this school. It was also the year in which the number of interventions dropped.

4.6.2 School number 2

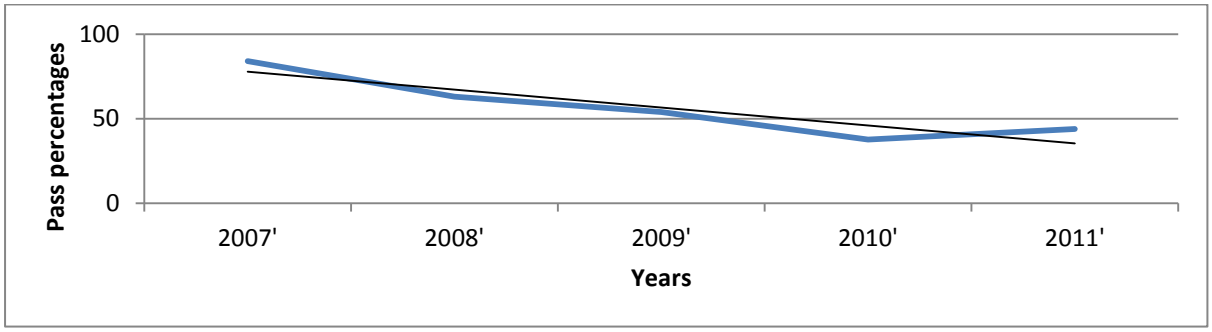
Figure 4.10: Line chart of school 2 matric pass rates



For this school the pass rates started with a high value. As it looks, every alternative year the rates go up and down. The numbers of interventions also fluctuate in the alternative years in the same way as the pass rates. The last year of the study was one of the years that showed low pass rates for the school. The high rates reach about 80%. The lower ones has a 75%, but all others are just below 60%. The trend line shows a decrease in the pass rates.

4.6.3 School number 3

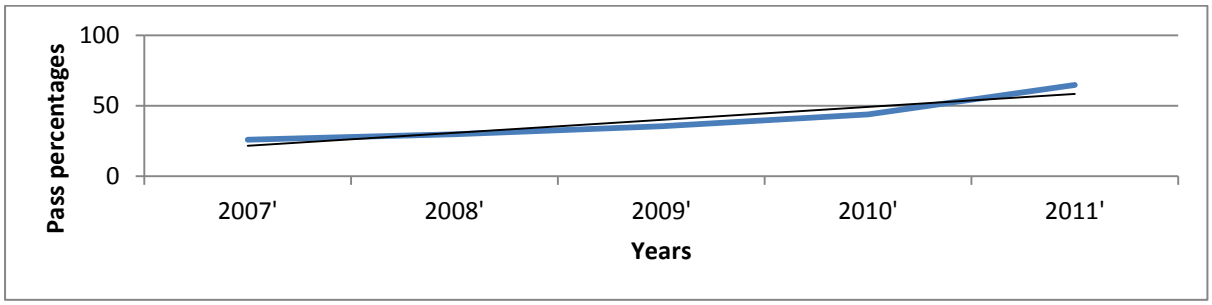
Figure 4.11: Line chart of school 3 matric pass rates



This school stared well in the first year of this study with a mark of about 85%. It was the year of most interventions in the study. From there onwards the other marks as well as the numbers of interventions have been constantly declining over the years. Even though the last year of the study showed a slight increase, it was high enough only to exceed the mark for its predecessor year. The trend line of pass rates was fast showing a decline.

4.6.4 School number 4

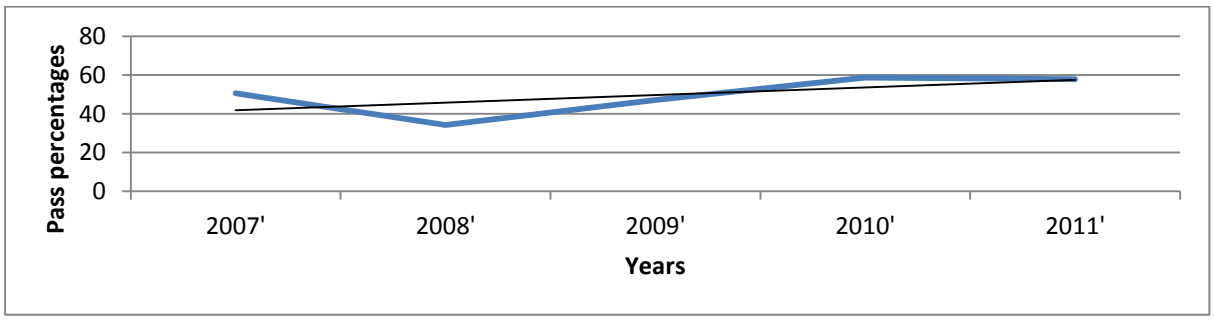
Figure 4.12: Line chart of school 4 matric pass rates



This one started badly. The numbers of interventions were also at its lowest. However, each year the interventions were also increased, and the pass rates have also been increasing. The highest increase occurred for the year in which the examination system was combined for all the schools of South Africa. The high the pass rates for this school have not matched the level of marks displayed by performing schools, but the increasing interventions indicate to have improved the pass rates. The trend line shows a fast upward trend.

4.6.5 School number 5

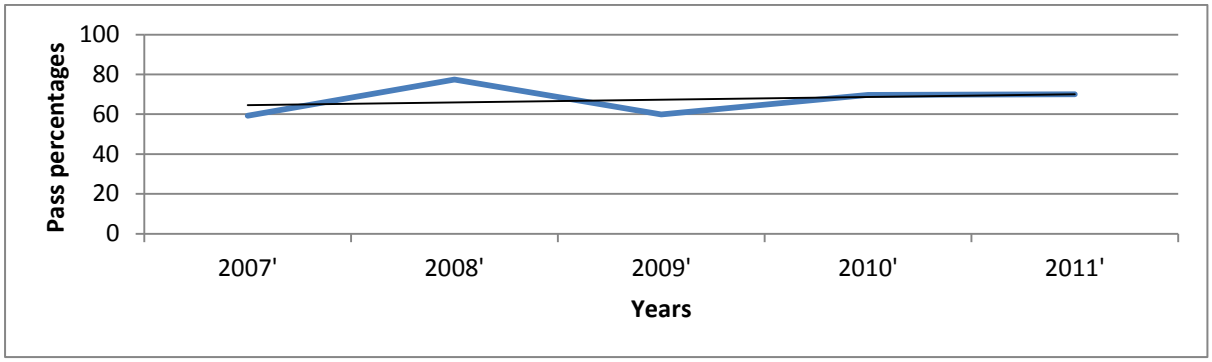
Figure 4.13: Line chart of school 5 matric pass rates



This school also show an increasing trend, after having started poorly with a very low pass rate in the first year of the study. The number of interventions was also low at the time. In the following year the interventions were reduced, and the lowest pass rates were also realised in that year. The interventions have since increased slowly over the years, and the resulting pass rates have also shown an increase over the same years. The trend line shows an increase of the pass rates over the years as well.

