Predicting Market Share of Packaged Food Product

Chad Huntebrinker

2024-12-10

# Title

Predicting Market Share of a Packaged Food Product using assorted variables from 36 consecutive months.

# Abstract

Any profitable company is concerned with market share, especially with market shares of their products. For this report, an examination into the factors influencing the market share of a packaged food product using a multiple linear regression model is done. The data is collected from a national database (Nielsen), spans 36 months, and includes variables such as product price and amount of advertising exposure. The analysis focuses on determining how these factors impact the product’s market share. By exploring the relationships between these variables, the report aims to provide valuable insights for company executives for their marketing strategies and improve market share. The conclusions drawn highlight the most significant influences on market performance, guiding future decision-making processes.

# Introduction

To consider what impacts market shares of a product, wide range of variables need to be considered. First off, market share itself is a measure of how much of a specific industry a product controls and is a percentage. For example, if a product sells 1,000 units of a product out of a total of 10,000 units in the market sector, then the market share for the product is 10%. The next variables that will be used to predict the market share of a product. Most of these variables are self explanatory. These would be the the month and year when this data was taken, the average monthly price (in dollars) of a product, whether there was a discount or some sort of promotion during the period. One variable, however, is the Gross Nielson Rating Points. The “[g]ross rating points (GRPs) quantify impressions as a percentage of a target audience, multiplied by the frequency of the audience seeing the ad. GRPs help marketers to understand the size of an audience that an ad impacts.”[1]. Or in other words, it’s a measurement used for the amount of advertising exposure a product receives. With these variables, this report aims to investigate whether market share of a product can be predicted.

## Primary Analysis Objectives

The Primary Analysis Objective is to investigate whether there is a linear association between market share of a product and variables related to a product and how it is advertised to see if accurate predictions can be made when given a set of values for these variables.

# Methods

## Data Sources

Data was collected from a national database (Nielsen) for 36 consecutive months. The product is from a large packaged foods manufacturer, although it is unknwon what the food is.

## Statistical Analysis

The data is avilable in .xlsx (Excel) format. The data analysis is completed using the software R (version 4.4.1) and this project focuses on multiple linear regression. Two variables were not included when creating the model: id number and month. Id number was not included due to it only being used to keep track of the row’s information. Month was not included as it seemed to have little no impact on the market share. No missing values were found in the dataset. Automatic model selection methods were used to arrive to the final model and the model assumptions are compared and confirmed while reviewing and using the final model.

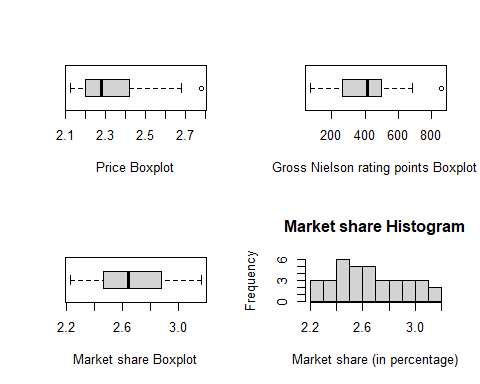
## Model Assumtions

All interfences are done using a significance value of 0.05.

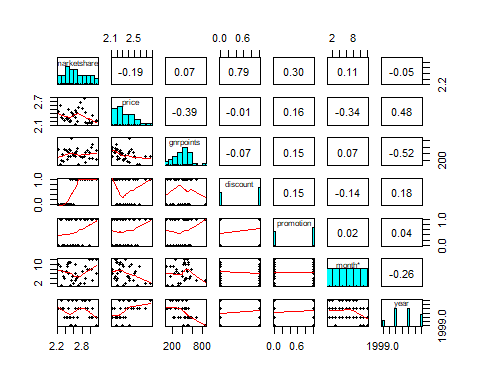
## Primary Objective Analysis

First, this report will investigate the individual predictors and the response variable to verify the skewness and identify any outliers. It will then complete a check for multicollinearity before utilizing automatic model selection to find the most effective predictor variables to include in the model. After that, the report will complete other checks for this final model (like homorskedasticity, normality of the errors, and independence of error terms) before seeing how accurate the model predicts.

### Individual Predictor and Response Variable Investigation

 After reviewing the boxplots for the continuous predictor variables (Price and Gross Nielson), the overall spread is good. They both seem to have an outlier on the far right side of the graph, something to be mindful of and check if any outliers are impacting the model if they are included in the final model. Next, a quick check on the spread of our response variable (market share) has shown no outliers.

