blx@bib1sorting

blx@biblist1sorting

blx@bib2prefixnumbers

blx@latem@fieldcompound blx@latem@fieldpcompound blx@latem@fieldpstrwidth

blx@refcontextprefixnumbersblx@refcontextlabelprefix=1

blx@bib1prefixnumbersprefixnumberlabelprefix

Forecasting Supermarket sales using a Generalized Linear Model

Kennedy Mwangi
and
Cynthia Chepkirui

A Project submitted in partial fulfilment of the requirements for the award of the Degree of

Bachelor of Science
in
Actuarial Science

Dedan Kimathi University of Technology

2022

Declaration by the Students

"We, *Kennedy Mwangi* and *Cynthia Chepkirui*, declare that this project entitled, *'Forecasting Supermarket sales using a Generalized Linear Model'* submitted in partial fulfilment of the degree of *Bachelor of Science in Actuarial Science*, is a record of original work carried out by us under the guidance of *Mr. Andrew Kinyita*, and has not formed a basis for the award of any other degree or diploma, in this or any other Institution or University. In line with the ethical practice in reporting scientific information, due acknowledgements have been made wherever the findings of others have been cited."

KENNEDY MWANGI
(S030-01-1572/2019)

Signature

Date

CYNTHIA CHEPKIRUI
(S030-01-1609/2019)

Signature

Date

Declaration by the Supervisor

This is to certify that the project entitled *'Forecasting Supermarket sales using a Generalized Linear Model'* submitted by *Kennedy Mwangi* and *Cynthia Chepkirui* to the Dedan Kimathi University of Technology, in partial fulfilment for the award of the degree of *Bachelor of Science in Actuarial Science*, is a bona-fide record of research work carried out by them under my supervision. The contents of this project, in full or in parts, have not been submitted to any other Institution or University for the award of any degree.

MR. ANDREW KINYITA
(Supervisor)

_____
Signature

_____
Date

DR. MAINA MUNDIA
(Project Coordinator)

_____
Signature

_____
Date

Acknowledgement

We would like to express our sincere gratitude to our supervisor, *Mr. Andrew Kinyita* for his excellent guidance during the times of doing the Project through helping us in doing and producing a good document. We would also like to appreciate the help of the lecturers from the department for helping out during the whole process of producing this work.

Dedication

Dedicated to our family, friends, lecturers and classmates who have really helped us out during the preparation of this project document.

Contents

List of Figures

List of Tables

Abbreviations

AIC	**A**kaike **I**nformation **C**riterion
AUC	**A**rea **U**nder the **C**urve
GLM	**G**eneralized **L**inear **M**odel
MLE	**M**aximum **L**ikelihood **E**stimator
MSE	**M**ean **S**quare **E**rror
OLS	**O**rdinary **L**inear **S**quares
pdf	**p**robability **d**ensity **f**unction
RMSE	**R**oot **M**ean **S**quare **E**rror
ROC	**R**eceiver **O**perating **C**haracteristic curve
se	**s**tandard **e**rror
VAR	**V**ector **A**utoregressive

Symbols

α	Level of significance
η	Linear predictor
ϕ	Dispersion parameter
χ^2	Chi square
$\hat{\beta}$	Estimated parameters
$g(\mu)$	Link function
$\hat{v}(\mu)$	Variance function
θ_i	Canonical Parameter
H_0	Null hypothesis
H_1	Alternative hypothesis
Σ	Summation
ϵ	Error term
L	Likelihood
Y	Dependent variable
X	Explanatory variable
∞	Optimal range

Abstract

It's vital for commercial enterprises to accurately forecast sales. Such predictive analytics is a crucial part of their decision support systems to increase the profitability of the company. In predictive data analytics, the branch of regression modeling is used to forecast numerical response variable like sales amount. In this category, linear models are simple and easy to interpret yet they permit generalization to very powerful and flexible families of models which are called Generalized linear models. The generalization potential oversimple linear regression can be explained. Generalized Linear Models relax the assumption of normally distributed error terms. Moreover, the relationship of the set of predictor variables and the response variable could be represented by a set of link functions rather than the sole choice of the identity function. The main objective of this study was to model sales amount forecast problem through the use of Generalized Linear Models. Company sales data are explored and the response variable, sales amount is fitted to the Gamma, Gaussian and Inverse Gaussian from the exponential family of distributions. The model selection was done using Akaike Information Criterion. The results showed that the Gamma distribution had the least Akaike Information Criterion hence was chosen. The model adequacy was assessed using the Chi-square goodness of fit at 5% significance level which led to the rejection of the null hypothesis that the model did not fit the data. This proved that the model was adequate enough to fit the data. The fitted Gamma model was then used to forecast the sales and the accuracy of the forecasted sales was checked using the mean square error and a value of 0.06 was obtained implying that the forecast did not disperse far from the observed values. In conclusion, the results showed that Generalized Linear model is reliable in sales forecasting and categorization of the predictor variables improve the model fitting results.

Chapter 1

Introduction

1.1 Introduction

Sales forecasting is the process of estimating future revenue by predicting the amount of the product a sales unit will sell in the next week, month, quarter, year or a certain period of time or a projected measure of how a market will respond to the company's go to market effect. Sales prediction plays an important role in many fields and helps improve the sales of a company by making future plans through predicting the sales of the company. Its an important prerequisite for enterprise planning and correct decision making allowing companies to better plan their business activities.

It is hard to overstate how important it is for the company to produce accurate sales forecasts. Privately held companies gain confidence in their business when leaders can trust the forecasts, while for publicly traded companies, accurate sales forecasts confers credibility in the market. Sales forecasts add value across the organization. Finance relies on forecasts to develop budgets for capacity planning and hiring, production uses sales forecasts to plan their circles, forecasts helps sales operations with territory and quota planning, supply chain with purchases and sales strategy with channels and partner strategies. An accurate sales forecasting process confers many benefits which may include: improved decision making about the future, reduction of sales pipeline and forecast risk alignment of sales quotas and revenue expectation, it's a benchmark that can be used in assessing future trends, ability to focus sales team on high revenue, high profit sales and pipeline opportunities resulting in improved win rates.

A major challenge to increasing sales lies in the ability to forecast sales patterns and know readily beforehand when to order and replenish inventories as well as plan for man power and staffs. The amount of sales data has steadily been an increase in recent years and ability to leverage this gold of data separates the high performing supermarkets from others. one of the most valuable assets a supermarket can have is data generated by customers as they interact with various supermarkets. Within these data lie important patterns and variables that can be modeled by a generalized linear model.

Previous studies on sales prediction has always used single prediction model that can perform best for all kinds of merchandise. The forecasts are generated using the flow of demand from the past as well as by considering other known factors in future.

1.2 Background of the Study

Regression is a statistical process of estimating relationship between two or more variables. Regression can be linear or nonlinear. In linear regression, the relationship is modeled by functions which are linear combination of variables. In regression one or more variables is used to predict another variable. The simplest form of regression involves two variables, the explanatory or independent variable used to predict another variable the response or the dependent variable. It is assumed that the two variables are linearly related.

This means that regression can tell us how much change we should expect in one variable if we alter the other by a certain amount. Where the response variable is linearly dependent on more than one explanatory variable. The assumptions made in a regression model includes; the mean of the error term is independent of the observed dependent variable,the error terms are uncorrelated and with a common variance that's independent. While linear models are practical for modeling real world phenomena because of their simplicity in training and model application, they assume normal distribution in the dependent variable and a linear impact of the independent variable on the dependent variable.