4.6.6 School number 6

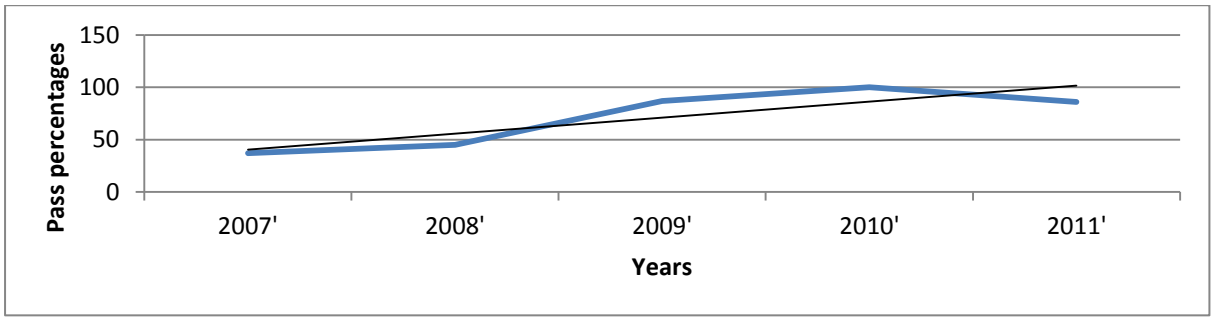
Figure 4.14: Line chart of school 6 matric pass rates



The pass rates in this school have been somewhat satisfactory. The trend line of the pass rates over the years shows a slight increase. The year showing the highest pass mark, which is the second year of the study, is also the year of the highest intervention.

4.6.7 School number 7

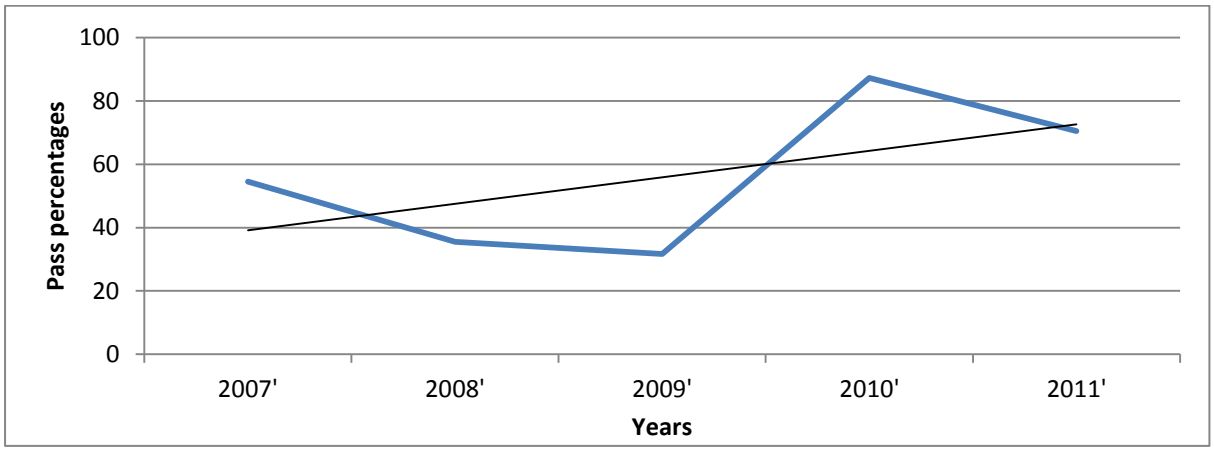
Figure 4.15: Line chart of school 7 matric pass rates



This school started very low on pass rates, and the number of interventions was also low. The pass rates have been going rapidly up over the years, and this pattern is also showing in the numbers of interventions in the same years. The trend line is showing a very fast increase. In one of the years the maximum ideal pass rate (100%) was also achieved. The increases and decreases in pass rates occur in similar patterns.

4.6.8 School number 8

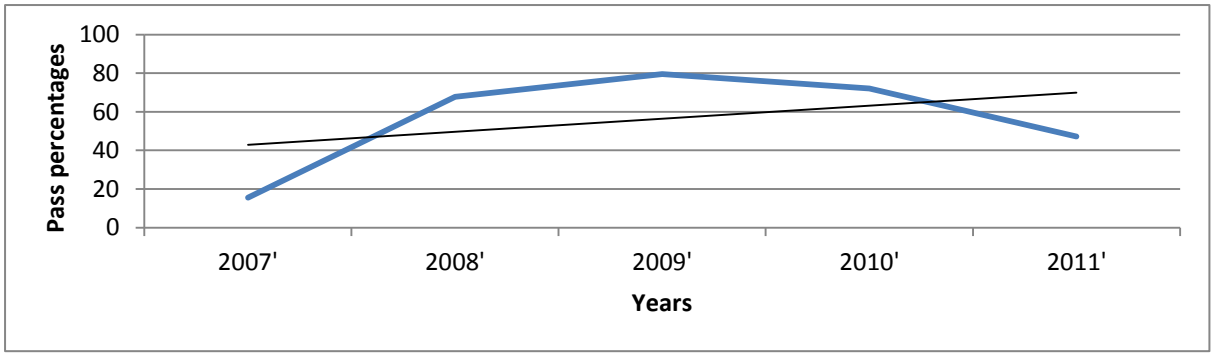
Figure 4.16: Line chart of school 8 matric pass rates



This school shows declines in pass rates in the earlier years and increases in the last years. In these same years the numbers of inventions were also declining in the early years and then increasing in the later years. The trend line displays an increase in the pass rates.

4.6.9 School number 9

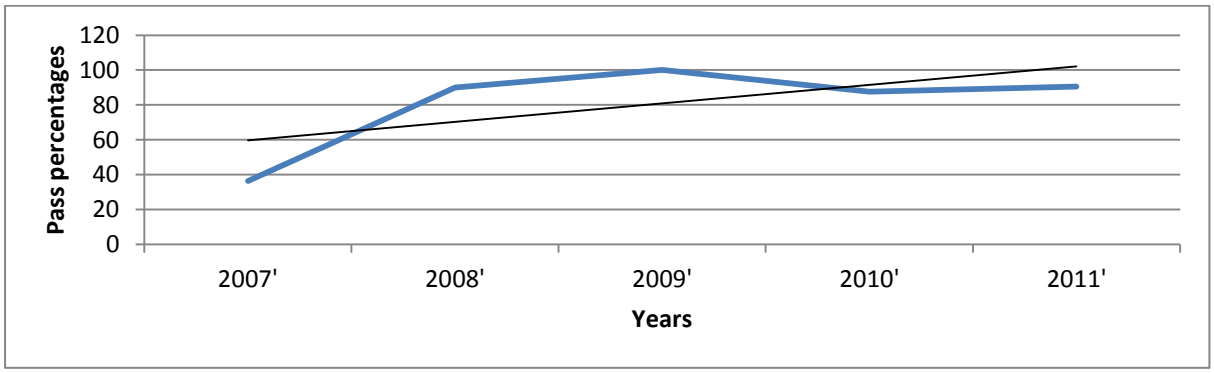
Figure 4.17: Line chart of school 9 matric pass rates



This school started badly, increased in the middle years showing progress in improving the pass rates. Then in a parabolic fashion the pass rates show a substantial decline. The numbers of interventions take the same pattern, they were low at the early and the late years of this study, and high in the middle years as for the pass rates. The trend line shows an increase though, which is an indication of improvement when compared with the earlier years.

