The Statistical Modeling of Retail Business Processes in Nyakach District

by

Omollo Thadius Oriema

A project report submitted in partial fulfilment
of the requirements for the degree of Master of Science in Applied Statistics

School of Mathematics, Statistics and Actuarial Science

MASENO UNIVERSITY

Abstract

Multivariate statistical analysis of retail business processes was done in this study. Multiple predictor variables were included in the regression model. Meta-analysis of the correlations were also used to explore heterogeneity and to estimate central tendency and variation in the effects. It was our observation that many retail businesses started in Nyakach District either don't do well or collapse within the first few years of their operation. Reasons for this trend were numerous and unsatisfactory besides being less quantitatively statistical.It was against this background that we were inspired to statistically model retail business processes in this region to investigate the effects of capital, business age and floor area among other factors, on business performance using regression and correlation analysis. We built models, vigorously analyzed and interpreted the same in an attempt to examine how value could be created and captured in a business unit. We have therefore provided a broad review of literature on business models in which we examined the business models concept through multidisciplinary and subject matter lenses.

Chapter 1

Introduction

1.1 background of the study

Models on regression analysis have been applied in many areas of study. Many applications involve situations where there are more than one regressor variable taking the form

$$Y = N_0 + N_1 x_1 + N_2 x_2 + ... + N_q x_q + \varepsilon,$$

with y being the predicted value,$x_i's$, the predictors,N_i's, the coefficients of predictors. ε takes care of independent variables affecting the model but not considered.Our predicted variable ,monthly profit is denoted by m_p .The predictors we considered are $x_1 = (f_a)$= business floor area in m^2, $x_2 = (a_g)$= business age in years/months , $x_3 = (c_a)$= capital invested in shillings , $x_4 = (m_s)$ =monthly sales in shillings, $x_5 = (t_r)$ = number of transactions in a month , $x_6 = (e_x)$ = monthly expenses in shillings.
The regression models are often used as approximating functions.

That is, the true relationship between m_p and $x_1, x_2.....x_q$ is unknown but over certain ranges of the independent variables the linear regression model is an adequate approximation. Multiple regression techniques could also be used to analyze cubic polynomial models in one regressor variable such as

$$m_p = N_0 + N_1 x + N_2 x^2 + N_3 x^3 + \varepsilon,$$

and even models with interaction effects such as

$$m_p = N_0 + N_1 x_1 + N_2 x_2 + N_{12} x_1 x_2 + \varepsilon.$$

In this study we focussed on retail business processes in Nyakach District in Kisumu County, by examining statistically the effects of capital, business age and floor area among other factors on business performance in this region using a model based on regression and correlation analysis. It was in the early 1970's when the word business models appeared for the first time as an economic keyword in Public talk . Ghaziani and Ventresca show that the categories of measuring used in business model discussions vary significantly with respect to the considered time period and the related community of discourse [19]. In the recent years the business model has been the focus of substantial attention both academic and practitioners. Since 1995 there have been 1177 papers published in peer reviewed academic journals in which the motion of business model is addressed [6]. Business models have also been the subject of a growing number of practitioner oriented studies. While there has been explosion in the number of papers published and an abundance of conference sessions and panels on the subject of business models, it appears that researchers and practitioners have yet to develop a common and widely accepted model, statistical or conceptual, that would allow researchers who examine the business model construct through different lenses to draw effectively on each others work. We have attempted in this study to build models and subject the same to vigorous analysis and interpretation scales for easy applicability. The majority of the study designs and statist-

ical analysis used these days control for the effects of multiple variables in the model and so most meta-analysis techniques focus on the synthesis of bivariate relationships. This is in part because methods to synthesize multivariate analysis including multiple regression models are not yet well understood. In addition several complications arise due to the characteristics of multiple regression analysis. For instance different models include different members of predictors and the predictors of interest are measured with different instruments and scales across the primary studies. As we are aware, statistics has proved to be an invaluable tool in our daily lives and many decisions, conclusions, predictions and forecasts by experts or laymen are usually based on some statistical data. Moreover nowadays virtually every area of serious scientific inquiry must be subjected to statistical analysis for validation. We want to say that the efforts of entrepreneurs at quality control, turnover maximization, cost minimization, product and inventory mix and many other business matters and processes can be effectively managed by using proven statistical procedures and models. It is our attempt in this study to equip retail business persons with quantitative and qualitative skills so as to make effective use of statistical data in the business work place to develop expertise in a standard set of statistical and graphical techniques which will be useful in analyzing data, and to learn to apply these techniques or models with respect to capital, business age and floor area. The use of regression analysis here is to enable the businessman estimate the nature or form of relationship between the variables / factors constantly affecting the business whereas, correlation analysis gives him an insight into the strength or degree of the relationship.

1.2 Statement Of the Problem

This study was done to assess the effects of capital, floor area and age of business among others on the overall business performance using models based on regression and correlation analysis. Prediction and decision making in business is clouded with subjectivity and gut feelings. In this study we built models and analyzed so as to replace the problem

of subjectivity in prediction and decision making with objectivity that relies on the available measurable data. This could help eliminate unforeseen risks that lead to collapse of businesses. We did this by using the information received from stratified sampling survey of retail business persons situated in the twenty five market centers in Nyakach District.

1.3 Objectives of the study

The objectives of the study were to:

(i)determine the effects of business age, floor area and capital among others on retail business performance.

(ii) use statistical data for creating models to assist retail business managers make decision.

(iii)build and analyze models in search of improvement in retail business performance in Nyakach using regression and correlation analysis.

(iv)use statistical and graphical techniques to present the analyzed data.

1.4 Significance of the Study

Research findings are highly relied upon by the government and serious individuals in various fields of profession. Therefore data collection and statistical analysis evident in this study could help improve business decision and reduce risk of implementing solutions that waste resources and effort in any area of business endeavor. The business community could as well be exposed to the extent of recognizing, developing and distinguishing between models for cross-sectional analysis at a single point in time and models for analysis at multiple points in time. The study was crucial as it aimed at increasing the capabilities of

businessmen to think statistically using data and use this capability to support business institution. Finally, the study was expected to help in building sufficient skills to provide leadership in statistical methods for staff in area of responsibility or specialty in the business enterprise.

1.5 Basic Concepts

(a)**Regression**

This term refers to the statistical methodology for predicting values of one or more response (dependent) variables, m_{pi}'s from a collection of predictor (independent) variables, x_i's. It can be used to assess the effects of predictor variables on the response.

The equation says that for a given value of the variable $X = x$, the actual value of m_p, is determined by the expression $N_0 + N_1 x$, plus some random variation ε, caused by other unmeasured factors. Thus if we know the values of N_0, the true population intercept, and N_1, the true population slope we can predict the value of m_p to within some random error ε. Incase of linear regression the relationship between x and m_p is represented by a straight line ie as x changes, m_p changes by a constant amount.

(b)**Regression Model**

It states that m_p is composed of mean, which depends in a continuous manner on the x_i's and the random variable ε which accounts for measurement error and the effects of other variables not explicitly considered in the model . In bivariate regression or simple regression, m_p is said to be a function of only one independent variable. It is described by;

$$m_p = N_0 + N_1 x + \varepsilon.$$

The equation says that for a given value of the variable $X = x$, the actual value

of m_p is determined by the expression $N_0 + N_1 x$, plus some random variation ε, caused by other unmeasured factors. Thus if we know the values of N_0, the true population intercept, and N_1, the true population slope we can predict the value of m_p to within some random error ε. In multiple regression model, m_p is a function of two or more independent variables i.e.

$$m_p = N_0 + N_1 x_1 + N_2 x_2 + ...N_q x_q + \varepsilon.$$

That is, the model states that the actual value of the variable m_p is determined by the equation $N_0 + N_1 x_1 + ... + N_q x_q$, plus some random variation, ε, caused by the other unmeasured factors.The coefficients $N_1, N_2.....N_k$ are similar to the slope coefficient N_1 in the univariate model, $mp = N_0 + N_1 x_1 + \varepsilon$, with a small difference. In the univariate model, N_0 represents the slope of the line, or the change in the dependent variable,m_p, for a unit change in the independent variable, X. In the multiple regression model, the parameters $N_1, N_2.....N_q$ are really not the slope because we are not talking about a line, but they have a similar interpretation. Each of the N_i coefficients represents the change in the dependent variable,m_p,if the variable associated with the coefficient of interest, X_i, is changed by one unit and all other variables in the model are held constant. Multiple regression equation describes how the mean value of m_p is related to x_1, x_2, x_q. Since we assume $E(\varepsilon) =$ 0,hence

$$E(m_p) = N_O + N_1 X_1 + N_2 X_2 + ... + N_q X_q.$$

Regression equation estimated from sample data is called Estimated regression equation

$$\hat{m}_p = n_0 + n_1 x_1 + n_2 x_2 + ... + n_q x_q,$$

where

$$n_o, n_1, n_1, ..n_6$$

are estimates of

$$N_O, N_1, N_2..N_6,$$

and q=6

(c)**Subjective /Intuitive Models**

These are the most abstract models that are characterized by inadequate formulation, limited data and insufficient testing. These involve the use of Gauntt charts,Flow charts,Functional Flow diagrams and Pert diagrams.

(d)**Business process**

Is a collection of related, structured activities or tasks that produce specific service or product for a particular customer or customers. In other words it can be viewed as a sequence of activities with interleaving decision points or with a process matrix as a sequence of activities with relevance rules based on the data in the process.There are various business processes but we looked at operational processes that constitute the core business and creates the primary value stream such as purchasing, marketing, sales and advertisement.

Chapter 2

Review of Related Literature

2.1 Introduction

Models on multivariate analysis have been witnessed in many fields including science and engineering. Regression is noted as a powerful tool in multivariate statistical analysis for understanding the relationship between dependent variable m_p and independent variable x. A commonly occurring situation is one where a random quantity m_p is a function of one or more independent variables $x_1, x_2.....x_6$ [1]. However many study designs and analyses control for the effects of multiple variables hence the need for good understanding of these multivariate techniques of which regression is one. Background knowledge on models reveal that they vary in their level of formality, explicitness, richness in details and relevance. They may have several functions including explaining phenomena, making predictions, making decisions and communicating knowledge among others [11].

Mathematical models have been in use for hundreds of years. For example Thales of Miletus at about 600 BC used a model to predict solar-eclipse of 585BC. He also devised a method of measuring heights by measuring lengths of shadows. In 250 BC Eratosthenes of Cyrene calculated the distances Earth-sun, Earth-moon and the circumference of the earth using mathematical geometric model. Diophantus of Alexandria in about 250 AD in his book Arithmetica, developed the beginning of algebra based on symbolism and notion of variable [4]. For Astronomy, Ptolemy, inspired by Pythagoras idea in 150 AD

developed a mathematical model of the solar system with circles and semicircles to predict the movement of the sun, moon and the planets. This model which was made simpler by Kepler in 1619 and later refined by Newton and Einstein is still valid today. In 1940's models of previously unknown sizes became tractable and it was possible to use mathematical modeling for solving practical problems of significant size. This has been boosted by improved technology in the early 1990's. Mathematical modeling has also been increasingly attractive in military industry and optimization problems. By 1940's there were models developed by Neumann and Charney in meteorology which enabled daily numerical prediction and forecasting [12].