Generalized Linear Models are an extension of simple linear regression models, which predict the response variable as a function of multiple predictor variables, they are empirical transforms of the classical linear (Gaussian) regression model and are distinguished from ordinary least squares by particular model, rather than data, transformations: specifically, a response distribution of one of the exponential family of distributions (normal, Poisson, gamma, binomial, inverse Gaussian) and a (monotonic) link function (identity, logarithmic, square root, logistic, power) which relates the mean of the response to a scale on which the model effects combine additively, If there exists an appropriate link function for fitting the GLM, then the goodness of fit of the GLM may produce better results than that of the linear regression models. GLMs relax the assumption of normally distributed error terms. Moreover, the individual values of the response variable are independent from each other.

Some of the advantages of using the GLM include, the response variable can have any form of exponential distribution type, it is able to deal with categorical predictors, it's relatively easy to interpret

and allows a clear understanding of how each of the predictors are influencing the outcome and also less susceptible to over fitting. The Gamma, Poisson, inverse Gaussian and exponential distributions, which are members of the exponential family are widely used to model physical quantities that take positive values. Sale amount is such a quantity and can be modeled as a random variable.

1.3 Statement of the Problem

Every organization faces constant change in planning and decision making process of the business. To meet the needs of the organization, a type of forecast is needed. The more reliable the forecast, the better the results for planning and decision making. Forecasting has been a challenge in most managements. An efficient forecasting system is a requirement in the supply chain management which will in turn aid in handling demand shifts of the products and resources. Every firm's goal is to hold enough inventory to meet their customers' demand and reduce cost of buying and stocking the inventory.

1.4 Justification of the study

Most shortcomings in any business enterprise are as a result of poor decisions which stems from poor or inefficient forecasting methods. Most of the researches that have been conducted suggests the use of generalized linear models with discrete distributions as the preferred models to use in sales prediction. This study aims to build a GLM model with continuous distribution which can help to improve the accuracy of the predictions. The value of sales forecasts is twofold one, if one is able to identify that he or she is going to achieve or even exceed his or her target for a given period it gives him or her the ability to employ more staff, purchase inventory ahead of time and cut down on the delays that the organization might encounter down the line. Two, if one runs an accurate forecast and realize he might not hit the target, he or she may proactively execute the remediation plans.

1.5 Objectives of the Study

1.5.1 General objective

The general objective of the study was to forecast supermarket sales data using a Generalized Linear Model.

1.5.2 Specific objectives

The specific objectives of the study were;

i. To fit a Generalized Linear Model to the sales data.

ii. To check the adequacy of the fitted model.

iii. To forecast the sales using the fitted model.

iv. To check the accuracy of the forecasted sales.

1.6 Significance of the Study

Knowledge about the future is the one sure way to having a bright future. To forecast means knowing about the future and this can be very helpful to individuals and businesses. Some of the beneficiaries of this process includes; operations management, they will be aware that there will be different levels of demand for products for example during festive season and thus will be able to schedule for more production. Finance and risk management would use the forecast to see whether the potential sales will meet the level of returns they seek and determine how much they should spend and what salaries to pay workers. If marketing analysis of of past transaction data reveals that there will be high level of demand festive season they can be able to push for more advertisement during that time. Consumers will also benefit from the study because the supermarkets will now be offering the high demand goods which they need. This will also help the supermarkets to gain more customers hence making huge profits from sales and also get rid of costly rush orders and uneven level of inventory. The manufacturing companies of the high demand goods will also benefit since the supermarkets will now have a constant order of the required products hence leading to the avoidance of large productions which might lead to excess supply and if the products are perishable will lead to losses due to low demand after expiration. For the scholars, they will also benefit from this study since they will be able to gain some extra knowledge concerning the incorporation of the GLM model to forecast.

Chapter 2

Literature Review

2.1 Introduction

This chapter gives the review of studies that have been carried out in the past related to sales forecasts. The methodology that has been used in these studies is the generalized linear model.

2.2 Empirical Review

Karlsson, (2020) did study whose aim was to build a model that can help in predicting the sales quantities of different product classes and identify which factors are the most significant in the different models. Generalized linear models with a Negative binomial distribution and Poisson distribution was applied to retrieve the predicted sales quantity. The variables considered significant for the predicted outcome of the sales quantity for each product class in the models were: original price, purchase month, color, cluster, purchase country and channel. Residual analysis showed promising results for the negative binomial which turned out to be more desirable and proved to be a good fit for the data as compared to Poisson distribution. From the findings, it can be established that a generalized linear model can be used to predict future sales quantities of the different product classes.

Müller, (2018) carried out a study whose aim was to predict company sales demand. A GLM with gamma distributed dependent variable was adapted. A comparative analysis was performed with the linear model. Unique company sales data are explored and the response variable, sales amount was fitted to the Gamma distribution which is a member of the exponential family. The model fitting results confirmed that sales amount distribution is best fit to a gamma distribution. The model selection was performed via the use of mean squared error and Akaike Information Criterion metrics respectively. The results showed that GLM is better than the linear regression. Moreover, categorization on the predictor variables improves model fitting results significantly.

Zelingher et al., (2020) assessed the impacts of yearly variation of maize production and yields on maize prices using a generalized linear model with binomial distribution a member of the exponential family and logit link. The model computed the probability of price increase given a regional yield. The GLM was fitted for each month using yield changes as inputs. The most influential inputs were assessed using the AIC and the quantitative price estimation was assessed by root mean squared error

(RMSE). The accuracy of the predictions were evaluated using ROC curves (AUC). Most accurate predictions of the month of October were obtained with least RMSE and highest AUC revealing that this model was best for quantitative maize price prediction.

Yilmaz, (2020) did a research whose aim was to forecast house prices in Turkey and provide sufficient evidence in support of the adequacy of estimated prices for Turkey houses. Macroeconomic indicators related to houses such as gold, interest rates, and currency were considered. Generalized linear model and Vector AutoRegressive model were compared. The analysis identifies forecasts of housing market index from the generalized linear model as accurate compared to VAR method based on R-squared and RMSE values. Turkey's housing market showed high dependence on the macro-economic indicators and hence GLM proved to be a better model.

Lasek et al., (2019) did a study on restaurant sales and customer demand forecasting. The main aim of the study was to get accurate sales and customer demand forecasts. Sales forecasting is crucial for an independent restaurant and for restaurant chains. The main problem that was being focused on was the issue of getting a good methodology for different kinds of analytical methods. With a good model, the sales that could be forecasted would be close to accurate and would help the restaurant to be at par in planning their management and also in keeping the business afloat. A generalized linear model was adopted. Although no single method is best in every situation, at the end of the study, the Poisson model gave the best results that were quite adequate in the forecasting.

Elliot, (2019) did a study to compare a generalized linear model to an ordinary linear model to predict stock prices. A generalized linear model with normal distribution and a log link function was adopted in the analysis. It was found that the linear model performed worst when the number of lags were increased and the model prediction diverged greater from the real data. Generalized linear model performed better in prediction with lower RMSE compared to the linear model. Empirical examinations of the forecasting precision for the stock prices showed that the proposed GLM improved the forecasts implying that GLM is a better model.

