Name: Kennedy Mwangi Reg Number: S030-01-1572/2019
Name: Cynthia Chepkirui Reg Number: S030-01-1609/2019

MODELING SUPERMARKET SALES USING GENERALIZED LINEAR MODEL

Introduction

Tracking sales in a business premise is of great importance.this project therefore specializes in modeling total sales in a supermarket taking the sales to be the dependent variable being affected by other several factors including: unit price, tax, cost of goods sold(cogs), gross margin, gross income.

Problem statement

Many supermarkets today do not have a good forecast of their sales. This is mostly due to lack of skills,resources and knowledge to make sales estimation. At best,most supermarkets chains store use ad hoc tools and predict their upcoming sales . The use of traditional statistical methods to forecast sales has left a lot of challenges unaddressed and mostly result in creation of predictive models that perform poorly.Our aim is to model the supermarket sales and identify the most important variables that will help us in better sales forecast.

Research Objectives

General objectives

The general objective will be to model supermarket sales data using a multiple linear regression model.

Specific objectives

The specific objectives will be:

1. Fit a multiple linear regression model to the data.
2. Check the adequacy of the fitted model.
3. Predict the sales using the fitted model.
4. Check the accuracy of the sales.

Methodology

Fit the Multiple Linear Regression model to use. The multiple linear regression will take the form of:

$$Y_i = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_{p-1} X_{p-1} + e_i \tag{1}$$

- i=0,1,2,...,n
- Y_i= dependent variable, i=0,1,2,...,n

- X_p= explanatory variables, with “p” predictor variables
- β_p=0,1,2,...,p-1, p values
- e_i= residuals (model’s error term), having a normal distribution with mean 0 and constant variance

Assumptions of the model

- The data should be independent and random.
- The response variable Y does not need to be normally distributed but the distribution is from the exponential family.
- The original response variable need not have a linear relationship with the independent variables but the transformed response variable is linearly dependent.
- Homoscedasticity need not be satisfied.
- Errors are independent but need not be normally distributed.

Data

The data consists of supermarket sales. The data has categorical variables and some continuous variables that affect the total sales. The independent continuous variables comprise of: unit price of products, tax, cost of goods sold(cogs), gross margin, gross income and the categorical variables include;Customer type: normal or member, Gender: male or female.

Invoice id: Computer generated sales slip invoice identification number

Branch: Branch of supercenter (3 branches are available identified by A, B and C).

City: Location of supercenters

Customer type: Type of customers, recorded by Members for customers using member card and Normal for without member card.

Gender: Gender type of customer

Product line: General item categorization groups - Electronic accessories, Fashion accessories, Food and beverages, Health and beauty, Home and lifestyle, Sports and travel

Unit price: Price of each product in

Quantity: Number of products purchased by customer

Tax: 5% tax fee for customer buying

Total: Total price including tax

Date: Date of purchase (Record available from January 2019 to March 2019)

Time: Purchase time (10am to 9pm)

Payment: Payment used by customer for purchase (3 methods are available – Cash, Credit card and Ewallet)

COGS: Cost of goods sold

Gross margin percentage: Gross margin percentage

Gross income: Gross income The data consists of 1001 observations. The data was obtained from; (https://www.kaggle.com/aungpyaeap/supermarket-sales)

References

1. Maxwell, O., Mayowa, B. A., Chinedu, I. U., & Peace, A. E. (2018). Modelling count data; a generalized linear model framework. Am J Math Stat, 8(6), 179-183.
2. McCullagh, P., & Nelder, J.A. (1983). Generalized Linear Models (2nd ed.). Routledge. https://doi.org/10.1201/9780203753736