The introduction of variables, function spaces and of all the mathematical structural theory has made mathematical models increasingly formal consisting of concepts such as variables, relations and data. Systematic use of variables was invented by Vieta (1540 - 1603), 300 years later Cantor and Russel clarified the true role of variables in the formulation of mathematical theory. Physics and the description of the nature's principles became the major driving force in modeling. Later economics joined in and now an ever increasing number of applications demand models and their analysis.

Our concern in this study is generally business modeling but more specifically statistical business modeling using regression and correlation analysis. In the last few decades studies involving multivariate techniques has increased notably making simpler univariate approaches to meta-analysis more difficult to apply and justify [7]. Consequently researchers conducting quantitative reviews face the alternatives of omitting a large number of multivariate primary studies or attempting to synthesize those multivariate studies and in some cases combining their results with ones obtained from univariate designs. However when dealing with effects from multivariate designs such as multiple regression models there is no universal approach on which meta-analysts agree. Thus new indices of effect magnitude and methods of synthesizing them are called for. Following is a review of business models and in particular statistical business models.

2.2 Business models

Our concern is generally business models and specifically building statistical business models based on regression and correlation analysis. Business may be defined as the art of getting things done presumably pursuing one's profit or the general interest of the society [4]. Business model is a framework for creating economic, social and other forms of value; it represents the core aspect of a business including purpose, strategies, organization structures, trading practices, operation, policies and practices . It is therefore a method of doing business by which an enterprise can sustain itself. It shows business data, organization and processes. A number of models are in existence ranging from conceptual to statistical models. Our main concern will be statistical models however lets have a glance at conceptual models. It was in the early 1970's that the word business model appeared for the first time as an economic keyword in public talk [19] Ghaziani and Ventresca say here that the major shift in the frequency of use of specific business model frames is accompanied with the advent of the New Economy in the mid 1990's. At the same time the business model concept increasingly gained importance on the research agenda of business and management science scholars. There is a wide spread acknowledgement implicit and explicit that a business model is a new unit of analysis in addition to the product, firm or industry or network levels, it is centered on a focal organization but its boundaries are wider than those of the organization. The business model constitutes a holistic concept consideration of all elements that build the anatomy of a firm's core logic for creating and appropriating value. There are a number of models classified as simplistic, complex and simple models [1]. These are not based on numerical data and measurable state but on personal opinion, experience and belief. The bait and the hook model was in existence in the early part of 20th century. It involves offering one product at a low price (bait) and another at a high price. These models were built using flow charts, functional flow block diagrams, Gantt Chart, Programme Evaluation and Review Technique diagram and Integration Definition modeling language [15]. These techniques

emerged in the beginning of 12th Century. Gantt Charts were the first to arrive around 1899, the flow charts in the 1970's, Functional Block Diagram and Programme Evaluation Review Techniques in the 1950's. Data flow Diagram and Integration Definition in the 1970's. Among the modern methods are unified modeling language and business process modeling notation. Still these represent just a fraction of the methodologies used over the years to document business processes. The term business modeling was coined in 1960's in the field of systems engineering by S. Williams. He stated in 1967 that business process modeling improves administrative control. His idea was that techniques for obtaining a better understanding of physical control systems could be used in a similar way for business as processes. It took until the 1990's before the term became popular. In 1990's the term process became a new productivity paradigm. Businesses were encouraged to think in processes instead of functions and procedures. Process thinking looks at the chain of events in the business from purchases to supply, from order retrieval to sales e.t.c. The first business modeling involving data collection, data flow analysis, process flow diagrams and reporting facilities were present in 1995. This dimension was a migration from the above mentioned subjective conceptual models to the more objective statistical business models.

2.3 Statistical business models

Statistics is the Science of mass phenomena under the condition of uncertainty [4]. The original idea of statistics was the collection of information about and for state. The word statistics derives directly from the Italian word for state. The earliest writing on statistics was found in the 9th Century book written by Al-Kindi (801-873 CE). He gave a detailed description analysis of how to use statistics and frequency analysis to decipher encrypted message. This was the birth of statistics and cryptanalysis. Some scholars pinpoint the origin of statistics to 1663 with the publication of natural and political observations upon the bills of mortality by John Gantt, a native of London. He analyzed this document

for the state using descriptive statistics. He showed that the optimum values of both the regression slope and correlation coefficient for a straight line could be evaluated from the product moment function [16] .He attributed the formula to August Bravai's work fifty years earlier. He noted Bravai did not demonstrate the use of product moment for calculating correlation coefficient but he never showed that it provided the best for the data. Mathematics foundations of statistics were laid in the 17th Century with the development of probability theory by Blaise Pascal and Pierre de Fermat . The method of least squares was first described by Carl Fredrich Gause in around 1794. The scope of this discipline has broadened in the early 19th century to include the collection and analysis of data in general. Today statistics is widely used in government business, natural science, Social Science, and virtually every area of serious scientific inquiry that must be subjected to statistical analysis for validation. This is the backbone of this study. Statistical modeling relies heavily on regression. Regression analysis and its multiple variations are by far the most widely utilized modeling approach. It is a strategic tool utilized by many of the world's top corporations for marketing mix models, data mining and volume forecasting based on some set of parameters and initial conditions [4].Sir Francis Galton (1892 - 1911) is thought to be the father of regression analysis after his experiment on sweet potatoes and hereditary patterns in heights of adult humans[2]. Later regression analysis exploded into one of the most powerful statistical tools at our disposal. The rigorous treatment of correlation and regression analysis was the work of Pearson in 1896. In the 1950's statistical methods prevailed in many areas of science for understanding disorderly phenomena. Statistical business modeling with managerial implications such as, 'what if' analysis rely on regression analysis, powerful techniques for studying relationships between dependent variables (i.e. output, performance measure and independent variables (i.e in put factors, decision variables). Summarizing relationships among the variables by the most appropriate equation. (i.e. model) allows us to predict or identify the most influential factors and study their impacts on the output for any changes in their current values. Most of the available literature on business models are on two extreme ends. One end purely de-

terministic models with qualitative variables with little or no mathematics [6] whereas the other end consists of serious mathematical/ statistical models with quantitative variables which are only good for a specialized group of users [19] .

In this study we intend to concentrate much on the latter category and then finally illustrate how to deal with the mixture of the two (qualitative and quantitative variables) in our model. Authors,[1, 13, 15] all approach regression analysis with much vigor and details but fail to interpret constantly each analyzed aspect of regression to end users or practitioners who could be engineers, meteorologists, medics etc. In this study we will be interpreting the meaning of each observation made so that the user, the business person in particular can make decision appropriately. We therefore intend to show that, "Applied statistics" is really applicable in business world, among other fields. Most of the work on modeling using regression and correlation analysis have the results mainly presented either analytically or tabular form [7]. In addition to these we in this project intend to involve a lot of graphs for easier interpretation by those with no in-depth knowledge in statistics. Further it is true that all authors look at regression analysis separately in two parts. First, simple linear regression then multiple regression. We also wish to progress likewise but continuously in one document for easy comparability and decision making on which model is better. Again fore mentioned authors do subject only one model in each case to the many analysis instruments. In this review of academic literature we have attempted to explore the origin of the construct and to examine the business model concept through multidisciplinary and subject matter lenses. This review revealed the following;

Despite the overall surge in the literature on business models scholars do not agree on what a business model is. Researchers tend to adopt idiosyncratic definitions that are difficult to reconcile with each other [6]. There is wide spread acknowledgement implicit and explicit that business model is a new unit of analysis in addition to the product, firm, industry or network levels, it is centered on a focal organization, but its boundaries are wider than those of the organization. Business models emphasize a system level hol-

istic approach towards explaining how firms do business. Organizational activities play an important role in the various conceptualizations of business models that have been proposed. Business models seek to explain both value creation and value capture [6].

In this study our intended contribution was to provide and document most comprehensive and up to date literature on models. Review on statistical business models based on regression and correlation analysis followed before such models were built and analyzed.Summary,conclusion and recommendations were made in an attempt to bridge the seemingly wide gaps between the conceptual and statistical business models.

Chapter 3

Research Methodology

3.1 Research design,Data and Sampling

This study was conducted in Nyakach District .Nyakach District is in Kisumu County, bordering Kericho and Homabay Counties. It is a designated hardship area based on insecurity and unfavourable economic environment. Scarcity of resources among others have therefore impacted negatively on businesses in in the area.

Data used was basically primary data collected from retailers who were interviewed from twenty five trading centers randomly selected.This comprised of over ninety percent of the trading centers.From Nyando county council records, retail population in this area stands at about six thousand,including those without shops and annual trading licenses. Only retailers with shops and trading licenses who had been in business for not less than five months were considered in this study. We discovered this was about ten percent of the retail traders population. Stratified random sampling with proportional allocation was used with the population divided into twenty five strata,each market center forming a stratum. Stratification allowed the heterogeneous retail population to be divided into sub-population, called strata each of which was a market that was internally homogenous. The written questionnaire (see appendix) was structured to capture information relating to the retailer's average monthly profit, capital employed in the business, business floor area, expenses, average number of transactions in a day, business age in months,average

monthly sales, education level, source of his capital, distance from main road and wholesaler, sex/age of a majority of his customers and the number of his employees among others.For semi illiterate retailers we conducted oral interview.These were not many but it appeared some were shy to say so.Nyando County Council offices helped us with the estimated number of retailers with trading licenses and the official names of the market centers .

Overall sample size

Most of the available literature recommend that if the target population is less than ten thousand then the required sample size would be smaller.Since our population was about 6000 we adopted the formula where the estimate of sample size is obtained as follows:

$$S = \frac{c}{1 + \frac{c}{\text{population}}} \tag{3.1.1}$$

where

$$c = \frac{z^2[p(1-p)]}{D^2}$$

- P is true proportion of factor in the population or the expected frequency value of retailers with shops.
- D is maximum difference between the sample mean and the population mean or the expected frequency value minus worst accepted value.
- Z is the area under normal curve corresponding to the confidence level of 95%.

With

$$Z = 1.96, \quad P = 10\% \quad D = 0.05, \quad c = \frac{(1.96)^2[0.1(0.9)]}{0.05^2} = 138.2976$$

From equation 3.1.1

$$S = \frac{138.2976}{1 + \frac{138.2976}{6248}} = 135.302$$

But Sample size per market, k_h , depended on the market weight $\frac{K_H}{6248}$ as $\sum K_H = 6248$ and not K so that $k_h = [\frac{K_H}{6248}]S$. $\sum k_h = k$, the overall sample size.These were calculated as in appendix no.2 ,slightly altering sample size to 148.

This was a fairly good sample considering our target population of 630.We saw this necessary so that the discrepancy (sampling error) between the sample characteristics and population characteristics could be reduced.We distributed out 180 questionnaires so that we could meet our target of 148 should some respondents fail to fill theirs.Out of 180 questionnaires given to 180 retailers only 151 were returned filled properly.From these we randomly eliminated 3 to leave us our goal of 148.