Sazontyev, (2018) did a study on the sales forecast in euros at 1115 stores owned by Rossmann, a European phamaceutical company. The main aim of the study was to give the best sales forecasts. The methodology that was used was the GLM and an algorith of a feed foward neutral net. It was found that the feed forward neutral net algorithm produced poor results while the GLM provided

better results. The GLM model was preferred as an appropriate model for forecasting sales according to the study.

Dane, (2018) did a research on price forecasts of apartments. The general purpose of the study was to develop a statistical model to forecast the listing prices of apartments in South Africa. Residential property is an important segment of property market in South Africa. And the large portfolio of residential property contributes significantly towards the wealth of the country where it's capitalized on the household balance sheet in the set of national accounts. Residential property transactions are typically infrequent and relate to a highly differentiated set of items, rendering effective measurement techniques complex and difficult. This study develops a technical price function using a generalized linear model based on the gamma distribution and log-link function. The generalized linear models use the iterative reweighted least squares algorithm to obtain maximum likelihood estimates of model parameters for observations that belong to an exponential distribution family, where the systematic effects can be made linear through a link function. The reason for using generalized linear models over the ordinary least squares is to correctly account for the error structure and through the appropriate link function, the standardized deviance residuals should be homogeneous. From the study, it was concluded that the gamma distribution is suitable for non-negative continuous data. The tests that were carried out showed that the GLM gamma distribution produced close to accurate forecasts rendering it effective in the forecasting of the listing prices of the apartments.

The reviewed literature showed the use of generalized linear models to predict sales. Previous studies clearly indicates that these methods are adequate in the prediction of sales for various institutions though most failed to address the problems of; low accuracy, inconsistency and redundancy and hence were not able to give the best forecasts. To address these gaps, this study built a GLM model with various continuous distributions that would help in getting more accurate sales forecasts. The Akaike information Criterion was used to compare the distributions and the Gamma model was chosen. The adequacy was checked using Chi-square and this showed that the chosen Gamma model was adequate. Model accuracy was assessed using the mean square error and it proved that the forecasts were accurate which implied that the model was able to address the problem of low accuracy.

Chapter 3

Methodology

3.1 Introduction

This chapter presents the methodology of the study. It gives a description of the definitions and notations used in generalized linear models. Section 3.2 introduces the generalized linear model, assumptions and other concepts used in the generalized linear models. Section 3.3 presents the estimation of the generalized linear model parameters. Section 3.4 assesses the fit of the generalized linear mode. Section 3.5 covers the the prediction using the fitted generalized linear model. Section 3.6 involves checking of the accuracy of the predicted sales.

3.2 Generalized Linear models

The generalized linear model is a generalized form of linear models that extends the scope of linear models to allow for non-normality in the response variable. It provides a mean for modeling the relationship between one or more explanatory variables and a response variable whose distribution can be expressed in the form:

$$Y_i = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \varepsilon \tag{3.2.1}$$

where $\boldsymbol{Y}_i$ is the response variable, $\boldsymbol{X}_i$ is the explanatory variable.

A GLM is used for determining the relationship between the mean of the response variable and the covariates. By setting the link function g(μ)=η, then, assuming that the link function is invertible, the mean μ can be made as the subject of the formula:

$$\mu = g^{-1}(\eta) \tag{3.2.2}$$

A GLM model follows certain assumptions such as: the linear component is retained, distributions are restricted to the exponential dispersion family, response variables must be independent, the mean response changes with the conditions; but the functional shape of the distribution remains fundamentally unchanged and the mean response changes in some linear way as the conditions change. A GLM consists of three components;

3.2.1 A random component

A random component specifies the conditional distribution of the response variable, $\boldsymbol{Y_i}$ (for the ith of n independently sampled observations), given the values of the explanatory variables in the model.

The Exponential Family

In Nelder and Wedderburn's original formulation (Nelder et al., 1972), the distribution of $\boldsymbol{Y_i}$ is a member of an exponential family, such as the Gaussian (normal), binomial, Poisson, gamma, or inverse-Gaussian families of distributions. Glm has been extended to multivariate exponential families (such as the multinomial distribution),to certain non-exponential families (such as the two-parameter negative-binomial distribution), and to some situations in which the distribution of $\boldsymbol{Y_i}$ is not specified. The probability mass function of the exponential family of distributions is defined in the following form

$$f(y;\theta,\phi) = \exp\left\{\frac{y\theta - b(\theta)}{\phi} + c(y,\phi)\right\} \tag{3.2.3}$$

where θ is the canonical parameter that depends on regressors via the linear predictor further, ϕ is the dispersion parameter often known. The functions a(.), b(.) and c(.) are known and determine which member of the family is used. The log likelihood of equation (3.2.5) expressed as a function ofθ_i andϕ is given by

$$l(\theta_i,\phi;y_i) = log f_y(y_i;\theta_i,\phi) \tag{3.2.4}$$

the log likelihood is given by

$$l(\theta_i,\phi;y_i) = \left\{\frac{y\theta_i - b(\theta_i)}{\phi} + c(y,\phi)\right\} \tag{3.2.5}$$

Expressions for the first and second derivatives of the log likelihood in terms of the mean and variances ofy_i and the function of $a(\phi)$ are required. The mean and variance can be derived using the results

$$E\left(\frac{\partial l(\theta_i,\phi;y_i)}{\partial\theta_i}\right) = 0 \tag{3.2.6}$$

$$E\left(\frac{\partial^2 l(\theta_i,\phi;y_i)}{\partial\theta_i^2}\right) = -E\left(\frac{\partial l(\theta_i,\phi;y_i)}{\partial\theta_i}\right)^2 \tag{3.2.7}$$

First by the partial derivative of the log likelihood with respect to θ_i, we obtain

$$E\left(\frac{\partial l(\theta_i,\phi;y_i)}{\partial \theta_i}\right) = \frac{y_i - b'(\theta_i)}{a(\phi)} \tag{3.2.8}$$

where $'$ denotes differentiation with respect to θ_i. By taking the expectation and setting the equation to zero,we have

$$E\left(\frac{y_i - b'(\theta_i)}{a(\phi)}\right) = 0 \tag{3.2.9}$$

$$\frac{\mu_i - b'(\theta_i)}{a(\phi)} = 0 \tag{3.2.10}$$

consequently, the mean is:

$$E(y_i) = b'(\theta_i) = \mu_i \tag{3.2.11}$$

for $a(\phi) \neq 0$

The variance can be derived by taking the second partial derivative of the log likelihood function with respect to θ_i, to obtain

$$\frac{\partial^2 l(\theta_i,\phi;y_i)}{\partial \theta_i^2} = \frac{-b''(\theta_i)}{a(\phi)} \tag{3.2.12}$$

substituting in equation (3.2.14) and equation (3.2.11) into equation(3.2.9)

$$E\left(\frac{-b''(\theta_i)}{a(\phi)}\right) = E\left(\frac{y_i - b'(\theta_i)}{a(\phi)}\right)^2 \tag{3.2.13}$$

Reordering and using $var(y_i) = E[(y_i - E(y_i))^2]$ where $E(y_i) = b'(\theta_i)$

$$\frac{-b''(\theta_i)}{a(\phi)} = \frac{-var(y_i)}{a(\phi)^2} \tag{3.2.14}$$

thus,

$$var(y_i) = b''(\theta_i)a(\phi) = V(\mu_i)a(\phi) \tag{3.2.15}$$

Gamma distribution

The gamma distribution can take on a pretty wide range of shapes and given the link between the mean and the variance through its two parameters. It deals with heteroskedasticity in non-negative

data. It works well for positive-only data with positively-skewed errors. The log link can represent an underlying multiplicative process.