4.6.10 School number 10

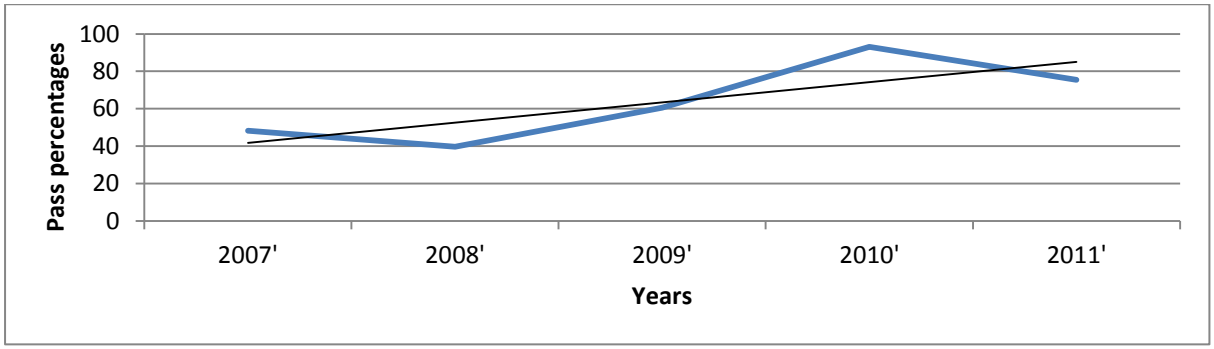
Figure 4.18: Line chart of school 10 matric pass rates



In this school improvement in pass rates is gradual over the years. Interventions have also been increased gradually. The school started very low, and then increased impressively. The years showing low marks coincide with those showing few interventions. The trend line also reveals a dramatic increase in pass rates that include 100% and over 90% rates.

4.6.11 School number 11

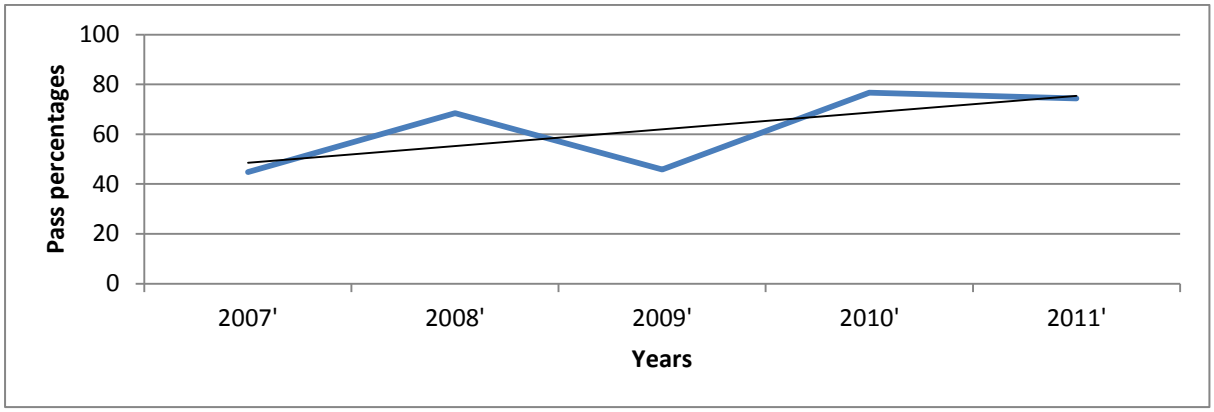
Figure 4.19: Line chart of school 11 matric pass rates



This is another school that shows an increase in pass rates. A pass rate of below 40% came earlier in the study, and close to 95% was achieved during the increase period. The year of lowest pass rate was the same one in which least interventions were shown. Also, the highest pass rate was shown in the year of most interventions. The trend line shows an increase. In the last year of the study the number of interventions was reduced. In this year as well, the pass rate dropped.

4.6.12 School number 12

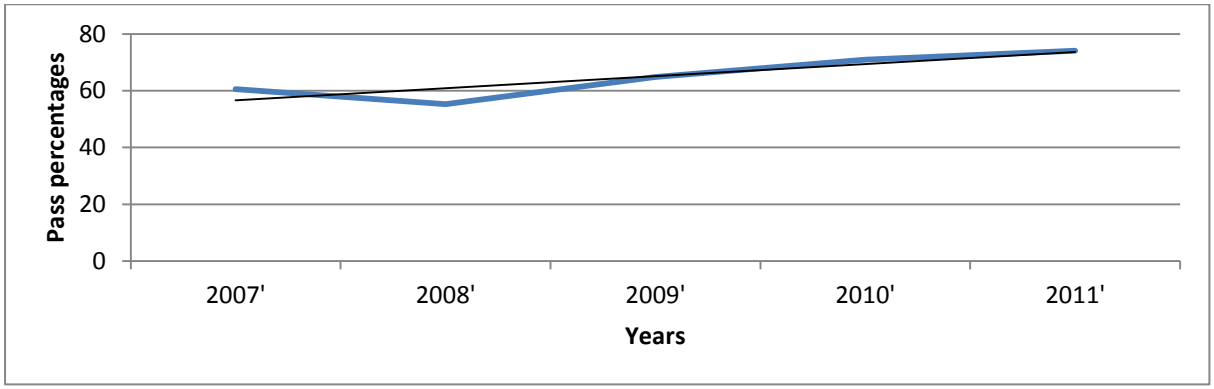
Figure 4.20: Line chart of school 12 matric pass rates



The pass rates of this school are also increasing impressively over the years. Alternative years show increase-decrease pattern, and this pattern occurs with the number of interventions as well. The trend line shows an increase.