3.2 Modelling process

Data was imported into Stata version 12.1 (Stata Corp.,College Station, Texas, USA) for analysis from Excel spreadsheet used for data entry. Descriptive statistics was done using frequencies and respective proportions and means with corresponding standard deviations or medians and respective inter-quartile ranges after assessing for normality of the particular covariates. Pairwise Pearson Correlation analysis of each predictor covariate versus the monthly profit was done reporting the correlation coefficient r and corresponding p values. Linear regression assumptions were assessed for using appropriate graphs and tests. We also checked for influential variables for exclusion. Univariate analysis was done using linear regression analysis and all covariates were included apriori into the multivariable linear regression model. Finally a stepwise regression model with a probability of inclusion important for removal from model set at 0.02 was used to determine the single most important factors influencing monthly profit. Regression coefficients, respective 95% confidence intervals and p values were reported for each of the covariates fitted in the model. Adjusted R^2 values were also reported to assess the amount of variance

accounted for by the covariates and the model as a whole for the multivariable model. The F value and its significance was also reported for the multivariable model. Results were presented in the form of tables, box plots and scatter plots with fitted regression lines as deemed appropriate. We used model building techniques to identify the best subset from the set of the six independent variables. We used forward selection, backward elimination, and stepwise regression techniques to select the best model based on largest R^2, smallest MSE, largest F statistic, and largest t statistic.

We had $x_1, x_2 ..... x_q$ as our q predictor variables related to the variable monthly profit,m_p where $q = 6$, m_p equals monthly profit in shillings and a sample size of $k = 148$. we considered the linear relationship between the dependent or response variables m_{pi} and one or more of the predictor variables then used a linear model to relate the m_p to the x's and were concerned with estimation and testing of the parameters in the model. One aspect of the study was choosing which variables to include in the model . We distinguished two cases according to the number of variables:

- Simple linear regression:
  one m_p and one x. For example, suppose we wish to predict m_p based only on the number of transactions made in the month.

- Multiple linear regression(univariate multiple regression):
  one m_p and several predictor variables $(x's)$.We could attempt to improve our prediction of m_p by using more than one independent variable,for example,capital invested,level of advertisement,sales or expenses incurred.

The second aspect of regression is not the same as the Multivariate multiple linear regression where there are several $m_{p's}$ and several $x_i's$. The classical linear regression model states that m_p is composed of a mean, N_0 which depends in a continuous manner on the independent variables x's and a random error ε.

$$m_p = N_O + N_1 f_a + N_2 a_g + N_3 c_a + N_4 m_s + N_5 t_r + N_6 e_x + \varepsilon.$$

With $k = 148$ independent observations on m_{pi} and the associated values of the independent variables, the complete model in matrix form was as shown below.

$$\begin{pmatrix} m_p\ 1 \\ m_p\ 2 \\ \vdots \\ m_p\ \mathrm{k} \end{pmatrix} =$$

$$\begin{pmatrix} \mathrm{N}_0 & +\mathrm{N}_1 f_{a11} + \ldots & \ldots & +\mathrm{N}_5 t_{r15} & +\mathrm{N}_6 e_{x16} + \varepsilon_1 \\ \mathrm{N}_0 & +\mathrm{N}_1 f_{a21} + \ldots & \ldots & +\mathrm{N}_5 t_{r25} & +\mathrm{N}_6 e_{x26} + \varepsilon_2 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \mathrm{N}_0 & +\mathrm{N}_1 f_{ak1} + \ldots & \ldots & +\mathrm{N}_5 t_{rk5} & +\mathrm{N}_6 e_{xk6} + \varepsilon_k \end{pmatrix} \tag{3.2.1}$$

that is,

$$\begin{pmatrix} m_{p1} \\ m_{p2} \\ \vdots \\ m_{pk} \end{pmatrix} = \begin{pmatrix} 1 & f_{a11} & a_{g12} & \ldots & e_{x16} \\ 1 & f_{a21} & a_{g22} & \ldots & e_{x26} \\ \vdots & & \vdots & & \\ 1 & f_{ak1} & a_{gk2} & \ldots & e_{xk6} \end{pmatrix} \begin{pmatrix} N_1 \\ N_2 \\ \vdots \\ N_6 \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_k \end{pmatrix} \tag{3.2.2}$$

Where the error terms $\varepsilon_{i's}$ were assumed to have the following properties:

- The error term ε is a random variable with the mean or expected value of zero;that is $E(\varepsilon_j) = 0$. This implies that the model is linear and that no additional terms are needed to predict m_p, all the remaining variation in m_p is purely random and unpredictable. Thus if $E(\varepsilon_i) = 0$,then

$$E(m_{pi}) = N_0 + N_1 f_{ai1} + N_2 a_{gi2} + \ldots\ldots + N_6 e_{xi6}.$$

and the mean of m_p is expressible in terms of these 6 predictor variables with no others needed.

- Var $(\varepsilon_i) = \sigma^2$(constant.)

  The variance of each (ε_i) is the same for all values of predictors $f_a, a_g, c_a, m_s, t_r, e_x$ which also implies that $var(m_{pi}) = \sigma^2$ and is also the same for all values of predictors $f_a, a_g, c_a, m_s, t_r, e_x$.

- Cov $(\varepsilon_i.\varepsilon_j) = 0, i \neq j$. The error terms are uncorrelated, from which it follows that $m_{p's}$ are also uncorrelated that is $\text{cov}(m_{pi}m_{pj}) = 0$. The value of ε for a particular set of values for the independent variables is not related to the value of ε for any other set of values.

- The error term is a normally distributed random variable reflecting the deviation between the m_p value and the expected value of m_p is given by;

$$N_0 + N_1 f_a + N_2 a_g + N_3 c_a + N_4 m_s + N_5 t_r + N_6 e_x.$$

  Implication:because $N_0, N_1 ..... N_6$ are constants for the given values of $f_a, a_g, ...e_x$, the dependent-variable m_p is also normally distributed random variable.

Thus we restate the three assumptions in terms of m_p as follows:

- $E(m_{pi}) = N_0 + N_1 f_{ai1} + N_2 a_{gi2} + ... + N_6 e_{xi6} \quad i = 1.....148$
- $\text{Var}(m_{pi}) = \sigma^2 \quad i = 1.....148$
- $\text{cov}(m_{pi}m_{pj}) = 0 \quad i \neq j$

3.3 Estimation Process

This is a procedure for using sample data to find the estimated regression equation, the technique that finds the equation of the line that minimizes the total or sum of the squared deviations between the actual data points and the line. Our concern was to develop an equation that would predict the response, m_p for the given values of the predictor variables $f_a, a_g, c_a, m_s, t_r, e_x$.

The line should be close to as many of the data points as possible. The distance from each data point to the line is called the deviation or errors of the line. We find a line that minimizes the overall deviation of the data points from the line,and the least squares criterion is:

$$Min \sum (m_{pi} - \hat{m}_{pi})^2.$$

Where m_{pi} is the observed value of the monthly profit for the i^{th}retailer. $\hat{m}_{pi}$ is the estimated value of the monthly profit for the i^{th} retailer

In bivariate situation ,we determine the quantities n_0 and n_1 such that the deviation

$$D = \sum (m_{pi} - n_0 - n_1 x_i)^2$$

is minimized. $X_{i's}$ are predictor variables:

$$m_{pi} - n_0 - n_1 x_i = (m_{pi} - \bar{m}_p) - n_1(x_i - \bar{x}) + (\bar{m}_p - n_0 - n_1 \bar{x})$$

Squaring both sides we obtain;

$$\begin{aligned}(m_{pi} - n_0 - n_1 x_i)^2 &= (m_{pi} - \bar{m}_p)^2 + n_1^2(x_i - \bar{x})^2 + (\bar{m}_p - n_0 - n_1\bar{x}_i)^2 \\ &\quad -2n_1(x_i - \bar{x})(m_{pi} - \bar{m}_p) - 2n_1(x_i - \bar{x}) \\ &\quad (\bar{m}_p - n_0 - n_1\bar{x}) + 2(m_{pi} - \bar{m}_p) \\ &\quad (\bar{m}_p - n_0 - n_1\bar{x})\end{aligned}$$

We now sum both sides over $i = 1.....148$ and note that the last two terms on the right hand side of the formula disappear after summation, because $\sum(x_i - \bar{x}) = 0$

and $\sum(m_{pi} - \bar{m}_p) = 0$ hence we have

$$D = S_{mp}^2 + n_1^2 S_x^2 + k(\bar{m}_p - n_0 - n_1\bar{x})^2 - 2n_1 S_{xmp}.$$

We can now arrange the terms and complete square with;

$$\begin{aligned}D &= 148(\bar{m}_p - n_0 - n_1\bar{x})^2 + n_1^2 s_x^2 - 2n_1 s_{xmp} + s_{mp}^2 \\ &= 148(\bar{m}_p - n_0 - n_1\bar{x})^2 + (n_1^2 s_x^2 - 2n_1 s_{xmp} + \\ &\quad \frac{s_{xmp}^2}{s_x^2}) + s_{mp}^2 - \frac{s_{xmp}^2}{s_x^2} \\ &= 148(\bar{m}_p - n_0 - n_1\bar{x})^2 + (n_1 s_x - \frac{s_{xmp}}{s_x})^2 \\ &\quad +(s_{mp}^2 - \frac{s_{xmp}^2}{s_x^2})\end{aligned}$$

The last term does not involve n_0 and n_1. The first two terms can be reduced to the smallest value of zero if we set

$$n_1 = \sum(x_i - \bar{x})(m_{pi} - \bar{m}_p) = \frac{s_{xmp}^2}{s_x^2}$$

and $n_0 = \bar{m}_p - n_1\bar{x}$

We note Properties of these estimators :

(a)The estimators are unbiased : that is,

$$E(n_0) = N_0 \quad \text{and} \quad E(n_1) = N_1.$$

(b)

$$Var(n_0) = \delta^2(\frac{1}{148} + \frac{\bar{x}^2}{s_x^2}) \quad \text{and} \quad var(n_1) = \frac{\delta^2}{s_x^2}.$$

(c) The distribution of n_0 and n_1 are normal with means of N_0 and N_1, respectively ; the standard deviations are the square roots of the variance given in(b).

(d) $s^2 = \frac{SSE}{k-2} = \frac{SSE}{146}$ is an unbiased estimator of σ^2. Also $\frac{(146)s^2}{\sigma^2}$ is . distributed as X^2 with $df = 146$ and it is independent of n_0 and n_1.