The probability density function of a gamma distribution is as follows:

$$f(x;\alpha,\beta) = \begin{cases} \frac{X^{\alpha-1}}{\gamma(\alpha)\,\beta^{\alpha}}\ exp(\frac{-x}{\beta}) & \alpha > 0,\ \beta > 0,\ x > 0 \end{cases} \tag{3.2.16}$$

To estimate the parameters of the gamma distribution that best fits this sampled data,the maximum likelihood of the parameters can be used.

From equation (3.2.16) the log-likelihood can be obtained resulting in

$$\begin{aligned} log(f(x;\alpha,\beta)) &= (\alpha-1)\sum logx_i - nlog\Gamma\alpha - n\alpha log\beta - \frac{1}{\beta}\sum x_i \\ &= n(\alpha\ -1)logx - nlog\Gamma\alpha - n\alpha log\beta - \frac{n\bar{x}}{\beta} \end{aligned} \tag{3.2.17}$$

The MLE which maximizes (3.2.16) is given as;

$$\begin{aligned} &\frac{d}{d\alpha}log(f(x;\alpha,\beta)) = 0 \\ &log(f(x;\alpha,\beta)) = n\alpha log\beta - nlog\Gamma(\alpha) + (\alpha-1)\sum_{i=1}^{n} logx_i - \beta\sum_{i=1}^{n} x_i \\ &0 = nlog\beta - \frac{n}{\Gamma(\alpha)}\Gamma'(\alpha) + \sum_{i=1}^{n} logx_i \\ &\phi(\alpha) = \frac{\Gamma'(\alpha)}{\Gamma(\alpha)} \\ &\phi(\alpha) = \hat{\alpha} \\ &\hat{\alpha} = log\beta + \frac{1}{n}\sum_{i=1}^{n} logx_i \end{aligned} \tag{3.2.18}$$

$$\begin{aligned} &\frac{d}{d\beta}log(f(x;\alpha,\beta)) = 0 \\ &0 = \frac{n\alpha}{\beta} - \sum_{i=1}^{n} x_i \\ &\hat{\beta} = \frac{\hat{\alpha}}{\frac{1}{n}\sum_{i=1}^{n} x_i} \\ &\hat{\beta} = \frac{\hat{\alpha}}{\bar{x}} \end{aligned} \tag{3.2.19}$$

The mean of the Gamma distribution can be obtained as follows;

$$\begin{aligned} E(X) &= \int_0^\infty x f_X(x) dx \\ &= \int_0^\infty x \cdot \frac{\lambda^\alpha}{\Gamma\alpha} x^{\alpha-1} e^{-\lambda x} \mathrm{d}x \\ &= \frac{\lambda^\alpha}{\Gamma(\alpha)} \int_0^\infty x \cdot x^{\alpha-1} e^{-\lambda x} \mathrm{d}x \\ &= \frac{\lambda^\alpha}{\Gamma(\alpha)} \int_0^\infty x^\alpha e^{-\lambda x} \mathrm{d}x \\ &= \frac{\lambda^\alpha}{\Gamma(\alpha)} \frac{\Gamma(\alpha+1)}{\lambda^{\alpha+1}} \\ &= \frac{\alpha\Gamma(\alpha)}{\lambda\Gamma(\alpha)} \\ &= \frac{\alpha}{\lambda}. \end{aligned} \tag{3.2.20}$$

Similarly the variance can be obtained by getting the second moment and subtracting the square of the mean. The second moment can be obtained as follows;

$$\begin{aligned} E(X^2) &= \int_0^\infty x^2 \mathrm{d}x \\ &= \int_0^\infty x^2 \cdot \frac{\lambda^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\lambda x} \mathrm{d}x \\ &= \frac{\lambda^\alpha}{\Gamma(\alpha)} \frac{\Gamma(\alpha+2)}{\lambda^{\alpha+2}} \\ &= \frac{(\alpha+1)\alpha\Gamma(\alpha)}{\lambda^2\Gamma(\alpha)} \\ &= \frac{\alpha(\alpha+1)}{\lambda^2}. \end{aligned} \tag{3.2.21}$$

The variance can be obtained as;

$$\begin{aligned} Var(X) &= E(X^2) - (E(X))^2 \\ &= \frac{\alpha(\alpha+1)}{\lambda^2} - \frac{\alpha^2}{\lambda^2} \\ &= \frac{\alpha}{\lambda^2}. \end{aligned} \tag{3.2.22}$$

Gaussian distribution

The Gaussian model is a generalized linear model form of regression analysis used to model continuous data. The Gaussian model assumes the response variable has a Gaussian distribution.

The Gaussian distribution is of the form;

$$f(x;\mu,\sigma) = \frac{1}{\sqrt{2\pi}\sigma}\, exp\left(-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right)\ \mu,\sigma > 0,\ -\infty < x < \infty \tag{3.2.23}$$

The likelihood of the Gaussian distribution can be obtained by the product of the functions;

$$L(x;\mu,\sigma) = \prod_{i=1}^{n}\left(\frac{1}{\sqrt{2\pi}\sigma}\, exp\left(-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right)\right) \tag{3.2.24}$$

The maximum likelihood estimates are given by;

$$\begin{aligned}\hat{\mu_n} &= \frac{1}{n}\sum_{j=1}^{n} X_j \\ \hat{\sigma_n^2} &= \frac{1}{n}\sum_{j=1}^{n}\left(X_j - \hat{\mu}\right)\end{aligned} \tag{3.2.25}$$

The mean of the Gaussian distribution is given by;

$$\begin{aligned}E(X) &= \int_{-\infty}^{\infty} xf(x)dx \\ &= \frac{1}{\sigma\sqrt{2\pi}}\int_{-\infty}^{\infty} xexp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)dx \\ &= \frac{1}{\sqrt{\pi}}\left(\sqrt{2\pi}\int_{-\infty}^{\infty} texp(-t^2)dt + \mu\int_{-\infty}^{\infty} exp(-t^2)dt\right) \\ &= \frac{\mu\sqrt{\pi}}{\sqrt{\pi}} \\ &= \mu\end{aligned} \tag{3.2.26}$$

The variance of the Gaussian distribution is given by;

$$\begin{aligned}Var(X) &= \int_{-\infty}^{\infty} x^2 f(x)dx - (E(X))^2 \\ &= \frac{1}{\sigma\sqrt{2\pi}}\int_{-\infty}^{\infty} x^2 exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)dx - \mu^2 \\ &= \frac{2\sigma^2}{\sqrt{\pi}}\int_{-\infty}^{\infty} t^2 exp(-t^2)dt \\ &= \frac{2\sigma^2}{\sqrt{\pi}}\frac{1}{2}\int_{-\infty}^{\infty} exp(-t^2)dt \\ &= \sigma^2\end{aligned} \tag{3.2.27}$$

Inverse Gaussian distribution

The inverse Gaussian distribution (also known as the Wald distribution) is a two-parameter exponential family of continuous probability distributions with support on (0,∞). X$\sim$ IG(μ,λ) can be written to indicate that a random variable is inverse Gaussian-distributed with mean μ and shape parameter λ.