4.6.13 School number 13

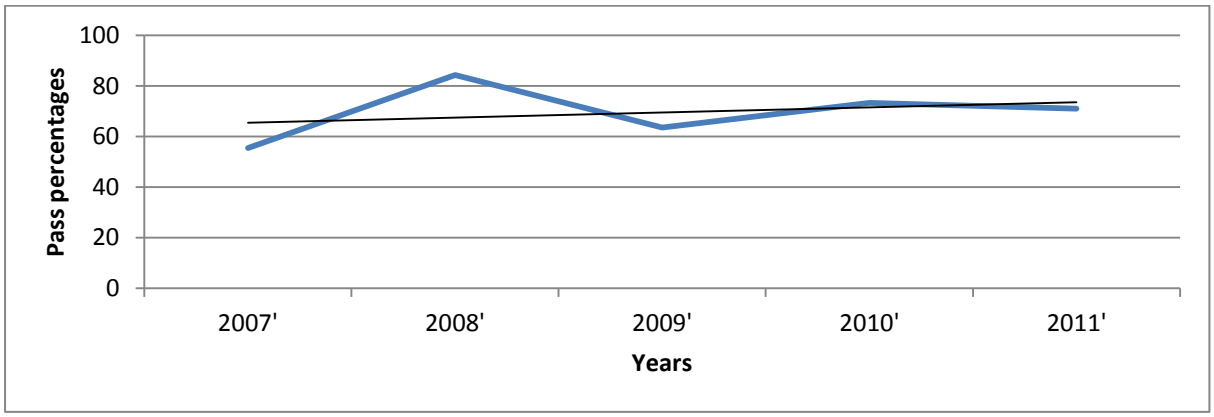
Figure 4.21: Line chart of school 13 matric pass rates



A gradual increase in the pass rates is shown in this school over the study period. The second year showed a decline in the pass rate though. In the same year, it was realised that the number of interventions had been reduced. In the other years the interventions were increased. The trend line displays an increase.

4.6.14 School number 14

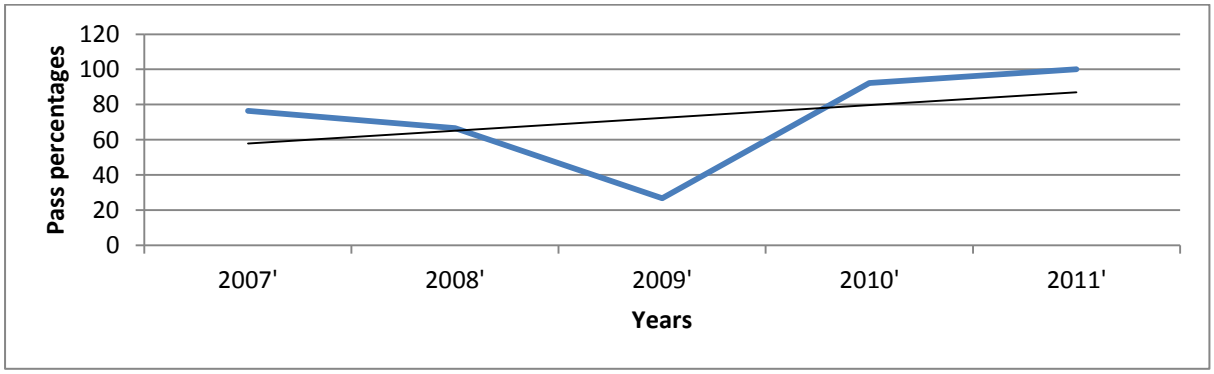
Figure 4.22: Line chart of school 14 matric pass rates



This school produce constantly good pass rates, with a slight increase over the years. In the second year of this study, most interventions were used. The pass rate for that year was also outstanding. In the years thereafter the interventions were immediately reduced but increased slowly compared to the first year. They have been increased gradually. The pass rates also show the same pattern. The years of more interventions are the ones of better pass rates. As for the trend line, a slow increase is shown in the pass rates.

4.6.15 School number 15

Figure 4.23: Line chart of school 15 matric pass rates

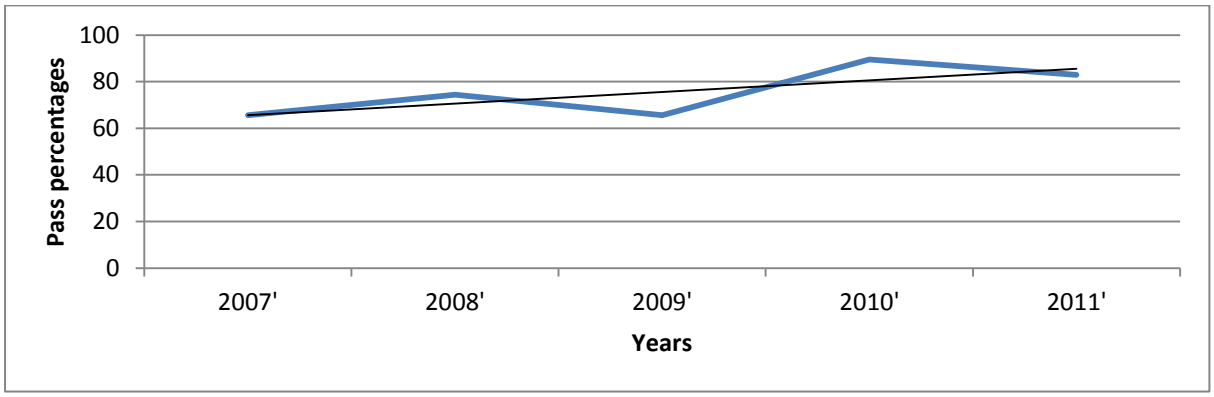


This is one school showing mixed results in different years, with an overall pattern displayed by the trend line as increasing pass rates. The third year had the least interventions. The same

year showed the poorest pass rates. In the years of more interventions, higher pass rates were also shown.

4.6.16 School number 16

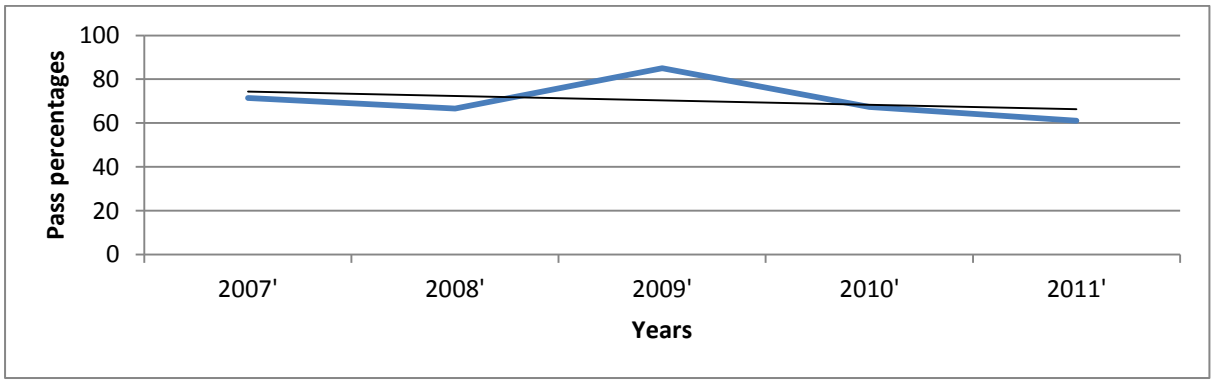
Figure 4.24: Line chart of school 16 matric pass rates