(e)Replacing σ^2 in (b) with its sample estimate s^2 and considering the square root of the variances, we obtain the estimated standard error of n_0 and n_1;

Estimated Standard Error of

$$n_0 = s\sqrt{(\frac{1}{148} + \frac{\bar{x}^2}{s_x^2})};$$

Estimated standard error of $n_1 = \frac{s}{s_x}$

(f)

$$\frac{s_x(n_1 - N_1)}{s}$$

has t distribution with df $= 146$

$$\frac{(n_0 - N_0)}{s\sqrt{(\frac{1}{148} + \frac{\bar{x}^2}{s_x^2})}}$$

has t distribution with df= 146

In the multivariate situation , we have

$$E(m_{pi}) = N_0 + N_1 x_{i1} + N_2 x_{i2} + ... + N_6 x_{i6},$$

given that $E(\varepsilon_j) = 0$. We seek to estimate the $N's$ and thereby estimate $E(m_{pi})$. If the estimates are denoted by $\hat{N}_0, \hat{N}_1 .....\hat{N}_6$ then

$$E(m_{pi}) = \hat{N}_0 + \hat{N}_1 x_{i1} + \hat{N}_2 x_{i2} + ... + \hat{N}_6 x_{i6}.$$

However, $\hat{E}(m_{pi})$is usually designated $\hat{m}_{pi}$. Thus $\hat{m}_{pi}$ estimates $E(m_{pi})$, not m_{pi}. We now consider the estimates of the $N's$. The least square estimates of $\hat{N}_0, \hat{N}_1 .....\hat{N}_6$ minimize the sum of squares of deviations of the 148 observed $m_{p's}$ from their modeled values, that is, from their values m_{pi} predicted by the model. Thus we seek $\hat{N}_0, \hat{N}_1 .....\hat{N}_6$ that minimize

$$\begin{aligned} SSE &= \sum_{i=1}^{148} \hat{\varepsilon}^2 = \sum (m_{pi} - \hat{m}_{pi})^2 \\ &= (m_{pi} - N_0 - N_1 x_{i1} - N_2 x_{i2} - ..... - N_6 x_{i6})^2. \end{aligned} \tag{3.3.1}$$

The value of $\hat{\mathbf{N}} = (\hat{N}_0, \hat{N}_1 .....\hat{N}_6)'$ that minimizes SSE in[3.3.1] was provided by a computer software.

3.4 selection of predictors

Model building techniques were used to identify the best multiple regression model from a set of independent variables based on criteria such as largest R^2,smallest MSE, largest F statistic, and largest t statistic.

There were more predictors than were needed for predicting m_p. Some of them were redundant and could be discarded. In addition to logistical motivations for deleting x variables, there are statistical incentives; for example, if an x is deleted from the fitted model, the variances of the N_j's and of the m_{pi} 's are reduced. The approaches to subset selection are: all possible subsets, stepwise selection,forward selection,and backward elimination.

In this study we mainly used the stepwise method particularly to come up with the reduced model.

3.4.1 All Possible Subsets

Given a set of candidate variables, regression models for all possible combinations of variables are run and then identification of the best model for each different size. For example, if five variables are under consideration for a model, the all subsets method runs all models with one variable, all possible combinations of two variables, all three-variable combinations, and so on. It then chooses, according to one or several criteria, the best one-variable the model, the best two-variable model, and so on. The user is then free to choose among the models presented. It should be noted that the number of possible combinations grows with respect to the number of variables.Since we had 6 variables under consideration, then there were $2^6 - 1$ combinations/subsets to consider.

All possible subsets of the x's are examined.This may not be computationally feasible if the sample size and number of variables are large. However computer program we used found the optimum sub set without examining all the subsets. The number of variables in a subset is denoted by $p - 1$, so that with the inclusion of an intercept, there are p parameters in the model. The corresponding total number of available variables from which a subset is to be selected is denoted by $q(= 6)$, with q+1(=7) parameters in the model. Our choice of the best model was guided by the following scales of measurements:

R_p^2 : By its definition, the proportion of total sum of squares accounted for by regression,

is a measure of model fit. The subscript p is an index of the subset size, since it indicates the number of parameters in the model, including an intercept. However, R_p^2 does not reach a maximum for any value of p less than 6 because it cannot decrease when a variable is added to the model. We find the subset with largest R_p^2 for each of $p = 2, 3, ..., t$ and then choose a value of p beyond which the increases in R^2 appear to be unimportant.

s^2p : Another criterion is the variance estimator for each subset.

$$s_p^2 = \frac{SSE_p}{148 - p} \tag{3.4.1}$$

In each of the p = 2, 3, . . . , t, we find the subset with smallest s_p^2. If t is large,a typical pattern as p approaches t is for the minimal s_p^2 to decrease to an overall minimum less than s_t^2 and then increase. The minimum value of s_p^2 can be less than s_t^2 if the decrease in SSE_p with an additional variable does not offset the loss of a degree of freedom in the denominator. We choose the subset with absolute minimum s_p^2. This procedure may fit some noise unique to the sample and thereby include one or more extraneous predictor variables.Alternatively we choose p such that $min_p\ s_p^2 = s_t^2$ or, choose the smallest value of p such that $min_p\ s_p^2 < s_t^2$ since there will not be a $p < 7$ such that $min_p\ s^2p$ is exactly equal to s_t^2.

The C_p criterion: This statistic is a measure of the total squared error.The quantity p is the number of terms in the model including the intercept N_o. The C_p statistic uses the SSE from the model being evaluated and the MSE from the full model with all the potential variables.When we use C_p the best model with p terms result in C_p=p. For this model the estimates of the coefficients are unbiased. The expected squared error,

$$E[\hat{m}_{pi} - E(m_{pi})]^2$$

is used in formulating the C_p criterion because it incorporates a variance component and a bias component. The goal is to find a model that achieves balance between the bias and variance of the fitted values $\hat{m}_{pi}$. Bias arises when the $\hat{m}_{pi}$ values are based on an

incorrect model, in which $E(\hat{m}_{pi}) \neq (m_{pi})$.

3.4.2 Stepwise selection

For many data sets, it may be impractical to examine all possible subsets. We used the stepwise approach, which has virtually no limit as to the number of variables or observations under study. We were concerned with selecting the independent variables $(x's)$ that best predict the dependent variable (m_p) in regression. We first review the forward selection procedure, which typically uses an F-test at each step. At the first step,m_p is regressed on each x_j alone, and the x with the largest F-value is centered into the model. At the second step, we search for the variable with the largest partial F-value for testing the significance of each variable in the presence of the variable first entered. Thus, if we denote the first variable to enter as x_1, then at the second step we calculate the partial F-statistic;

$$F = \frac{MSR(x_j|x_1)}{MSE(x_j, x_1)}$$

for each $j \neq 1$ and choose the variable that maximizes F, where $MSR = \frac{SSR_f - SSR_r}{h}$ and $MSE = \frac{SSE_f}{141}$ are the mean squares for regression and error, respectively. In this case,$SSR_f = SSR(x_1, x_j)$ and $SSR_r = SSR(x_1)$. We note also that h = 1 because only one variable is being added, and MSE is calculated using only the variable already entered plus the candidate variable. This procedure continues at each step until the largest partial F for an entering variable falls below a preselected threshold F-value or until the corresponding p-value exceeds some predetermined level.

The stepwise selection procedure similarly seeks the best variable to enter at each step. Then after a variable has entered, each of the variables previously entered is examined by a partial F-test to see if it is no longer significant and can be dropped from the model. The backward elimination procedure begins with all $x's$ in the model and deletes one at a time.

The partial F-statistic for each variable in the presence of the others is calculated, and the variable with smallest F is eliminated. This continues until the smallest F at some step exceeds a preselected threshold value.Since these sequential methods do not examine all subsets, they will often fail to find the optimum subset, especially if k is large. However, R_p^2, s_p^2, or C_p may not differ substantially between the optimum subset and the one found by stepwise selection. There are some possible risks in the use of stepwise methods. The stepwise procedure may fail to detect a true predictor (an x_j for which $N_j \neq 0$) because s_p^2 is biased upward in an under specified model, thus artificially reducing the partial F-value. On the other hand, a variable that is not a true predictor of m_p (an x_j for which $N_j = 0$ may enter because of chance correlations in a particular sample. In the presence of such noise variables, the partial F-statistic for the entering variable does not have an F-distribution because it is maximized at each step. The calculated p-values become optimistic. This problem intensifies when the sample size is relatively small compared to the number of variables. At times large values of R^2 can occur, even when there is no relationship between m_p and the $x's$ in the population. Sometimes if authentic x contributors as well as noise variables are included ,for most samples, a large percentage of the selected variables is noise, particularly when the number of candidate variables is large relative to the number of observations. The adjusted R^2 of the selected variables is highly inflated.

3.5 Measuring Goodness of Fit

The quantity R^2 - coefficient of determination gives the proportion of the total variation in the $m_{p's}$ explained by or attributable to the predictor variables $f_a, a_g, c_a, m_s, t_r, e_x$. We showed how well the estimated regression equation fitted the data using R^2.[See 3.5.3].Here R^2 equals 1 if the fitted equation passes through all the data points so that the estimated $\varepsilon_i = 0$ for all i's.At the other extreme ,$R^2 = 0$ if $N_0 = m_p$ and $\hat{N}_1 = \hat{N}_2 = ... = \hat{N}_6 = 0$. In

this case the predictor variables have no influence on the response. $i^{th}residual = m_{pi} - \hat{m}_p$

$$SSE = \sum (m_{pi} - \hat{m}_{pi})^2 \tag{3.5.1}$$

SSE measures the error in using the estimated regression equation.

$m_{pi} - \hat{m}_{pi}$ is the error involved in using $\hat{m}_p$ to estimate

$$SST = \sum (m_{pi} - \bar{m}_p)^2 \tag{3.5.2}$$

SST is a measure of how well the observations cluster about the $\bar{m}_p$ line and SSE is a measure of how well the observations cluster about the $\hat{m}_p$ line.

$$SST = SSR + SSE \; and \; \frac{SSR}{SST} = R^2 \tag{3.5.3}$$

which is used to evaluate goodness of fit for the estimated regression equation where

- $0 \leq R^2 \leq 1$
- $R^2 = 0$ poorest fit
- $R^2 = 1$ best fit

For many independent variables we have multiple coefficient of determination R squared, measuring goodness of fit for the estimated multiple regression equation $R^2 = \frac{SSR}{SST}$. R^2 may be adjusted for the number of independent variables to avoid overestimating the impact of adding an independent variables on the amount of variability explained by the estimated regression equation with k denoting the number of observations and q denoting the number of independent variables, the adjusted multiple coefficient of determination is computed as follows:

$$R^2_{adj} = \frac{(1 - (1 - R^2)k - 1)}{k - 1 - q} = \frac{1 - (1 - R^2)147}{141} \tag{3.5.4}$$

3.6 Testing for significance

With $m_{pi} = n_0 + n_i x_i$,

$$SSE = \sum (m_{pi} - \hat{m}_{pi})^2 = \sum (m_{pi} - n_0 - n_i x_i)^2,$$

$x_i's$ the predictor variables, n_i's the estimates of N_i's. Mean square error (estimate of σ^2;)

$$S^2 = MSE = SSE = \frac{SSE}{146}. \tag{3.6.1}$$

And standard error of the estimate;

$$S = \sqrt{MSE} = \sqrt{\frac{SSE}{146}} \tag{3.6.2}$$

The t- test

If x and m_p are linearly related then $N_1 \neq 0$. t test helps to conclude weather $N_1 \neq N_0$

$H_0 : N_1 = 0$

$H_1 : N_1 \neq 0$

$t = \frac{n_1}{s_{n1}}$. The rejection rule : p value approach : reject H_0 if p-value$\leq \alpha$ critical value approach : reject H_0 if $t \leq -t\frac{a}{2}$or if $t \geq t\frac{a}{2}, t\frac{a}{2} = t_{0.05}$ where $t_{0.05}$is based on a t distribution with 146 degrees of freedom.