The distribution function for the Inverse Gaussian distribution can be written as;

$$f(x;\mu,\lambda) = \sqrt{\frac{\lambda}{2\pi x^3}} exp\left(\frac{-\lambda(x-\mu)^2}{2\mu^2 x}\right) \tag{3.2.28}$$

The model where $X_i \sim \text{IG}(\mu, \lambda w_i)$ with all w_i known, (μ,λ) unknown and all X_i independent has the likelihood;

$$L(\mu,\lambda) = \left(\frac{\lambda}{2\pi}\right)^{n/2} \left(\prod_{i=1}^{n} \frac{w_i}{X_i^3}\right)^{1/2} exp\left(\frac{\lambda}{\mu}\sum_{i=1}^{n} w_i - \frac{\lambda}{2\mu^2}\sum_{i=1}^{n} w_i X_i - \frac{\lambda}{2}\sum_{i=1}^{n} w_i \frac{1}{X_i}\right) \tag{3.2.29}$$

Solving this yields;

$$\begin{aligned} \hat{\mu} &= \frac{\sum_{i=1}^{n} w_i X_i}{\sum_{i=1}^{n} w_i} \\ \frac{1}{\hat{\lambda}} &= \frac{1}{n}\sum_{i=1}^{n} w_i \left(\frac{1}{X_i} - \frac{1}{\hat{\mu}}\right) \end{aligned} \tag{3.2.30}$$

3.2.2 The linear predictor

That is a linear function of regressors

$$\eta = \beta_0 + \beta_1 X_1 + \cdots + \beta_{p-1} X_{p-1} \tag{3.2.31}$$

As in the linear model, and in the logit and probit models, the regressors X_{ij} are prespecified functions of the explanatory variables and therefore may include quantitative explanatory variables, transformations of quantitative explanatory variables, polynomial regressors, dummy regressors, interactions, and so on. Indeed, one of the advantages of GLMs is that the structure of the linear predictor is the familiar structure of a linear model.

3.2.3 Link function

A smooth and invertible linearizing link function $g(.)$, which transforms the expectation of the response variable, $\mu_i = E(Yi)$, to the linear predictor:

$$g(\mu) = \beta_0 + \beta_1 X_1 + \ldots + \beta_{p-1} X_{p-1} \tag{3.2.32}$$

Some common link functions include;

TABLE 3.1: Common link functions for the GLM models

link	$\eta = g(\mu_i)$	$\mu_i = g^{-1}\eta_i$
Identity	μ_i	η_i
Log	$log_e \mu_i$	e^{η_i}
Inverse	μ_i^{-1}	η_i^{-1}
Inverse-square	μ_i^{-2}	$\eta_i^{-1/2}$
Square root	$\sqrt{(}\mu_i)$	η_i^2
Logit	$log_e \frac{\mu_i}{1-\mu_i}$	$\frac{1}{1+e^{\eta_i}}$
Probit	$\phi^{-1}(\mu_i)$	$\phi(\eta_i)$

Because the link function is invertible, we can also write

$$\mu_i = g^{-1}(\eta_i) = g^{-1}(\beta_0 + \beta_1 X_1 + \ldots + \beta_{p-1} X_{p-1}) \tag{3.2.33}$$

and, thus, the GLM may be thought of as a linear model for a transformation of the expected response or as a nonlinear regression model for the response. The inverse link $g^{-1}(.)$ is also called the *mean function*. Note that the identity link simply returns its argument unaltered, η_i=$g(\mu_i)$=μ_i and thus μ_i=$g^{-1}(\eta_i)$=η_i

3.3 Parameter Estimation of the fitted GLM model

The parameters of the generalized linear model will be estimated using the maximum likelihood method. When estimating the parameters, the intention is to search for the values that maximize the log likelihood. Likelihood is written as $L(\theta|y)$ and is considered as a function of θ, which gives the probability how likely the parameters θ are for the observed data. The estimation is for the values that

maximize this probability. The set of observations $y = (y1, y2, ..., yn)$ are independently distributed with a distribution from the exponential family.

$$f(y;\theta,\phi) = exp\left(\frac{y\theta - b(\theta)}{a(\phi)} + c(y,\phi)\right) \tag{3.3.1}$$

The likelihood function can be written as the product of the distributions:

$$L(\theta,\phi) = f(\theta,\phi|y) = \prod_{i=1}^{n} f_i(\theta,\phi|y) \tag{3.3.2}$$

and on the logarithm scale this independence gives us an additive property. The log likelihood can be written as:

$$l(\theta,\phi) = log\ f_y(\theta,\phi|y) = \sum_{i=1}^{n} log\ f_i(\theta,\phi|y) \tag{3.3.3}$$

3.4 Model selection

This section introduces the method that will be used in choosing the best distribution to use.

3.4.1 Akaike information criterion (AIC)

The AIC is a measure of fit that is often used as a means for choosing between models is the following expression

$$AIC = -2\ln(L) + 2q \tag{3.4.1}$$

where L is referred to the maximum value of the likelihood function for the model and again, q is the number of parameters included in the model. This measure can be used when deciding if or not to include an explanatory variable in a model since the model with the lowest AIC value is preferred over larger values when comparing full and nested model, where one or more variables which are included in the full model have been removed.

3.5 Assessing the fit of the model

When assessing the fit of a statistical model, the model is evaluated on the discrepancy between the observed values and the predicted values, ergo, how well the model results corresponds to the true values. The measures of assessing the fit of the model chosen for this study is a Wald test, Pearson χ^2 statistics, Akaike's Information Criterion and multicollinearity.

3.5.1 Chi-square goodness of fit test

The way of testing the goodness of fit of the model is the generalized Pearson χ^2 test, which has the following definition

$$\chi^2 = \sum_{i=1}^{n} \left(\frac{(O_i - E_i)^2}{E_i} \right) \tag{3.5.1}$$

The hypothesis of the chi-square test can be written as;

$$\begin{aligned} H_0 &: The\ model\ does\ not\ fit\ the\ data \\ H_1 &: The\ model\ fits\ the\ data \end{aligned} \tag{3.5.2}$$

where, **O** is the observed value and **E** is the expected value. The value of χ^2 is then compared against the critical value of χ_c^2 on a given significance level and with given degrees of freedom. If the critical value is less than the calculated value, then the null hypothesis that the distribution is true can be rejected, hence the model does not fit the data well based on the given level of significance.

3.6 Prediction using the fitted model

The acquired supermarket sales data includes the total sales of a certain product. Hence, this can be referred to as a response variable that counts the occurrences of a specific event, and therefore the data can be considered as count data since the observations of the sold quantity is restricted to non-negative integer values. The collected data needs to be restructured in order to be able to perform the analysis. The independent variables to be included in the model will be chosen out of the total collected data. Subsequently, each of those explanatory variables will be clustered together in order for better classifying the characteristics for the variables. The equation of the model that will produce the forecast will follow the form of equation (3.6.1)

$$\hat{P}(Y|X) = \frac{exp(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p)}{1 + exp(-(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p))} \tag{3.6.1}$$

3.7 Checking the Accuracy of predicted sales

This section presents a way in which the predicted sales are analyzed to check their accuracy. The mean squared error will be used to assess the accuracy of the forecasts and the observed observations. The MSE is calculated as follows;

$$MSE = \frac{1}{n} \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 \tag{3.7.1}$$

where Y_i are the observed values and $\hat{Y}_i$ are the predicted values. The MSE measures how close the regression line is to a set of data points. A large MSE indicates that there is a wide dispersion of data around its mean while a smaller MSE indicates that the data are closely dispersed around the mean. A smaller MSE is preferred since it shows that the forecasted values are more accurate than in a larger MSE.