The pass rates in this school were satisfactory from the start. A slow increase in the number of interventions was shown, except in the third year in which the interventions were few. In this year the pass rate dropped as well. The interventions have since been increased slowly, and the pass rates are gradually showing to increase. The trend line confirms this pattern

4.6.17 School number 17

Figure 4.25: Line chart of school 17 matric pass rates

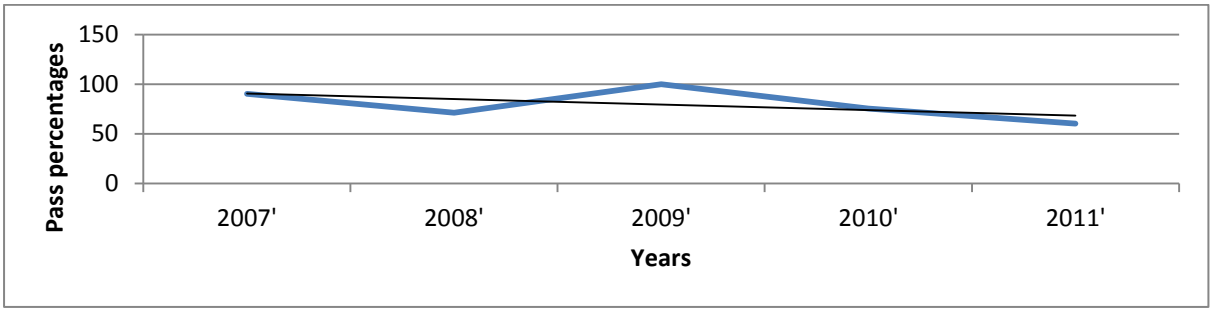


In this school the pass rates are showing to drop. They have a maximum pass rate of about 85% in the third year when the interventions were many. The minimum occurs at 60% when

the interventions were fewest. The trend line confirms the decreasing pattern. In the third year of this study, interventions were increased expressively to a high number. An excellent pass rate was achieved. Since then the interventions were reduced gradually over the years. These same years also showed lower pass rates.

4.6.18 School number 18

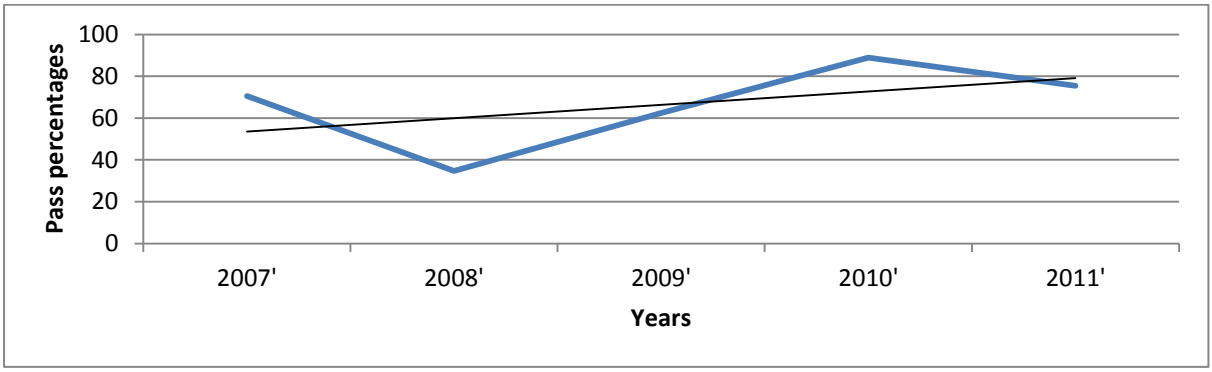
Figure 4.26: Line chart of school 18 matric pass rates



In this school as well, the pass rates are showing to drop, and the trend line confirms this pattern. In the third year of this study, interventions were increased expressively to a high number. An excellent pass rate of 100% was achieved in the year of most interventions. Since then the interventions were reduced gradually over the years. The minimum pass rate of 60% was reached in the fourth year of fewest interventions. These same years of declining pass rates also showed lower pass rates.

4.6.19 School number 19

Figure 4.27: Line chart of school 19 matric pass rates



The second year of this study had very few interventions. The pass rate for the year also went to its lowest. In other years the numbers of interventions were increased slowly. The pass rates in these years also increased gradually. The trend line also shows an increasing pattern.

Observation

Mixed pass rates are shown for the different schools as some generally do well while others cannot be classified this way. Also, some pass rates increase over the years while others decline over the same period. However, all the schools have shown a consistent pattern in that in the year of many interventions the pass rate also becomes higher. This study was aimed at examining the relationship between the number of interventions and the pass rates in the schools of this study.

4.7 Significance of Correlations

Correlational research assesses if the two population being related are correlated, hence in this study it would assess the existence of a relationship between pass rates and numbers of interventions are correlated. Curwin and Slater (2002) point out that the null hypothesis denies the existence of a linear relationship. Denote the correlation coefficient between the two populations by ρ. Hence, the null hypothesis is

$$H_0 : \rho = 0 \tag{4.35}$$

Bless and Kathuria (2003) illustrates that in cases where $n \geq 30$, the test statistic used is the z given by the formula

$$z = r\sqrt{n-1} \tag{4.36}$$

At the 5% level of significance the critical values for this two sided test is $z = 1.96$. The test is significant if the calculated value exceeds the critical value. The smallest R-squared is the one obtained for the logarithmic regression of 0.8829. The estimated correlation coefficient is about the square root of this value. Hence, the test statistic for the logarithmic regression is the value

$$z = r\sqrt{n-1} = \sqrt{0.8829} \times \sqrt{94} = 9.11.$$

The test is significant and the null hypothesis is rejected. Hence, in all the regression forms considered, the possibilities of the existence of all these relationships exist. This confirms that the measures of bias and precision considered earlier in the chapter were based indeed on feasible relationships.

4.9 Conclusion

The chapter used regression methods and time series analysis to establish a relationship between pass rates and the numbers of interventions in the various schools of Letlhabile area in the North-West Province of South Africa. The various relevant methods were considered and those that showed to be inadequate were eliminated in the early stages of the analyses. A statistical test was used to determine the suitability of the remaining methods for use in predicting the pass rates. The suitability was established, but the best from these methods was needed. The others were compared more comprehensively using the statistical measures. In the next chapter the analyses are consolidated and the best method is identified. That chapter will close with recommendations for increasing the pass rates, limitations of the study and the recommendations for further studies.