Sampling distribution of n_1 is such that; $E(n_1) = N_1$. Standard deviation of σ_{n1} is

$$\sigma_{n1}; = \frac{\sigma}{\sqrt{\sum (x_i - \bar{x})^2}} \tag{3.6.3}$$

Estimated standard deviation of s_{n1}

$$s_{n1} = \frac{s}{\sqrt{\sum(x_i - \bar{x})^2}} \tag{3.6.4}$$

In the multivariate case H_0: $N_i = 0$; H_1 : $Ni \neq 0$; $t = \frac{n_i}{Sn_i}$. p value approach : reject H_0 if p-value $\leq \alpha$; critical value approach : reject H_0 if $t \leq -t\frac{a}{2}$ or if $t \geq t\frac{a}{2}, t\frac{a}{2} = t_{0.05}$ where $t_{0.05}$is based on a t distribution with 141 degrees of freedom.

The F Test

For Multivariate Case F-test can is used to test for an overall significant relationship. F test for the bivariate case :$H_0 : N_1 = 0$; $H_1 : N_1 \neq 0$;

$$F = \frac{MSR}{MSE} \tag{3.6.5}$$

Rejection Rule: p-value approach: Reject H_0 if p value $\leq \alpha$: a Critical value approach: Reject H_0 if $F \geq F_{0.05}$, where $F_{0.05}$ is based on an F distribution with 1 degree of freedom in the numerator and 146 degree of freedom in the denominator. For many x variables, F test for overall significance; $H_O : N_1 = N_2 ..... N_6 = 0$; H_1: one or more of the parameters is not equal to zero; $F = \frac{MSR}{MSE}$.

Rejection Rule: p-value approach: Reject H_0 if p value ≤ 0.05. Critical value approach: Reject H_0 if $F \geq F_{0.05}$,where $F_{0.05}$ is based on an F distribution with p degree of freedom in the numerator and $k - q - 1 = 141$ degree of freedom in the denominator.

Chapter 4

Data Analysis,Presentation and Discussion

4.0.1 Descriptive Statistics

The first step in the analysis was to explore the data. This was done by calculating the median, mean,inter-quartile range and standard deviation for all the fields (columns) in the data. The data was collected from 148 businesses. The mean business floor area was 19.7 (SD = 5.7) and median business age of 5.3 years (IQR = 2.5 - 10).There was a median capital investment of 90000 shillings (IQR = 25000 - 231000). The median monthly sale was 26100 shillings (IQR = 7500 - 69000) given a median number of 318 transactions (IQR = 233 - 539) in a month. The median monthly expense was 16000 shillings (IQR = 4060 - 40425) while the median profit was 2500 (IQR = 850 - 6500). They are tabulated as below.

Table 4.180

Descriptive Statistics

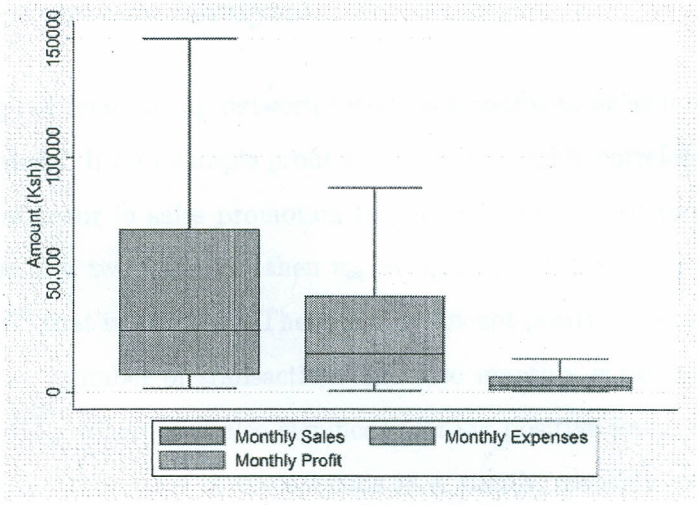
Variable	Obstn	Mean	S.D	Min	Median	Max	IQR
Area	148	19.74	5.69	9	18	36	8.375
Age	148	88.74	79.8	6	64	300	90.5
Capital	148	151510.8	168922.1	3000	88500	890000	208499.8
M sals	148	44308.65	50492.62	1500	25000	267000	61800
No Tran	148	451.89	589.8	126	317	7000	312.25
Exp	148	26173	29168.6	500	15925	173250	36348.75
Profit	148	4557.73	5043.09	300	2500	25000	5612.5

Box plots

We illustrated the above information using boxplots.A boxplot is a graphical technique that displays the distribution of variables. It helps us see the location, skewness , spread, tail length and outlying points. It is particularly useful in comparing batches. The rectangles extend from the first to the third sample quartile thereby drawing attention to the central 50 % of the data .Thus the length of the rectangle equals the sample interquartile range . The location of the sample median is identified.Whiskers extend from the ends of the rectangles to the extreme values of the data,beyond which lie the outliers(see appendix). The largest business had floor area of $36m^2$ with a median of $18m^2$ and a mean of $17.7m^2$. Lower quartile was $15.5m^2$ and upper quartile $24m^2$, giving an interquartile range of $8.4m^2$ .The standard deviation of $5.7m^2$ implies most of the retailers in the region have a floor area of $\pm 5.7m^2$ around $17.7m^2$. The oldest business observed was 25 years with a median age of 5.3 years and a mean of 7.4 years. Lower quartile was about 2.5 years and upper quartile 10 years, giving an interquartile range of 7.5 years.The standard deviation of 6.6 years implies most of the retailers in the region have been in the business for a period ± 6.6 years around 7.4 years. The highest amount of capital invested was eight hundred and ninety thousand shillings with a median of ksh.88500 and a mean of ksh.151511. Lower quartile was ksh.25000 and upper quartile ksh.231000 giving an interquartile range of ksh.208500 .The standard deviation of ksh.168900 implies most of the retailers in the region start businesses with $\pm ksh.168900$ around ksh.151500.

There are many outlying capital values ,considering the minimum,mean and sd. The most number of business deals made was 7000 with a median of 317 and a mean of 452. Lower quartile was 233 and upper quartile 539 giving an interquartile range of 312 .The standard deviation of 589 implies that most of the retailers in the region have minimal number of transactions in a month ,that is ± 589 around 451.

Figure 4.0.1: Box plot of Sales, expense and profit



Sales, expense and profit in one graph(in that order)are seen to display some pattern and decrease in spread and magnitude in terms of figures in shillings they assume.Probably this could be the reason why they have remained in the reduced model. Highest expense made was ksh.173250 per month with a median of ksh.15925 and a mean of ksh.26173. Lower quartile was ksh.4060 and upper quartile ksh.40425, giving an interquartile range of ksh.36350 .The standard deviation of ksh.29168 implies there is not much of outlying expenses incurred. Maximum sales made was observed at ksh.267000 per month with a median of ksh.25000 and a mean of ksh.44300. Lower quartile was ksh.7500 and upper quartile ksh.69000, giving an interquartile range of ksh.61800 .The standard deviation of

ksh.50490 implies most of the retailers in the region make monthly sales of $\pm ksh.50490$ around ksh.44300. Maximum profit realised was ksh.25000 with a median of ksh.2500 and a mean of ksh.4560. Lower quartile was ksh.850 and upper quartile ksh.6500, giving an interquartile range of ksh.5610 .The standard deviation of ksh.5043 suggests that most of the retailers realise monthly profits below the average.Key variables noted to affect profit are the sales and expenses.

4.0.2 Correlation Analysis

The knowledge of relationship between variables is useful to us as it enables us to predict and control events. If for example profit and sales are highly correlated we can know how much to use or incur in sales promotion to generate the desired profit. If we generalize u and v to be any two variables then $r_{uv} = s'_{uv}s_u s_v$. Multiple correlation coefficient is derived from R^2,that is $\pm N_0\sqrt{R^2}$. There was significant positive linear correlation between business age vs. number of transactions and also monthly expenses,$r = 0.483$ and $r = 0.235$ respectively. There was also significant strong positive linear correlations between capital invested vs. number of transactions in a month, monthly expenses and monthly profit ; $r = 0.990, 0.735$ and 0.914 respectively. We had significant strong positive linear correlations between monthly sales vs. monthly expenses and profit ,$r = 0.708, 0.923$ respectively. Monthly expenses were also positively correlated to monthly profit , $r = 0.644$.

Table 4.190

Correlation Analysis

	Area	Age	Capital	Sales	Transactions	Expenses	Profit
Area	1						
Age	-0.0036	1					
p-value	0.9659						
Capital	-0 .1142	0 .1358	1				
P-value	0.1715	0.1035					
Sales	-0 .1394	0.1325	0.9901*	1			
p-value	0 .0945	0.1121	.001				
Trans.	0 .0448	0.4831*	0.1132	0.1108	1		
p-value	0 .593	0.001	0.1754	0.1846			
Expe.	0.0649	0.2347*	0.7354*	0.7077*	0.1409	1	
p-value	0.4381	0.0045	0.001	0.001	0.091		
Profit	-0 .106	0.1121	0.9143*	0.9225*	0.0884	0.6440*	1
p-value	0.2044	0.1974	0.001	0.001	0.2901	0.01	

* Significant correlation($p < 0.05$)

Scatter Plots

This information on correlation is well illustrated using scatter plots.

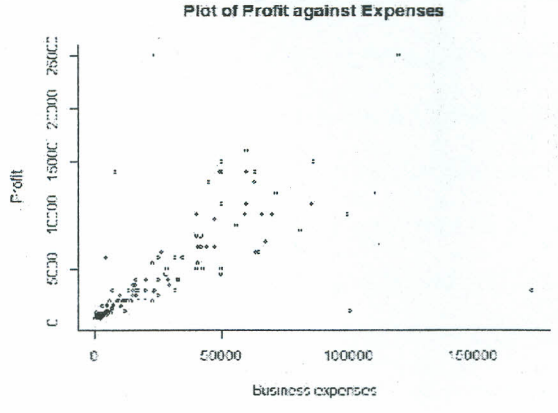
These are bivariate or trivariate plots of variables against each other. Scatter plots help us understand the relationships among the variables of the data set .A downward slopping scatter indicates that as we increase the variable on the horizontal axis the variable on the vertical axis decreases. An analogous statement can be made for upward slopping scatters.