Chapter 4

Results and Discussions

4.1 Introduction

This chapter presents the results and discussion of the findings of the research. Section 4.2 gives the descriptive statistics of the supermarket sales data. Section 4.3 shows the estimates of the fitted generalized linear models. Section 4.4 involves the selection of the best GLM model. Section 4.5 deals with the adequacy of the fitted generalized linear model. Section 4.6 shows the prediction of the sales using the fitted GLM model. Section 4.7 deals with the accuracy of the sales forecasts.

4.2 Data source and description

The data to be used in this study is readily available in the Kaggle website under the link $https://www.kaggle.com/datasets/aungpyaeap/supermarket-sales$ consisting of 1000 observations. The dataset is one of the historical sales data of supermarket company which has recorded in 3 different branches for 3 months. The variables consist of 10 variables; customer type; either a member or normal, the gender; male or female, the unit price of the goods purchased, the quantity, tax at 5%, the total amount of the purchased items, cost of goods sold, gross margin, gross income and the rating that was provided by the serviced customer.

TABLE 4.1: Descriptive statistics of supermarket sales

	min	Std	Mean	Median	Max	Kurtosis	Skewness
Unit Price	10.1300	26.4700	55.7500	55.7500	99.9600	1.7600	0.00707
Quantity	1.0000	2.9100	5.5500	5.0000	10.0000	1.8000	0.0129
Tax	0.5085	11.7300	15.5000	12.0880	49.6500	2.9100	0.8912
Rating	4.0000	1.7200	6.9730	7.0000	10.0000	1.8400	0.0090
cogs	10.1700	234.1800	307.5900	993.0000	241.7600	2.9100	0.8900
Gross margin %	4.7600	0.0000	4.7600	4.7600	4.7600	0.0000	0.0000
Gross income	0.5100	11.7100	15.3800	49.6500	12.0900	2.9100	0.8900

Table 4.1 shows the measure of central location of each variable using the mean where values of each variable are added up and divided by the total sample space.Similarly, standard deviations, skewness,

kurtosis, minimum and maximum were used to measure the variability of the data. The standard deviation measures the dispersion of a subject set of data from the mean. The higher the standard deviation, the more dispersed is the subject set of data from the mean. For example, the standard deviation of the unit price indicates that the dispersion of the price is 26.47 times from the mean. The results of the skewness show that all the other variables apart from the rating are positively skewed. This implies that their right-hand tails are longer than left-hand tails implying that the data was not normally distributed. In addition to that, all the variables exhibit leptokurtic distribution since their kurtosis are greater than zero.

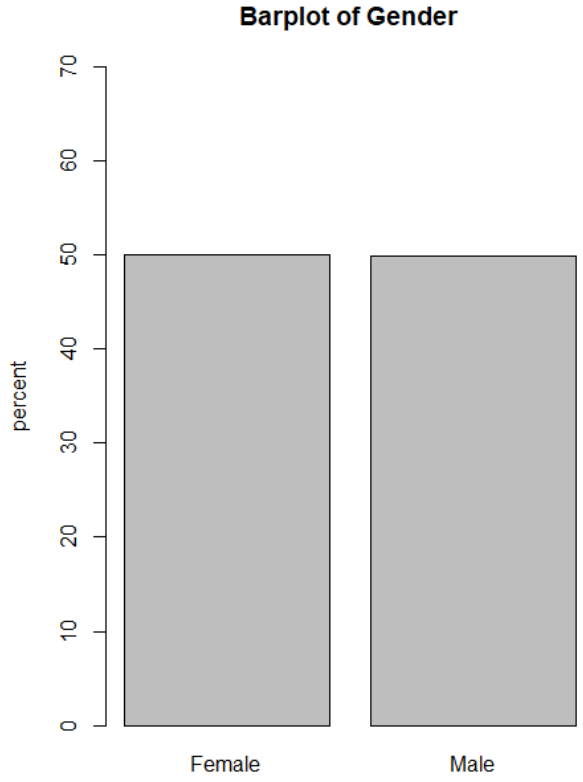

FIGURE 4.1: Gender statistics barplot

Figure 4.1 shows a barplot of the gender. It shows that the males have a percentage of 49.9%; a total of 499 and the females have a percentage of 50.1%; a total of 501.

4.3 The estimates of the fitted GLM models

The parameters of the fitted GLM model are estimated using the maximum likelihood estimation method. The fitting was done and summaries where total sales is the response variable. A generalized linear model with Gamma distribution was fitted to the data and summary of estimates was obtained.

TABLE 4.2: Parameter estimates of the fitted Gamma model

Variable	Parameter	se(parameter)	p-value	exp(parameter)
Intercept	3.0425	0.0559	0.0000	20.9573
Unit price	0.0209	0.0559	0.0000	1.0211
Quantity	0.2220	0.0070	0.0000	1.2485
Tax	0.0003	0.0023	0.8800	1.0003
Rating	0.0039	0.0051	0.4400	1.0039
Gender	0.0064	0.0175	0.7160	1.0064

Table 4.2 shows the estimated parameters of the total sales using GLM with Gamma model. The coefficient estimate in the output indicates the average change in the log odd of the response variable with a unit increase in each predictor variable for example, a unit increase in the predictor variable unit price is associated with an average change of 0.0208 in the log odds of the response variable. This means higher values of the unit price are associated with a higher likelihood of the response variable. The standard error gives an idea of the variability associated with coefficient estimate. This essentially tells us how well each predictor variable is able to predict the value of the response variable in the model. The variables unit price, quantity and tax are significant at 5% level. The coefficients for unit price, quantity, rating, gender and tax estimates are positive indicating positive correlation with the response variable.

TABLE 4.3: Parameter estimates of the fitted Inverse Gaussian model

Variable	Parameter	se(parameter)	p-value	exp(parameter)
Intercept	0.0215	0.0005	0.0000	1.0218
Unit price	-0.0002	0.0000	0.0000	0.9998
Quantity	-0.0019	0.0001	0.0000	0.9981
Tax	0.0003	0.0000	0.0000	1.0003
Rating	-0.0000	0.0000	0.4660	1.0000
Gender	-0.0000	0.0001	0.7610	1.0000

Table 4.3 shows the estimated parameters of the total sales using a GLM with Inverse Gaussian distribution. The p-value is used in checking the significant variables. The variables; unit price, quantity and tax are significant at 5% level of significance. The significant variables shows there is correlation

with the response variable. The coefficients for unit price, quantity, rating and gender are negative implying negative correlation with the response variable while the coefficient for tax estimate is positive indicating positive correlation with the response variable. The standard error provides the absolute measure of the typical distance that the data points fall from the regression line.