CHAPTER 5: CONCLUSION AND RECOMMENDATIONS

5.1 Introduction

This study evaluated the impact that the numbers of interventions have in matric classes of the Letlhabile area of the North-West Province of South Africa. Bivariate data consisting of the pass rates (Y) and the numbers of interventions (X) were obtained from the circuit office in Letlhabile. These data were provided for the 19 schools in tabular formats. Since the data required relationships between the pass rates and the numbers of interventions, regression methods were needed. These were discussed in Chapter 2. Also, the data were presented over five years. This made it necessary to establish the pattern over the five years involved, to determine how the numbers of interventions over these years for each of the 19 schools in the five years could have affected the pass rates. This was explainable using time series analysis since the data were time based. Hence, time series analysis was presented ion Chapter 3. The previous chapter introduced various tentative prediction models that seemed relevant for use in the prediction of the pass rates from numbers of interventions. Graphs and statistical tests were used to analyse the data. Preliminary tests were used to eliminate models that were showing inadequacy. The chapter then presented analyses of the remaining methods using regression analysis and time series. The various comprehensive comparisons were also presented. In addition, graphs were used to determine the way the numbers of interventions affected the pass rates. The findings obtained are explained in the forthcoming sections of this chapter:

5.2 Selection of the Best Method

Time series methods showed the changes over the different periods in which the study was undertaken. There were years in which the numbers of interventions were low for some schools, and the years in which they were high. Time series analysis was used to detect the patterns of the pass rates data during these changes. The methods that were included in the contest for determining the best method were the exponential, linear, logarithmic and power functions. This stage occurred after eliminating curvilinear methods which were indicating to be inadequate as they were showing multicollinearity of the variables involved. In this section the already clearly leading model for predicting pass rates from the numbers of interventions

is formally selected. All the five methods were found sensible to use when assessed through the correlation test. They were then compared using the measures of bias and precision, as well as arguing using the coefficient of determination (R-squared). The illustrations are in Table 4.2 of Chapter 4. The contents of the table are reproduced below.

Table 5.1: Summary table of comparison statistics

	Exponential	**Linear**	**Logarithmic**	**Power**
ME	0.122799	**-0.00219**	0.989645	-19.7444
s_e	194.6688	**33.91778**	41.19732	75.65476
MAD	4.504603	**3.461642**	4.162263	19.91056
MAPE	492087.4	**10.95757**	43510.31	29.95186
MSE	31.78981	**18.45349**	29.50305	484.533
R-squared	0.8859	**0.9498**	0.9225	0.9125

The bold values in Table 5.1 indicate that the values are the most desirable according to the way the measures are used. For example the first five measures should be low for a measure to be considered desirable while the last one (the R-squared) should be high for being desirable. The discussions and summary were presented below Table 4.2 of the previous chapter. That explanation is clarified in the next section. The bold numbers are shown to be all allocated to the linear function.

5.3 Verdict from comparisons

The measures of precision were used to compare the methods. These methods depend on error analysis which is used to determine the method leading to least error when used in the prediction of future values. A measure of quality of fit to the data was also calculated for the methods being compared. In brief, the linear function emerged as the leading model in terms of all the criteria used to determine the most accurate method. As a result the linear regression method has been selected as the model for use in predicting pass rates based on the number of interventions for the schools in the area of Letlhabile in the North-West Province of South Africa. In addition, the quality of this model, based on the R-squared value, has been estimated to about 95%, which is higher than the values obtained for the other methods. The 95% quality measure means that in using the linear equation to predict the matric pass rates in

the high schools in the Letlhabile area, the number of intervention would be able to account for about 95% of the changes in the matric pass rates.

5.4 Observations from numbers of interventions

The schools varied in pass rates patterns over the five-year period of this study. Some schools showed increasing pass rates while others showed decreasing pass rates over the period. However, more schools showed to have improving pass rates compared to very few that showed declining pass rates. In all the schools involved though, the pass rates increased each time the number of interventions increased and decreased when interventions decreased. This was the case irrespective of the nature of the trend of the pass rates. That means, for the pass rates that were decreasing, during the time of more intervention the pass rates went high for that year while it went lower down when interventions decreased. For the schools that had the increasing trend, in the years of more interventions the pass rates went even higher while the years of reduced interventions showed reduction in pass rates.

5.5 Limitations

The problem with the data appearing as the number of interventions does not provide all the necessary details of the types of interventions. The interventions method in the Letlhabile area is usually in the forms of Saturday classes, camps where students are taught prior to final year-end examinations, as well as specialised revision classes. However, it fails to mention the calibre of people who are involved in such interventions. Also, the data did not state the way each intervention is counted. It thus gives the impression that one event of intervention is be counted as one case. The people who provided the data were not the ones who collected them and there was no one to provide details.

It seems that certain small numbers of interventions could not make a sizable improvement in the pass rates. A minimum number of interventions that can make pass rates improvement possible should be known so that interventions can be geared to achieving improvement. This is because if a minimum is not reached the pass rates will not be improved. The other limitation is that it is not known how high the number of intervention will be to reach exhaustion. There could be a number of interventions beyond which no further improvement

in pass rates can be achieved. If this can be known, it will help so that no resources are provided. This is because any attempt to improve pass rates cannot happen if exhaustion is reached and will be wasted.

5.6 Recommendations

5.6.1 Recommendations for the study

The education administration in the Letlhabile area should

- Identify low performing schools in matric and introduce intervention programmes;
- Ensure that schools involved in intervention programmes are supported for sustained interventions over the years;
- Consider intervention at earlier schooling, not only at matric stage; and
- Encourage research around intervention strategies to improve the methods.

5.6.2 Recommendations for further research

More research should be embarked on to determine:

- The types of interventions that ensure pass rates improvements
- The minimum number of interventions required to ensure improvements in pass rates
- The maximum number of interventions to reach exhaustion
- The caliber of people used in the interventions

REFERENCES

Africa, H.P. (2005). *Audit: student failure, Report to UKZN*, University of KwaZulu Natal, Durban, South Africa.

Aldrich, J. (2005). Fisher and regression. *Statistical Science*, 20(4): 401-417.

Allen, L.J.S. (2010). *An introduction to stochastic processes with applications to biology, 2nd edition*. New York: Chapman and Hall.

Bless, C. & Kathuria, R. (1993). *Fundamentals of social statistics: an African perspective*. Wetton: Juta Company.

Bloomfield, P. (1976). *Fourier analysis of time series: An introduction*. New York: Wiley.

Boashash, B. (ed.), (2003) *Time-frequency signal analysis and processing: A comprehensive reference*. Oxford: Elsevier Science.

Box, G. & Jenkins, G. (1976). *Time series analysis: forecasting and control, rev. ed.*, Oakland, California: Holden-Day.