### Multicollinearity



## marketshare price gnrpoints discount promotion  
## marketshare 1.00000000 -0.188506951 0.07256395 0.790730217 0.3049951  
## price -0.18850695 1.000000000 -0.38777648 -0.008473014 0.1569611  
## gnrpoints 0.07256395 -0.387776478 1.00000000 -0.072632893 0.1466687  
## discount 0.79073022 -0.008473014 -0.07263289 1.000000000 0.1511858  
## promotion 0.30499506 0.156961061 0.14666874 0.151185789 1.0000000

The next step is to look at both the relationship between market share and all the predictor variables along with the multicolineratiy between the predictor variables. First, there seems to be no or very little multicolineratiy between the predictor variables. When it comes to the relationship between market share and the continuous predictor variables like gnrpoints and price, there doesn’t seem to be much of a linear relationship between them individually. There seems to be some between market share and discount.

### Automatic Model Selection

## [1] 8.051324 5.889624 3.367129 4.130333 6.000000

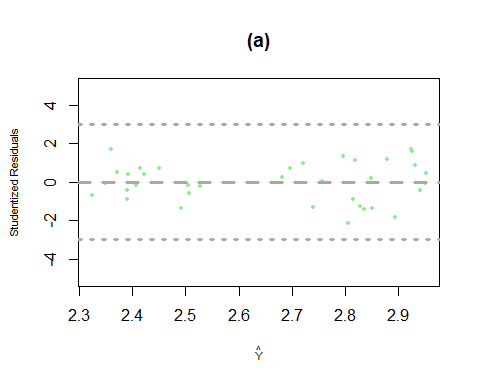
## [1] 0.6252543 0.6641938 0.7065091 0.7180814 0.7193009

## [1] -28.16723 -28.53340 -29.79864 -27.66333 -24.23587

## (Intercept) price gnrpoints discount promotion year   
## TRUE TRUE FALSE TRUE TRUE FALSE

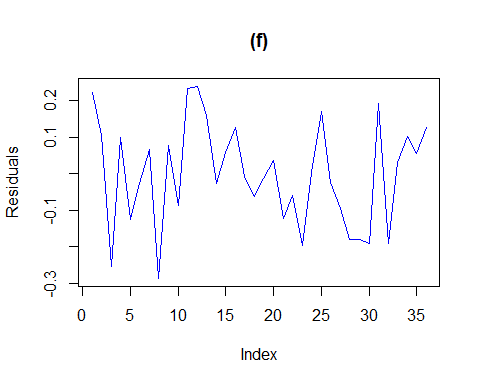
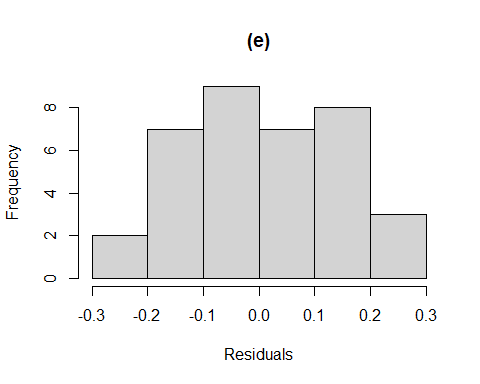
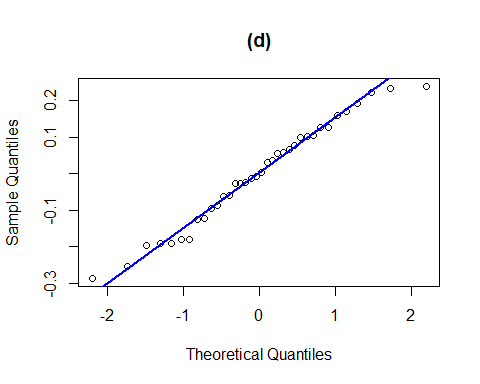
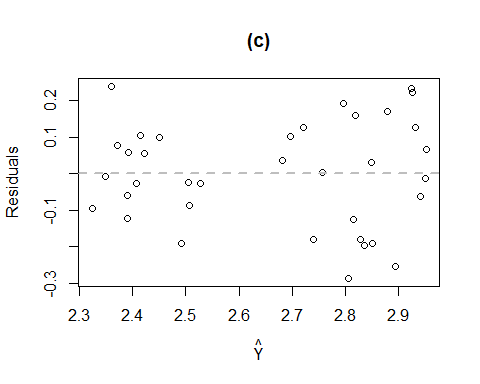
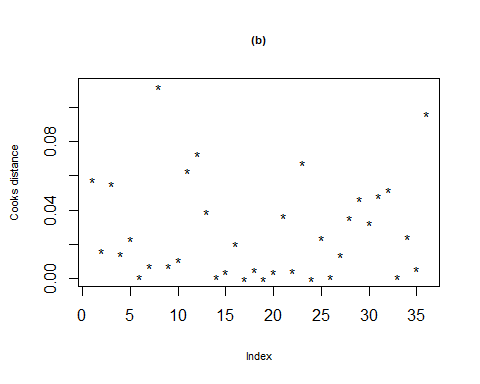
The next step is to find a model via automatic variable selection. This report utilized the “regsubsets” function from the “leap” package in R. This report also utilizes Mallow’s Cp, adjusted R^2, and BIC. Both for Cp and BIC, lower values mean a better model, while a higher adjusted R^2 means a better model. When looking through the scores, this investigation found that the third model had the best results. After looking at what the third model includes, it was found that including price, discount, promotion, and an X-intercept would lead to the best model.