As can be seen area and profit have no significant relationship.From the table the correlation coefficient stands at -0.106 with a p-value of 0.2044.This means a retailer can't rely on area alone to predict profit. Correlation between profit and age of business is not significant(r=0.1121)with a p-value of 0.1974.Business age helps the retailer very little in predicting the profit. There is a strong positive correlation between profit and capital.The

higher the capital the higher the profit as can also be seen in figure 4.1.16 with the fitted regression line.Retailers should strive to boost their capital base to enjoy economies of scale which could be the possible cause of the high profit. There is a strong positive correlation between profit and monthly sales.Also Figure 4.0.3 with the fitted line attests to this.The higher the monthly sales the higher the profit.This means a lot of effort should be put on sales promotion activities such as advertisement. It can be observed most of the retailers hardly had the monthly transactions exceeding 1000.Correlation between profit and transaction is not significant.It is not easy to tell the nature of profit using the magnitude of transactions alone.

This information was well illustrated using scatter plots. The combinations above are many; therefore the plotting was done for selected combinations. Since the variable profit is of main interest, it was plotted against the rest of the variables, whether they had a statistically significant correlation coefficient or not, as shown below(see appendix also).

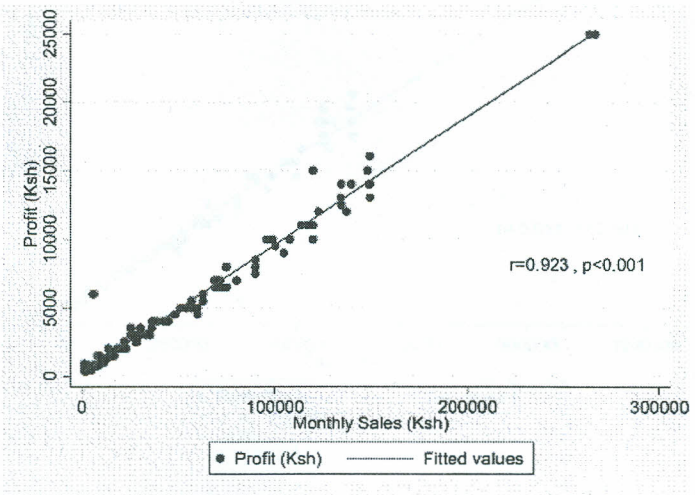
Figure 4.0.2: Scatter plot of profit against expenses



Correlation between profit and expenses is significant(0.01).The higher the expenses the higher the profit.This means most retailers dont spend business money carelessly but

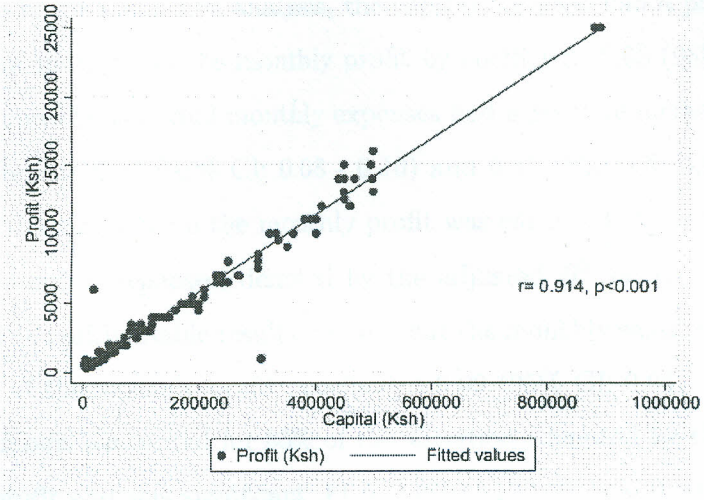
on core business activities that eventually lead to high profit.

Figure 4.0.3: Scatter plot of profit against monthly sales in kshs



Scatter plot of profit against monthly sales in shillings showing fitted regression line.

Figure 4.0.4: Scatter plot of profit against capital



Scatter plot of profit against capital showing fitted regression line.

The scatter plot of transaction against age of business with fitted regression line above shows that transaction and age of business have a significant linear relationship as the fitted line indicates.On the basis of this retailers can be advised to be patient for sometime upon entry into market before they can realise a stream of customers.

Generally the correlation coefficients between profit and capital and profit and sales have values very close to positive one. In addition they are statistically significant. When the two combinations are plotted, the points almost form a straight line ascending from the bottom left to the top right corner of the box. The points also form a near straight line when profit is plotted against expenses (correlation coefficient is 0.6).This could account for the reduced model containing sales and expenses only. The points scatter more for the other coefficient values since they are close to zero.

4.0.3 Regression Diagnostics

On univariable linear regression analysis, the capital invested had a positive and statistically significant increase on the monthly profit by coefficient 0.03 (95% CI: 0.02 - 0.03). Similarly the monthly sales and monthly expenses had a positive increase on the monthly profit by coefficients 0.09 (95% CI: 0.08 - 0.10) and 0.11 (95% CI: 0.09 - 0.13) respectively. Most of the variation in the monthly profit was captured by monthly sales, capital invested and monthly expenses; denoted by the adjusted R^2 values 0.85, 0.84 and 0.42 respectively. The multivariable results showed that the monthly sales retained significance as a predictor of the monthly profit , coefficient 0.09 (95% CI: 0.04 - 0.14). The overall multivariable model accounted for 85% of the variation (denoted by the adjusted R^2) in the monthly profit and was significant ,$F(6, 148) = 132.62$, $p < 0.001$. The coefficients from the multivariable linear regression model was used to come up with a market model to determine the monthly profit of businesses.

table 4.200

Linear Regression Analysis

	Univar. Model			Multiva. model			
Covar.	Adj R^2	$N_1(CI)$	PVal.	Adj R^2	$N_i(CI)$	PVal	
f_a	0.01	-93.84(-239.32 -51.65)	0.204	0.85	24.09(-35.84 -84.04)	0.428	
a_g	0.01	85.26(-39.66 -210.18)	0.179		-0.17(-57.7 -57.37	0.995	
c_a	0.84	0.03(0.02-0.03)	* < 0.01		0.01(-0.01 0.02)	0.00976*	
m_s	0.85	0.09(0.08-0.10)	* < 0.01		0 .09(0.04 -0.14)	* < 0.01	
t_r	0.01	0.75(-0.65-2.16)	0.29		-0.12(-0.75 0.51)	0.712	
e_x	0.42	0.11(0.09-0.13)	* < 0.01		-0.01(-0.02 -0.01)	0.0069*	
Constant					61.05(-1259.19 1381.3)	0.927	

* Significant , p < 0.05

4.0.4 The Fitted full model

From table 4.200 above we have our final model formulated as :

$$m_p = N_0 + N_1(f_a) + N_2(a_g) + N_3(c_a) + N_4(m_s) + N_5(t_r) + N_6(e_x) + \varepsilon$$

$$= 61.5 + 24.09\ (f_a) - 0.17\ (a_g) + 0.01\ (c_a) + 0.09(m_s) - 0.12\ (t_r) - 0.01\ (e_x) + \varepsilon.$$

We,however noted with concern that some observations had large influence on the

model either because of a large residual or because of x values that were significantly larger. One method that is used to detect influential values uses a statistic called Cooks distance, D. The Cooks distance method compares the values of the regression coefficients with all observation to the values when the i^{th} observation is removed from the model. If the i^{th} observation is influencing the model , then the difference in coefficients between the two models is large and the corresponding values of cooks distance is also large . A suggested definition of large is to compare D with $f_{0.5,q+1,k-q-1}$, and f statistic of a tail area of 0.50 and with $q+1$ degrees of freedom in the numerator and k-q- 1 in the denominator . A rougher rule of thumb is to compare cooks distance with 1. If, in either case D exceed the critical value , then the observation in question should be removed from the model.

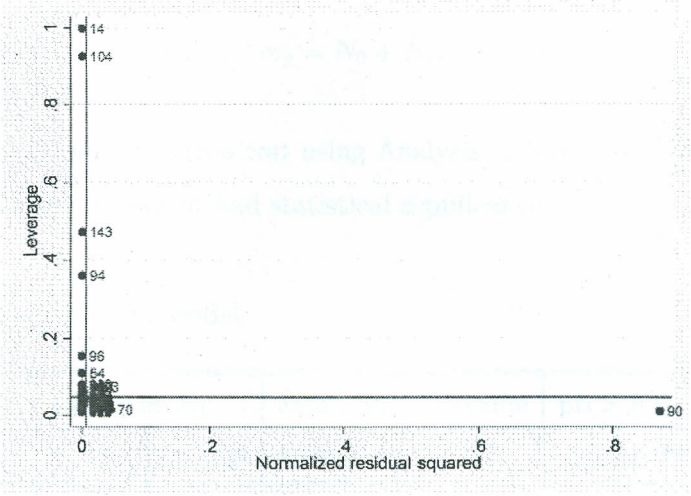
We identified the Unusual and influential data by using outlier , leverage and influence values.We used a leverage versus residual plot that shows the leverage by the residual squared and look for observations that are jointly high on both of these measures. It is a way of checking potential influential observations and outliers at the same time.This identified observation 90 as having a high residual value and observation 14 and 104 as having high leverage values. A closer look at the values reveal inconsistencies (see table below). The two red reference lines[figure 4.0.5] are the means for normalized residual squared (horizontal) and the means for leverage (vertical).

Table 4.210

Unusual and influential data

Number	Area	Capital	Sales	Trans	Exp	Profit	Age
14	17.5	300000	90000	7000	42000	8000	16.25
90	24	94000	28200	390	22560	25000	5.83
104	33	306000	9180	596	100980	1000	9.67

Figure 4.0.5: Plot of leverage against normalised residuals squared



The lowest value that Cook's D can have is zero, and the higher the Cook's D is, the more influential the value. The conventional cut-off point is $\frac{4}{k}$ where k (148) is the number of observations. It revealed the same observations (90,14 and 104) as influential.

Table 4.220

Cook's distance for Unusual and influential data

obs.no.	Area	Capital	Sales	Trans	Exp	Profit	Age	Cooks D
14	17.5	300000	90000	7000	42000	8000	16.25	4.1718
90	24	94000	28200	390	22560	25000	5.83	1
104	33	306000	9180	596	100980	1000	9.67	2.82

For the univariate regression models, the monthly sales were first used as the explanatory variable for the monthly profits,considering its significance as shown in the table above.

The following general liner model was used:

$$m_p = N_0 + N_1 sales + \varepsilon$$

This model was subjected to a test using Analysis of Variance. This test would indicate whether the slope coefficient had statistical significance.

Table 4.230

Anova for the univariate model

	Df	sum sq	Mean Sq	F-value	pr(> F)
M sals	1	3.185e+09	3.185e+09	840.7	<2e-16***
Residuals	146	5.532e+08	3.789e+06		

Note that 2e+05 means 200000, 4e+05 means 400000, and so on.

With a p-value less than $2X\ 10^{-16}$ the model was statically significant. The coefficients were calculated and are summarized in the table below. The null hypotheses for the terms in a regression equation are usually that "the coefficients are equal to zero". In this case, both the intercept and the slope were statistically significantly different from zero.