Brillinger, D. R. (1975). *Time series: Data analysis and theory*. New York: Holt, Rinehart. & Winston.

Butcher, D.F. & Muth, W.A. (1985). Predicting performance in an introductory computer science course. *Communications of the ACM, 28*(3), 263-268.

Campbell, F.P. & McCabe, G.P. (1984). Predicting the success of freshmen in a computer science major. *Communications of the ACM, 27*(11), 1108-1113.

Chatfield, C. (1993). Calculating interval forecasts, *Journal of Business and Economic Statistics*, 11: 121–135.

Chiang, C.L. (2003) *Statistical methods of analysis*. Singapore: World Scientific.

Cook, R.D. & Weisberg, S. (1982). Criticism and influence analysis in regression. *Sociological Methodology*, 13: 313-361.

Cressie, C. (1996) Change of support and the modiable areal unit problem. *Geographical Systems* 3:159–180.

Curwin, J. & Slater, R. (2002). *Quantitative methods for business decisions*, 5th edition. Singapore: Thomson Learning.

Draper, N.R. & Smith, H. (1998). *Applied regression analysis*. New York: Wiley Series in Probability and Statistics.

Durbin, J. & Koopman S.J. (2001). *Time series analysis by state space methods*. Oxford University Press.

Engle, R.F. & Granger, C.W J. (1987). Co-integration and error correction: Representation, estimation and testing. *Econometrica*, 55(2), 251-276.

Farrar, D.E. & Glauber, R.R. (1967). Multicollinearity in regression analysis: The problem revisited. *The Review of Economics and Statistics* 49(1):92-107.

Fisher, R.A. (1922). The goodness of fit of regression formulae, and the distribution of regression coefficients. *J. Royal Statist. Soc.* (Blackwell Publishing), 85(4): 597-612.

Fotheringham, A.S. & Wong, D.W.S. (1991,) The modifiable areal unit problem in multivariate statistical analysis. *Environment and Planning A*, 23(7): 1025-1044.

Fotheringham, A.S., Brunsdon, C. & Charlton, M. (2002). *Geographically weighted regression: the analysis of spatially varying relationships.* New York: Wiley.

Fox, J. (1997). *Applied regression analysis: Linear models and related methods.* London: Sage Publications.

Freedman, D.A. (2005). *Statistical models: Theory and practice*. Cambridge: Cambridge University Press.

Galton, F. (1989). Kinship and correlation (reprinted 1989)". *Statistical Science,* 4(2): 80–86.

Gardiner, C. (2004). *Handbook of stochastic methods: for Physics, Chemistry and the Natural Sciences, 3rd edition*. London: Springer.

Gershenfeld, N. (1999). *The nature of mathematical modeling*. Cambridge: Cambridge University Press.

Gershenfeld, N. (2000). *The nature of mathematical modeling*. Cambridge: Cambridge University Press.

Golding, P & McNamarah, S. (2005). *Predicting academic performance in the School of Computing and Information Technology.* Paper presented at the 35th ASEE/IEEE Frontiers in Education Conference, October 19 – 22, 2005, Indianapolis, IN., S2H16-S2H20. Retrieved June 13, 2011 from http://fie.engrng.pitt.edu/fie2005/papers /1195.pdf.

Good, P.I. & Hardin, J.W. (2009). *Common errors in statistics (and how to avoid them),* 3rd edition. Hoboken, New Jersey: Wiley.

Granger, C. (1981). Some properties of time series data and their use in econometric model specification. *Journal of Econometrics,* 16: 121-130.

Granger, C. & Newbold, P. (1974). Spurious regressions in econometrics. *Journal of Econometrics,* 2:111-120.

Gregory, A.W. & Hansen, B.E. (1996). Residual-based tests for cointegration in models with regime shifts. *Journal of Econometrics,* 70(1): 99-126.

Hamilton, J.D. (1994), *Time series analysis.* Princeton: Princeton University Press,

Hardle, W. (1990). *Applied nonparametric regression.* Cambridge: Cambridge University Press.

Heaton, H. (1896) *A method of solving quadratic equations*, *American Mathematical Monthly*, 3(10), 236–237.

Higham, N. (2002). *Accuracy and stability of numerical algorithms*, 2nd edition. SIAM

Lindley, D.V. (1987). Regression and correlation analysis. The New Palgrave: A dictionary of economics, 1st edition. (In Eatwell, J., Milgate, M. & Newman, P. (eds.). Palgrave Macmillan, The New Palgrave Dictionary of Economics Online. Palgrave Macmillan.

Kruck, S.E. & Lending, D. (2003). Predicting academic performance in an introductory college-level IS course. *Information Technology, Learning and Performance Journal*, *21*(2), 9-15, Fall.

Kutner, M.H., Nachtsheim, C.J. & Neter, J. (2004). *Applied linear regression models*, 4th edition. Boston: McGraw-Hill/Irwin.

Maharaj, M.S. & Gokal, H. (2006). *An investigation into the performance of first year, first entry students in information systems and technology in relation to their matriculation results.* 36th Annual Conference of the Southern African Computer Lecturers Association: Conference Proceedings, Irwin Brown, Cape Town (South Africa), 2006, pp.1-10.

Meade, N. & Islam, T. (1995). Prediction intervals for growth curve forecasts. *Journal of Forecasting*, 14: 413–430.

Mogull, R.G. (2004). *Second-semester applied statistics*. Kendall/Hunt Publishing Company.

Nikolić D, Muresan RC, Feng W. & Singer, W (2012). Scaled correlation analysis: a better way to compute a cross-correlogram. *European Journal of Neuroscience*, pp. 1–21,

O'Brien, R.M. (2007). A caution regarding rules of thumb for variance inflation factors, *Quality and Quantity,* 41(5): 673-690.

Ramcharan. R. (2006). *Regressions: Why are economists obsessed with them?* March 2006. Accessed 2012-06-03.

Rauchas, S., Rosman, B., Konidaris, G. & Sanders, I. (2006). Language performance at high school and success in first year computer science. Retrieved on June 13, 2011 from http://www.db.grinnell.edu/sigcse/sigcse2006/Program/View

Ravishankar, N. & Dey, D.K. (2002), A first course in linear model theory. Boca Raton: Chapman and Hall/CRC.

Rich, B. & Schmidt, P. (2004). *Schaum's outline of theory and problems of elementary algebra*. New York: McGraw-Hill.

Scott, A.J. (2012). Illusions in regression analysis. *International Journal of Forecasting*, (forthcoming).

Shasha, D. (2004). *High performance discovery in time series*, Berlin: Springer,

Shumway, R. H. (1988). *Applied statistical time series analysis*. Englewood Cliffs, NJ: Prentice Hall.

Strutz, T. (2010). *Data fitting and uncertainty (A practical introduction to weighted least squares and beyond)*. Vieweg+Teubner.