TABLE 4.4: Parameter estimates of the fitted Gaussian model

Variable	Parameter	se(parameter)	p-value	exp(parameter)
Intercept	0.0113	0.0004	0.0000	1.0113
Unit price	-0.0001	0.0000	0.0000	0.9999
Quantity	-0.0008	0.0001	0.0000	0.9992
Tax	0.0001	0.0000	0.0000	1.0000
Rating	0.0000	0.0000	0.8320	1.0000
Gender	0.0000	0.0001	0.8960	1.0000

Table 4.4 shows the estimated parameters of the total sales using GLM with Gaussian distribution. The coefficient estimate in the output indicates the average change in the response variable with a unit increase in each predictor variable for example, a unit increase in the predictor variable unit price is associated with an average change of -0.0001 in the response variable. This means that the higher values of the unit price are associated with lower values of the response. The variables unit price, quantity and tax are significant at 5% significance level. The coefficients for unit price and quantity are negative implying negative correlation with the response variable while the coefficient for tax estimate is positive indicating positive correlation with the response variable. The standard error gives an idea of the variability associated with coefficient estimate. For example the variability associated with unit price is 0.004.

4.4 Model selection

The AIC is a metric used to compare the fit of different regression models. The lower the value the better the regression model is able to fit the data.

TABLE 4.5: AIC values for selecting the model

	Gamma	Inverse Gaussian	Gaussian
AIC	8846	10183	9489

Table 4.5 shows the Akaike Information Criterion values for the three generalized linear models that is Gamma, Inverse Gaussian and Gaussian models. It can be seen that the model with inverse Gaussian has the highest value and the model with Gamma has the least AIC. This further implies that the Gamma model is considered to be a better fit for the data.

4.5 Adequacy of the model

To assess the adequacy of the fitted model, a Chi-square goodness of fit test at 5% significance level was used. The model has a chi-square statistic 149019 and 799 degrees of freedom. The chi-square distribution returns a p-value of zero. Consequently the null hypothesis, in equation (3.5.2) is rejected at 5% level of significance that the deviation between the observed and the expected is not significantly large and thus the model is a good fit to the data.

4.6 Prediction using the fitted model

A GLM gamma model was chosen to forecast the sales. The gamma model took the equation of the form;

$$\hat{P}(Y|X) = \frac{exp((3.0420 + 0.0210X_1 + 0.2220X_2 + 0.0003X_3 + 0.0039X_4 + 0.0064X_5))}{1 + exp(-(3.0420 + 0.0210X_1 + 0.2220X_2 + 0.0003X_3 + 0.0039X_4 + 0.0064X_5))} \quad (4.6.1)$$

where g is the associated link function. The Gamma GLM model was used to predict the sales and a sample of the predictions are given.

TABLE 4.6: Actual and Predicted sales values

No.	Actual	Predicted
801	144.963	133.514
802	253.680	238.089
803	495.317	421.726
804	462.672	446.291
805	714.326	781.299
806	325.374	262.224
807	195.678	236.977
808	210.966	151.038

Table 4.6 shows the first eight values of the predicted and actual sales using GLM with Gamma model. The values show that the model generated values that reflect the actual values; implying that the forecasts were accurate.

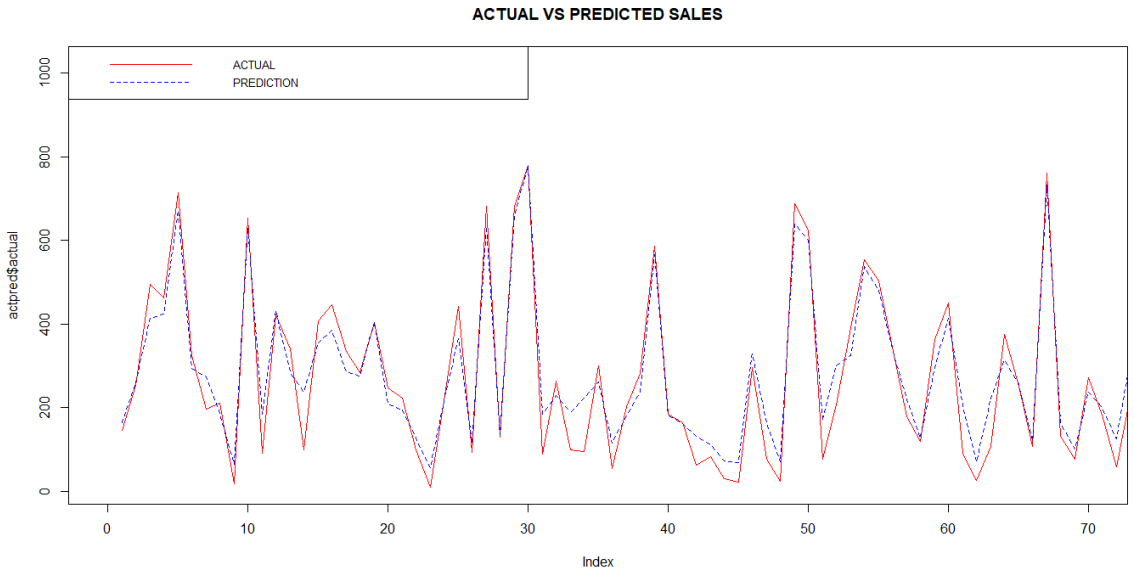

FIGURE 4.2: Plot of Actual and predicted sales

Figure 4.2 shows the plot of the actual and forecasted sales. The plot shows that the sales forecast are not far from the actual observed sales. This shows that the GLM Gamma model fitted was a good choice for the sales data.

4.7 The accuracy of the sales forecast

The accuracy of the sales forecast was checked using the mean square error formula in equation (3.7.1). Mean square error measures the average of squares of errors. A large MSE indicates wide dispersion from the actual values while lower values indicates closeness to the actual values which further implies the prediction model is accurate. An MSE value of 0.0603 was obtained which shows that the forecast were accurate.

Chapter 5

Summary and Conclusion

5.1 Introduction

This chapter gives the summary and conclusions made according to the obtained results of the sales forecast and also the recommendations of the study. Section 5.2 gives the summary of the results. Section 5.3 gives the conclusions drawn from the findings of this study. Section 5.4 outlines the recommendations suggested for further research.

5.2 Summary

The generalized linear models offers certain advantages over ordinary linear model: the separate specification for distribution and link function offers flexibility for achieving linearity, the link functions and its inverse functions allows for interpretation in both the transformed scale and original scale, it allows for specifications of distributions that explicitly model non normality when strict statistical assumptions are met and finally, allows for variety of research designs with any number of potential predictors.

This study focused on fitting generalized linear models to the supermarket sales using the Gaussian, Gamma and Inverse Gaussian distributions;the Gaussian and Gamma both having the log link function and the Inverse Gaussian having the inverse link function. After the models were fitted, a suitable had to be chosen. The Akaike Information Criterion was used in selecting a suitable model from the three model fits. The Gamma model proved to be the most suitable model since it had the least AIC. The Gamma model fit coefficients were all positive indicating positive correlation to the response variable; sales. So as to use the model in forecasting, the model's adequacy had to be assessed.

A chi-square goodness of fit test was carried out at a 5% level of significance to assess the Gamma model's adequacy. The chi-square test proved that the Gamma model was adequate to carry out the forecast. The model produced forecasts which were then assessed using the Mean square Error. The MSE showed that there was minimal error in the forecast hence the forecast was accurate.