Table 4.240

P-value and F-value for the univariate model

	Coeff. Estimate	Std Error	t-value	pr(> $\|t\|$)
Intercept	472.97716	213.19353	2.218	0.0281*
M sals	0.09219	0.00318	28.994	$< 2e-16$***

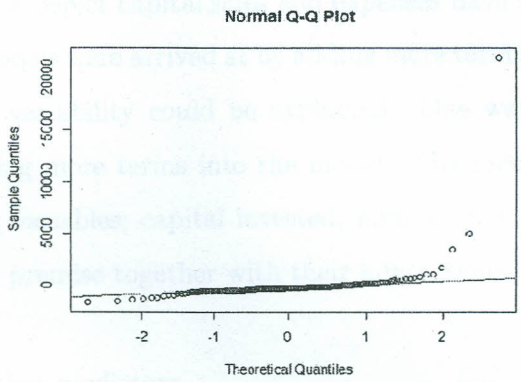
The regression equation can therefore be written as:

$$m_p = 472.8 + 0.09 m_s + \varepsilon.$$

Since a general linear model was used, the residuals were tested for normality. This was done by the use of a residual plot,to check if the assumptions hold.

The line in fig.4.0.6 is called the Q-Q line.The more the deviation of plots from the red line, the less normal the residuals are. In this case, the residuals are close to the normal line hence the general linear model was suitable for use in this model. See below.

Figure 4.0.6: Residual plot of sample quantiles against theoretical quantiles



Through residual analysis,we can tell whether we are violating the assumptions or using the linear model incorrectly.This can be done in other ways such as checking if the residuals have a mean of zero or randomly dispersed around this value,by plotting residuals(or standardised residuals) against the fitted valuesm_p or x. When the model is significant, m_p and x are related and so the plots will not have different shapes.Further,we can check the shape of the distribution.The normal probability plot is also important.It is a plot of the ordered data against their expected values under normal distribution.When the data are normally distributed ,the plot is a straight line.

It is however also important to look at how much variability the model explained. This

was done by working out the R^2.The formula for calculating R^2 is given by(or see 3.1.3)

$$R^2 = (1 - \frac{SS_r}{SS_T}) = (1 - \frac{5.532e+08}{(5.532e+08+3.185e+09)}) = 0.85201$$

The R^2 gives the amount of variability that the regression equation explained. The regression equation above explained up to 85.201 % of the total variability. This is a good amount of variability explained.

The other univariable models are as per table 4.200 above with their respective R^2 and p-values.These models depict capital,sales and expenses have significant effect on profit. The multivariable models were arrived at by adding more terms into the univariable model above so that more variability could be explained. One way to have more variability explained is by adding more terms into the model. The forward selection was used for this. The remaining variables; capital invested, number of transactions, age of business and area of business premise together with their interactions were added and tested.

Table 4.250

significance of the other predictors

	Df	Deviance	AIC	F-value	pr(F)
$< None >$		553208180	2665.8		
area	1	551154933	2667.3	0.5402	0.4635
age	1	552806088	2667.7	0.1055	0 .7458
Capital	1	553204898	2667.8	0.0009	0.00976
No Tran	1	552486620	2667.7	0.1894	0.6641
Exp	1	552616465	2667.7	0.1553	0.006941

From the above it can be seen that there is no other variable that has a statistically significant effect on the profit except for capital and expenses. For instance, in case we select capital and add it to the model so that we have the equation

$$m_p = N_0 + N_1 m_s + N_2 c_a + N_3 m_s : c_a + \varepsilon$$

. And we do similar analysis on it as done for the previous model, we get that again the sales have higher statistically significant effect at the alpha equals 0.05 level (p value is less than 0.01) compared to Capital and the interaction between capital and sales.

Table 4.260

Anova:Capital and the interaction between capital and sales

	Df	Sum Sq	Mean Sq	F Value	pr(> F)
M sals	1	3.185e+09	3.185e+09	829.449	$< 2e-16^{***}$
Capital	1	3.282e+03	3.282e+03	0.001	0.00976
M sals: Capital	1	1.875e+05	1.875e+05	0.0490	0.0825
Residuals	144	5.530e+08	3.840e+06		

Signif. Codes $(0,^{***}), (0.001,^{**}), (0.01,^{*}), (0.05,^{'.'}), (0.1,^{''})$.

dding the new variables in the model added little explanation in the variability. The R^2 changed to 0.85207 from 0.85201 when capital was not included.

$$R^2 = (1-\frac{SS_r}{SS_T}) = (1-\frac{5.530e+08}{(5.530e+08+3.185e+09+3.282e+03+1.875e+05)}) = 0.85207$$

Finally, coefficients for the model together with their 95% Confidence Interval were calculated. The tests showed that sales had statistically significant effect on the profit. Capital and the interaction between capital and sales did not have this statistically significant difference.

Table 4.270

Capital and the interaction between capital and sales:confidence intervals and their significance

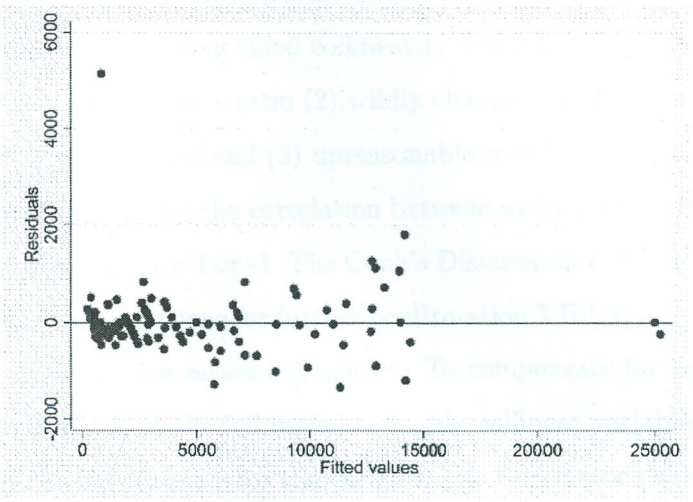
	Coefficients	lower 0.95 ci	upper0.95 ci	pr(> |t|)
Intercept	5.027993e+02	1.427369e+00	1.004171e+03	4.935991e-02
M sals	9.135841e-02	5.811481e-02	1.24602e-01	2.315888e-07
Capital	-2.144138e-04	-9.962455e-03	9.133627e-03	9.639021e-0001
M sals;Capital	2.628380e-09	-2.088433e-08	2.614109e-08	8.254425e-001

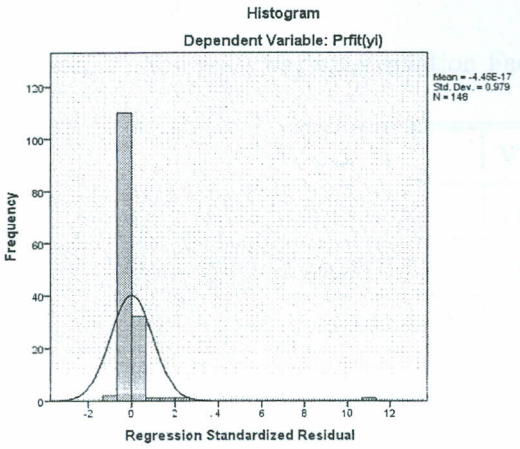
4.0.5 checking multivariate model adequacy

Homoscedasticity

For the multivariate model also, normality assessment was done.Usually residuals are plotted in various ways to detect possible anomalies for example in plotting residuals against predicted values departure from the assumptions is depicted by dependence of the residuals on the predicted value and non constant variance. An assessment of homoscedasticity indicated using a graphical method plotting the residuals versus fitted (predicted) values. Both Cameron Trivedi's decomposition of IM-test (p value =1.000) and Breusch-Pagan / Cook-Weisberg test for heteroscedasticity (p value =0.315) confirm that there is homogeneity in the residual distribution. Both test the null hypothesis that the variance of the residuals is homogenous (see appendix).

Figure 4.0.7: Residuals vs Fitted values





Assessment of Multi-Collinearity

The correlation among independent variables was investigated in the models built. It is a problem with being able to separate the effects of two (or more) variables on an outcome variable, that is, correlation among independent variables. We expected to find dependencies among the independent variables, x_i's , such that a linear constraint holds

approximately among the columns of the **X** matrix. If two variables are significantly alike, it becomes impossible to determine which of the variables accounts for variance in the dependent variable. We concluded collinearity based on such symptoms as (1) having significant F,but no significant t-ratio (2) wildly changing coefficients when an additional (collinear) variable is included and (3) unreasonable coefficients eg large R^2. To diagnose multicollinearity we calculate the correlation between every pair of independent variables and look for large values near 1 or -1 .The Cook's Distance and the other test, the variance inflation factor,vif, are also done for further confirmation. VIF(N_j)=$\frac{1}{1-R_j^2}$, j=1,2,..k. Large V(b_j)=$\delta^2(1-R_j^2)^{-1}$ implies severe collinearity. To compensate for this problem we eliminated some variables from the full model as highly collinear variables contain redundant information. As a rule of thumb for the variance inflation factor (VIF)., a variable whose VIF values are greater than 10 may merit further investigation. Tolerance, defined as $\frac{1}{VIF}$, is used to check on the degree of collinearity. A tolerance value lower than 0.1 is comparable to a VIF of 10.

Table 4.280

Multicollinearity:Variance inflation Factor and tolerance values

Variable	VIF	1/VIF
Capital*	1153.13	0 .000867
Sales*	1151.4	0 .000869
Transactions**	106.04	.00943
Age**	101.85	0 .009819
Area	4.65	0.214914
Expenses	2.4	0 .416848
Mean	419.91	

*/**collinear

We identified the following pairs; Capital and Sales ; Age and number of transactions as being collinear ($VIF > 10$). Only one from each pair was included in the stepwise model.

Linearity Assessment

Figure 4.0.8: Augmented components plus residuals vs monthly sales

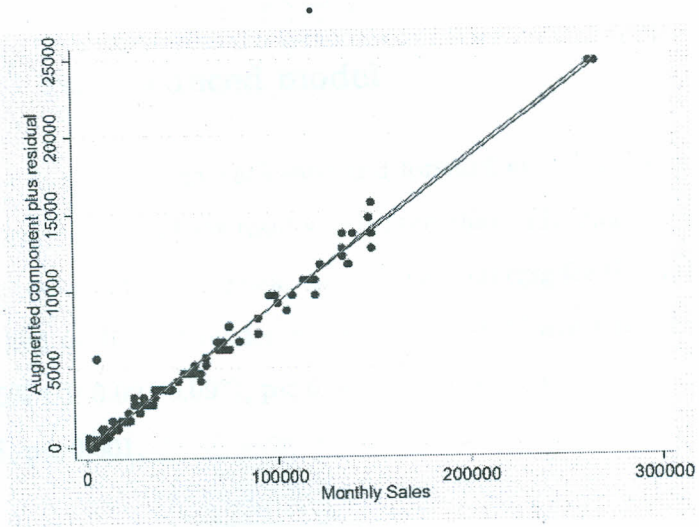
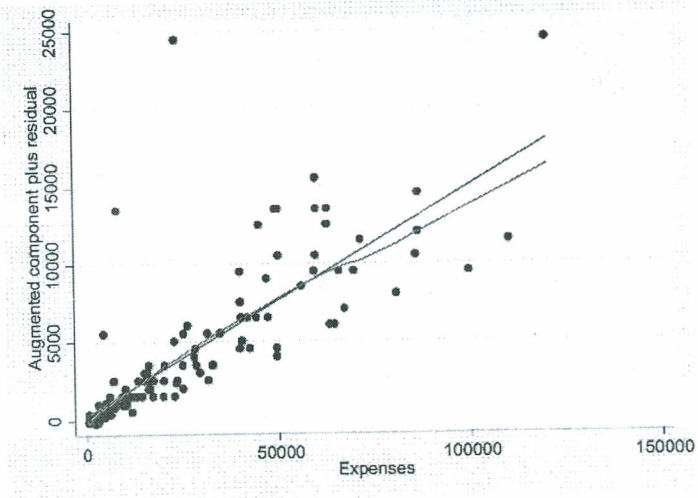


Figure 4.0.9: Augmented components plus residuals vs expenses



Linearity assessment for monthly sales, number of transactions(appendix) and monthly expenses all show no or very little deviation from normality which was ignorable.Note that monthly sales and expenses have significant effect on profit as seen in the reduced model.