Tofallis, C. (2009). Least squares percentage regression. *Journal of Modern Applied Statistical Methods,* 7: 526–534.

Van den Poel, D. & Larivière, B. (2004). Attrition analysis for financial services using proportional hazard models. *European Journal of Operational Research*, 157(1): 196-217.

Yang-Jing, L. (2009). Human age estimation by metric learning for regression problems. *Proc. International Conference on Computer Analysis of Images and Patterns*: 74–82.

ANNEXURES

Annexure A: Original Pass Rates Data with Numbers of Interventions

School no.		2011	2010	2009	2008	2007
1	Pass %	77.38	90.63	89.78	71.35	79.8
	No of intervention	7	9	9	7	8
2	Pass %	59.86	78.44	59.64	80.33	75.2
	No of intervention	6	8	6	8	7
3	Pass %	43.90	37.71	53.85	63.08	84.1
	No of intervention	4	4	5	6	8
4	Pass %	64.81	43.75	35.5	29.75	25.83
	No of intervention	7	5	4	4	4
5	Pass %	57.81	58.65	47.11	34.15	50.55
	No of intervention	6	6	4	4	5
6	Pass %	70.08	69.7	59.9	77.4	59.27
	No of intervention	7	7	6	7	6
7	Pass %	86.11	100	86.94	45.16	37.1
	No of intervention	8	9	8	5	4
8	Pass %	70.51	87.29	31.71	35.48	54.5
	No of intervention	7	8	3	3	5
9	Pass %	47.14	72.06	79.45	67.7	15.5
	No of intervention	5	7	8	7	3
10	Pass %	90.48	87.5	100	90	36.36
	No of intervention	9	8	9	9	4
11	Pass %	75.51	93.02	60.63	39.63	48.24
	No of intervention	7	9	6	4	5
12	Pass %	74.32	76.64	45.86	68.4	44.87
	No of intervention	7	8	5	7	4
13	Pass %	74.04	70.97	64.79	55.26	60.5
	No of intervention	7	7	6	6	6
14	Pass %	70.93	73.26	63.46	84.3	55.37
	No of intervention	7	7	6	8	5
15	Pass %	100.00	92.32	26.67	66.67	76.5
	No of intervention	9	9	3	7	8
16	Pass %	82.93	89.47	65.57	74.36	65.6
	No of intervention	8	9	7	7	7
17	Pass %	61.02	67.35	85	66.61	71.43
	No of intervention	6	7	8	7	7
18	Pass %	60.38	75.61	100	71.4	90.38
	No of intervention	6	7	9	7	9
19	Pass %	75.38	88.89	62.38	34.64	70.5
	No of intervention	7	9	6	3	7

Annexure B: Regression Data

Observation number	Pass (%) rate (Y)	Number of interventions (X)
1	77.38	7
2	90.63	9
3	89.78	9
4	71.35	7
5	79.8	8
6	59.86	6
7	78.44	8
8	59.64	6
9	80.33	8
10	75.2	7
11	43.9	4
12	37.71	4
13	53.85	5
14	63.08	6
15	84.1	8
16	64.81	7
17	43.75	5
18	35.5	4
19	29.75	4
20	25.83	4
21	57.81	6
22	58.65	6
23	47.11	4
24	34.15	4
25	50.55	5
26	70.08	7
27	69.7	7
28	59.9	6
29	77.4	7
30	59.27	6
31	60.38	6
32	75.61	7
33	100	9
34	71.4	7
35	90.38	9
36	75.38	7
37	88.89	9
38	62.38	6
39	34.64	3
40	70.5	7

41	86.11	8
42	100	9
43	86.94	8
44	45.16	5
45	37.1	4
46	70.51	7
47	87.29	8
48	31.71	3
49	35.48	3
50	54.5	5
51	47.14	5
52	72.06	7
53	79.45	8
54	67.7	7
55	15.5	3
56	90.48	9
57	87.5	8
58	100	9
59	90	9
60	36.36	4
61	75.51	7
62	93.02	9
63	60.63	6
64	39.63	4
65	48.24	5
66	74.32	7
67	76.64	8
68	45.86	5
69	68.4	7
70	44.87	4
71	74.04	7
72	70.97	7
73	64.79	6
74	55.26	6
75	60.5	6
76	70.93	7
77	73.26	7
78	63.46	6
79	84.3	8
80	55.37	5
81	100	9
82	92.32	9
83	26.67	3
84	66.67	7

85	76.5	8
86	82.93	8
87	89.47	9
88	65.57	7
89	74.36	7
90	65.6	7
91	61.02	6
92	67.35	7
93	85	8
94	66.61	7
95	71.43	7

Annexure C: Multiple Time Series Data

School no.		2011	2010	2009	2008	2007
1	Pass %	77.38	90.63	89.78	71.35	79.8
	No of intervention	7	9	9	7	8
2	Pass %	59.86	78.44	59.64	80.33	75.2
	No of intervention	6	8	6	8	7
3	Pass %	43.90	37.71	53.85	63.08	84.1
	No of intervention	4	4	5	6	8
4	Pass %	64.81	43.75	35.5	29.75	25.83
	No of intervention	7	5	4	4	4
5	Pass %	57.81	58.65	47.11	34.15	50.55
	No of intervention	6	6	4	4	5
6	Pass %	70.08	69.7	59.9	77.4	59.27
	No of intervention	7	7	6	7	6
7	Pass %	86.11	100	86.94	45.16	37.1
	No of intervention	8	9	8	5	4
8	Pass %	70.51	87.29	31.71	35.48	54.5
	No of intervention	7	8	3	3	5
9	Pass %	47.14	72.06	79.45	67.7	15.5
	No of intervention	5	7	8	7	3
10	Pass %	90.48	87.5	100	90	36.36
	No of intervention	9	8	9	9	4
11	Pass %	75.51	93.02	60.63	39.63	48.24
	No of intervention	7	9	6	4	5
12	Pass %	74.32	76.64	45.86	68.4	44.87
	No of intervention	7	8	5	7	4
13	Pass %	74.04	70.97	64.79	55.26	60.5
	No of intervention	7	7	6	6	6
14	Pass %	70.93	73.26	63.46	84.3	55.37
	No of intervention	7	7	6	8	5
15	Pass %	100.00	92.32	26.67	66.67	76.5
	No of intervention	9	9	3	7	8
16	Pass %	82.93	89.47	65.57	74.36	65.6
	No of intervention	8	9	7	7	7
17	Pass %	61.02	67.35	85	66.61	71.43
	No of intervention	6	7	8	7	7
18	Pass %	60.38	75.61	100	71.4	90.38
	No of intervention	6	7	9	7	9
19	Pass %	75.38	88.89	62.38	34.64	70.5
	No of intervention	7	9	6	3	7