5.3 Conclusion

This study aimed at forecasting supermarket sales. These sales forecasts play an important role in helping the supermarkets to run and manage their day to day activities. Apart from supermarkets, the study can also be applied to other commercial institutions that deal with sales. From the findings a generalized linear model proved to be reliable in sales forecasting since it was easy to use, handled the data easily and it gave forecasted sales values which were a reflection of the actual sales values.

5.4 Recommendations for Further Research

This study was based on supermarkets sales. A similar study is recommended for other commercial firms. Also, other exponential distributions instead of Gamma, Gaussian and Inverse Gaussian can be applied together with other predictor variables to evaluate their impact on total sales. One difficulty that reoccurred on several occasions during the pre-phase of the study was to handle the flaws of the structure of the initial data thus it recommends that constructing a better way of sorting and cleaning the initial data not only be time consuming actions be reduced but also the building of the model will benefit. Since shifts of customer interests is disregarded in this study and considering the rapid movements in the sector, for further research this study recommends adding more flexibility regarding behavior changing factors such as existing and future trends, campaigns and marketing when modeling.

References

Dane, Bax (2018). “Listing price estimation of apartments: A generalised linear model”. In: *Journal of Economic and Financial Sciences* 12.1, pp. 1–11.

Elliot, Cheng Hua (2019). “Time series prediction: Predicting stock price”. In: *Predicting stock price* 6, pp. 50–70.

Karlsson, Sofia (2020). *Purchase analysis in the retail industry using Generalized Linear Models*.

Lasek, Agnieszka, Nick Cercone, and Jim Saunders (2019). “Restaurant sales and customer demand forecasting: Literature survey and categorization of methods”. In: *Smart city 360*, pp. 479–491.

Müller, Marlene (2018). “Generalized Linear Models”. In: *parametric and a semiparametric model*, pp. 160–250.

Nelder, John Ashworth and Robert WM Wedderburn (1972). “Generalized linear models”. In: *Journal of the Royal Statistical Society: Series A (General)* 135.3, pp. 370–384.

Sazontyev (2018). “Rossmann store sales quantity prediction”. In: *Sales prediction* 50.1, pp. 137–175.

Yilmaz (2020). “Forecasting house prices in Turkey: GLM, VaR and time series approaches”. In: *Journal of Business Economics and Finance* 9.4, pp. 274–291.

Zelingher, Rotem, David Makowski, and Thierry Brunelle (2020). “Assessing the sensitivity of global maize price to regional productions using statistical and machine learning methods”. In: *Frontiers in Sustainable Food Systems* 5, p. 171.

Hothorn, T., Hornik, K., Zeileis, A. 2006. Unbiased Recursive Partitioning: A Conditional Inference Framework. Journal of Computational and Graphical Statistics, 15(3), 651-674.

P. Jong, G. Z. Heller, “Generalized Linear Models for Insurance Data,” International Series on Actuarial Science, Cambridge University Press, 2008.

Razzaghi, M. 2013. The Probit Link Function in Generalized Linear Models for Data Mining Applications. Journal of Modern Applied Statistical Methods, 12(19), 164-169.

Wamwea, Charity Mkajuma, Benjamin Kyalo Muema, Joseph Kyalo Mung’atu, et al. (2019). “Modelling a pay-as-you-drive insurance pricing structure using a generalized linear model: Case study of a company in Kiambu”. In: American Journal of Theoretical and Applied Statistics 4.6, pp. 527– 533.

Friedman, J., Hastie, T., Tibshirani, R. 2010. Regularization Paths for Generalized Linear Models via Coordinate Descent. Journal of Statistical Software, Articles, 33(1), 1-22.

Appendix

A.1 R Codes

```
library(dplyr)
library(ggplot2)
library(GGally)
library(ROCR)
library(readr)
library(mlbench)
library(gamlss)
library(magrittr)
library(moments)
library(fpp)
supermarket_sales1=read_csv("C:/Users/mkene/Downloads/New folder/
                    supermarket_sales male change.csv")
#CHECKING FOR CONTINUOUS VARIABLES
continuous <-select_if(supermarket_sales1, is.numeric)
summary(continuous)
summary(supermarket_sales1)
table(supermarket_sales$Gender)
n <- nrow(supermarket_sales)  # Number of students
(percent_gender <- table(supermarket_sales$Gender)/n * 100)
barplot(percent_gender,ylim=c(0,70), ylab="percent",main="Barplot of Gender")
skewness(supermarket_sales$Quantity)
##
Xg=supermarket_sales1$Gender
X1=factor(Xg)
X2=supermarket_sales1$`Unit price`
X3=supermarket_sales1$Quantity
X4=supermarket_sales1$`Tax 5%`
X5=supermarket_sales1$Rating
X6=supermarket_sales1$`gross income`
```

```
#Extracting the needed columns
data=data.frame(unitprice=X2,quantity=X3,tax=X4,rating=X5,gender=X1)
##Training/test sets
set.seed(100)
create_train_test <- function(data, size = 0.8, train = TRUE) {
 n_row = nrow(data)
 total_row = size * n_row
 train_sample <- 1: total_row
 if (train == TRUE) {
  return (data[train_sample, ])
 } else {
  return (data[-train_sample, ])
 }
}
data_train <- create_train_test(data, 0.8, train = TRUE)
data_test <- create_train_test(data, 0.8, train = FALSE)
summary(data_train)
#Total sales training and testing
V=supermarket_sales1$Total
Y=data.frame(Sales=V)
sales_create_test <- function(Y, size = 0.8, train = TRUE) {
 n_row = nrow(Y)
 total_row = size * n_row
 train_sample <- 1: total_row
 if (train == TRUE) {
  return (Y[train_sample, ])
 } else {
  return (Y[-train_sample, ])}}
sales_train <- sales_create_test(Y, 0.8, train = TRUE)
sales_test <- sales_create_test(Y, 0.8, train = FALSE)
##Building the model
formula=sales_train~.
```

```
# Gamma model
model_gamma=glm(formula, data=data_train, family=Gamma(link="log"))
summary(model_gamma)
# Inverse Gaussian model
model_inverse.gaussian=glm(formula, data=data_train,
                        family=inverse.gaussian(link = "inverse"))
summary(model_inverse.gaussian)
# Gaussian model
model_gaussian=glm(formula, data=data_train,
                family=gaussian(link = "inverse"))
summary(model_gaussian)
##Chi-square Test
chisq.test(sales_train, correct=FALSE)
pchisq(149019,df=799,lower.tail = FALSE)
#Forecasting the sales
newdata= data.frame(data_test)
predOut=predict(model_gamma, newdata, se.fit = FALSE, scale = NULL, df = Inf,
interval =  "prediction",
level = 0.95, type = "response")
View(predOut)
#MSE of the forecast
actpred=data.frame(pred=predOut, actual=sales_test)
MSE=mean((actpred$actual-actpred$pred)^2)
accuracy(actpred$pred,actpred$actual)
mean(model_gamma$residuals^2)
##Ploting actual vs predicted sales
plot(actpred$actual,type = "l",lty= 1,
          main="ACTUAL VS PREDICTED SALES",col = "red",xlim=c(0,70))
lines(predOut, type = "l", lty=2, col = "blue",xlim=c(0,70), axes=F)
legend("topleft", legend=c("ACTUAL", "PREDICTION"),
col=c("red", "blue"), lty=1:2, cex=0.8)
```