4.0.6 The Fitted reduced model

Further stepwise analysis (both backward and forward at a p value of 0.20 for inclusion and removal) regression analysis results indicated that only monthly sales and expenses remained in the two models giving near complete accounting for the variance in the model, R^2= 0.98.As noted in the first model, monthly sales still have a significant impact on the profit 0.094 (95% CI: 0.091 0.097), p< 0.001. The overall model fit was very good, F(2, 138) = 3962.68, p< 0.001.

Table 4.290

Fitted reduced model excluding 3 outliers identified

	Univar. Model			Multiva. Model			
Covar.	Adj R^2	$N_1(CI)$	PVal.	Adj R^2	$N_i(CI)$	PVal	
f_a	0.01	-93.84(-239.32 -51.65)	0.204	0.85	-	-	
a_g	0.01	85.26(-39.66 -210.18)	0.179		-	-	
c_a	0.84	0.03(0.02-0.03)	* < 0.01		-	-	
m_s	0.85	0.09(0.08-0.10)	* < 0.01		0 .094(0.091 -0.097)	* < 0.001	
t_r	0.01	0.75(-0.65-2.16)	0.29		-	-	
e_x	0.42	0.11(0.09-0.13)	* < 0.01		-0.004(-0.009 -0.001)	0.0137*	
Constant					311.492(167.73 455.25)	* 0.927	

* Significant(p < 0.05)

Hence the fitted reduced model becomes;

$$m_p = N_0 + N_4(m_s) + N_6(e_x) + \varepsilon$$

$=311.492 + 0.094(ms) - 0.004(ex) + \varepsilon$

4.0.7 Using the model

We can now use the reduced market model for estimation and prediction . We can do this by using the model to develop a point estimate of the mean value of m_p for a particular

value of x or to predict an individual value of m_p corresponding to a given value x. We then substituted the given values of $x_1x_2, ...x_6$ (the predictor variables) into the estimated regression equation and use the corresponding value of of $\hat{m}_p$ as the point estimate.

The measure Confidence interval Provides an estimate for the mean value of m_p at a particular value of x $(\mu_{m_p/x})$. The value of m_p obtained from the regression model is an estimate of the mean value of m_p for a given value of X; that is, value m_p is an estimate of the true mean, $(\mu_{m_p/x})$. If the variable x represents monthly sales or expenditures,then the regression model that relates x to profit m_p , predicts the average profit for a given level of monthly sales and expenditure. Because the estimate is based on sample data, there is some error in that estimate.

To find a confidence interval for a population parameter we use the sampling distribution of the estimate and the standard error of the sample statistic. The standard error is related to the standard error of the estimate, S_{m_p}/x. The formula for the 95% confidence interval for the mean value of m_p for a given value of $X = x_i$, $(\mu_{m_p/x})$ is:

$$\hat{m}_{pi} - t_{0.025,146}, s_{mp.x}\sqrt{\frac{1}{148} + \frac{(x_i - \bar{x})^2}{\sum x^2 - \frac{(\sum x)^2}{148}}} \leq \mu_{m_p.x_i} \leq \hat{m}_{pi} + t_{0.025,146}, s_{mp.x}\sqrt{\frac{1}{148} + \frac{(x_i - \bar{x})^2}{\sum x^2 - \frac{(\sum x)^2}{148}}} \tag{4.0.1}$$

Confidence interval for the mean estimate from the regression model gives the retailer an idea of how accurate the estimate is over a long period of time ;that is,it is an interval estimate for the mean.

Another important measure,the Prediction Interval provides an estimate for an individual value of m_p at a particular value of x. The formula used to calculate the 95% prediction interval for our regression model is almost identical to the formula for the confidence interval:

$$\hat{m}_{pi} - t_{0.025,146}, s_{mp.x}\sqrt{1 + \frac{1}{148} + \frac{(x_i - \bar{x})^2}{\sum x^2 - \frac{(\sum x)^2}{148}}} \leq \mu_{m_p.x_i} \leq \hat{m}_{pi} + t_{0.025,146}, s_{mp.x}\sqrt{1 + \frac{1}{148} + \frac{(x_i - \bar{x})^2}{\sum x^2 - \frac{(\sum x)^2}{148}}} \tag{4.0.2}$$

and

$$Sm_{pi} = \sqrt{\frac{(\sum m_{pi} - \hat{m}_{pi})}{146}} \tag{4.0.3}$$

The only difference is the 1 in the square root part of the formula. The impact of this change is that the prediction intervals for a specific value of X are wider than the corresponding confidence intervals given that the distribution of individual observations is always more variable than the distribution of the sample means.

The prediction intervals are important when the individual values of the m_p variable are critical to the decision process. If for example the retailer is interested in estimating cash flow, then an average price does not help him much but the specific estimate.

Chapter 5

Summary,Conclusion and Recommendations

The process of finding a multiple regression model may be easy but finding a model that is useful for decision making is an art.We further note that modeling is an iterative process that has no single correct answer.The answer one chooses depends on what one knows and understands about the problem to be solved and how one intends to use the model. There were several steps to the modeling process:Identifying potential independent variables,collecting data and finding a potential model were the beginning steps.Once a potential model was found the process of model building took place to find the best model.The objective of this process was to find a model that does an acceptable job of explaining or predicting the dependent variable with as few independent variables as possible.Some of the model building techniques used were forward selection,backward elimination,stepwise regression,and all possible regressions.Once the best model was identified it was checked for problems such as violation of assumptions,influential observations and multicollinearity before it could be used for decision making,description,control,and prediction.Description is important when trying to understand the way the variables are related.Control describes when the model is used to set standards or reduce variability .Prediction is when the model is used to determine what the resulting m_p value should be when x takes on certain values.

In the analysis, data was first explored using box plots, and tabulation of summary statistics. The data were then tested for correlation. To aid this, profit was plotted against the variables recorded using scatter plots . It could be seen that sales and capital had strong correlation with the profit. The variables were then used as explanatory variables to explain the profit encountered. The univariate model selected showed no variable had statistically significant effect on the profit except monthly sales and capital to some extent(see tables 4.240 and 4.250).The profit increased by 0.09219 for a unit increase in sales.Addition of new variables added very little to the variability(table 4.260) This model was only valid for sales between 1500 and 267000.The forward selection was used to add the remaining variable to the model but their coefficients were not statistically significant. There was high positive correlation between capital and profit though this did not have a fairly statistically significant effect on the profit probably due to the presence of outlier capital values of 880000 and 890000. Monthly sales,capital and expenses have effect in the univariable models as exhibited in table[4.200].These variables must be taken seriously by the businessman due to their overall impact on the profit.The full model in [4.0.4] illustrates the contribution of each predictor variable for a unit change in monthly profit.Table [4.280] shows that some predictors are redundant,leading to a reduced model under table [4.290] using stepwise analysis.Capital is dropped because of its collinearity with sales[table 4.280] ,otherwise it could replace it in the model. With sales and expenses as the only predictors ,the model proved to be so efficient as this saw R^2 adjusted moving from 0.85 to 0.98 which was very good.It means the six predictor variables we initially had are not all statistically significant in determining the monthly profit .The businessman can therefore choose his model based on environmental and economic circumstances surrounding the predictor variables.

Finally we are saying that the statistical model results may strongly contradict domain knowledge or expectation of various experts in as much as the same research findings are relied upon. A statistician may for example say advertising, capital, sales or number of transactions do not work based on a given model. On the other hand a marketing officer

may contradict this assertion, based on his experience or gut feelings. Our statistical modeling in this case has attempted to replace subjectivity with objectivity in business decision making processes in as far as quantitative variables are concerned .

We recommend that apart from the quantitative variables, further research be done with qualitative variables such as security or education level of retailers then a combination of both so as to fully capture the plight of retailers and the possible solutions.

References

[1] Alvin C.:*Methods of Multivariate Analysis*: (Second Edition): Wisley ,Canada(2002).

[2] Allen Webster:*Applied Statistics For Business & Economics an Essential Version*(Third Edition), McGrow Hill,Singapore (2005)

[3] Allen Webster:*Applied Statistics For Business & Economics*(First Edition) Irwic ..Inc, USA (1992)

[4] Amy Dahan Dalmedico:*History and Epistemology of Models on Meterology*,(1963)

[5] Ariel M. Aloe:*A Partial Effect Size For The synthesis of Multiple Regression Models*Florida, USA (2009).

[6] Anderson T.W.:*An Introduction To Multivariate Statistical Analysis*(Third Edition) Willy , New York (2004)

[7] Atkinson A.C:*Plots,Transformations and Regression:An Introduction to Graphical Methods of Diagnostic Reggression.*OUP England (2010)

[8] Belsley D.A *Regression:Identifying Influential Data and Sources of Collinearity*, Willy-Interscience, New York, USA (2004)

[9] Bowerman B.L:*Linear Statistical Models:An Applied Approach* Second Edition, Thompson Brookes/Cole, Belmont (2000).

[10] Cook R.D.:*Residual and Influence In Regression*, Chapman Hall, New York. (1982)

[11] David J.H.K:*Anatomy of Business Models: Asyntatical Review and Resaerch Agenda* London Business School (2010)

[12] Inmaculda Arostegui:*Statistical approaches to Analyze patient Reported Outcome As Response Variables: An application To health related Quality Of Life*, Sage Publications , United Kingdom. (2010)

[13] Mc Clave James T:*Annotted Instructors Edition(Statistics)*Seventh Edition,Prentice Hall,New Jersey,USA (1997).

[14] Rao C.R.:*Linear Statistical Inference and Its Application*, Second Edition, Willy-Interscience, New York (2002).

[15] Richard A.J :*Applied Multivariate statistical Analysis*,Sixth Edition, Pearson/Prentice Hall,USA (2007).

[16] Soong T*Fundamentals of Probability and statistics For Engineers* Willy ,New York USA (2004).

[17] Sweeney ..et..al: *Fundamentals of Business Statistics*,Fifth Edition. South Western Cengage Learning (2009).

[18] Williams S: *Business Process Modelling Improves Administrative Controls in Automation* (1967).

[19] Wolfgang, H:*Applied Multivariate Statistical analysis*(2003).