### Final Checks on the Model



## [1] "Value to use for Cook's Distance:"

## [1] 0.5722362



## fit lwr upr  
## 1 2.815898 2.495096 3.136701

## [1] "Actual value: 2.69"

## fit lwr upr  
## 1 2.506228 2.184902 2.827555

## [1] "Actual value: 2.42"

## fit lwr upr  
## 1 2.392427 2.073172 2.711683

## [1] "Actual value: 2.45"

## fit lwr upr  
## 1 2.682583 2.352566 3.012599

## [1] "Actual value: 2.72"

## fit lwr upr  
## 1 2.878559 2.563785 3.193334

## [1] "Actual value: 3.05"

## [1] TRUE

## [1] TRUE

## [1] TRUE

The first thing the report checked with this new model is the residuals. A graph is created for the studentized residuals (graph a) and Cook’s Distance (graph b) to see if any data points have a huge influence on the model. After looking for any points within a range of ±3 for graph a or for any points with a yellow marks indicating they’re outliers, but there are no such points.

The next checks this report looks at is the equal variance (graph c), the normaility of the errors (graph d and e), and the independence of the residuals (graph f). All of these graphs indicate that everything is okay.

The last checks required are the goodness of fit of the model and the accuracy. First, the p-test is done on the three different coefficients. A comparision between their p-value and 0.05 is done and it’s found that each of their values are less than or equal to 0.05. Thus confirming the goodness of fit with the model. The final check this investigation does is with predictions. The model tries to predict what the market share of the product by giving data that’s in the data. What’s found is most of the predictions are close to the actual values (off by 0.1 or less) or within the upper and lower bounds. As a result, it seems to indicate the accuracy of the final model.

# General Discussion and Conclusions

So the final estimated regression function from this investigation would be:

## Ŷ = 3.19 + -0.35 X1 + 0.4 X2 + 0.12 X3

Where Ŷ is the market share, X1 represents the price, X2 indicates whether there was a discount, and X3 indicates whether there was a promotion. The report found that a multiple linear regression model faithfully explains the relationship between the predictor variables and the market share. The MSE (which is found in the Anova table below) is 0.02244, indicating a relatively good fit. The report also looked at different tests and checks like residual plots, Cook’s Distance, and homorskedasticity (to name a few) to verify the model is both accurate and reliable. The model also predicted some of the values in the data and was found to be accurate as well. As a result, this report shows that the price of the product along with offered discounts and promotions impact the market share of a food product. This company could utlize this report as a way to make logical and data-driven decisions that will positively impact the market share of a product.

## Analysis of Variance Table  
##   
## Response: marketshare  
## Df Sum Sq Mean Sq F value Pr(>F)   
## price 1 0.08693 0.08693 3.8745 0.05774 .   
## discount 1 1.52347 1.52347 67.9027 2.057e-09 \*\*\*  
## promotion 1 0.11791 0.11791 5.2552 0.02860 \*   
## Residuals 32 0.71795 0.02244   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

# Appendix: R-code

#Chad Huntebrinker  
#Use signifigance value of 0.05  
  
library(readxl)  
library(leaps)  
library(psych)  
library(lmtest)  
  
excel\_data <- read\_excel("market\_share.xlsx")  
  
  
#Check for outliers for continuous predictor variables (price, and gnrpoints)  
par(mfrow = c(1, 2))  
boxplot(excel\_data$price, xlab = "Price Boxplot", horizontal = TRUE)  
boxplot(excel\_data$gnrpoints, xlab = "Gross Nielson rating points Boxplot", horizontal = TRUE)  
par(mfrow = c(1, 1))  
#Boxplots either show only potentially one outlier.  
  
#Check boxplot for the response variable (marketshare)  
par(mfrow = c(1, 2))  
boxplot(excel\_data$marketshare, xlab = "Market share Boxplot", horizontal = TRUE)  
hist(excel\_data$marketshare, main = "Market share Histogram", xlab = "Market share (in percentage)")  
par(mfrow = c(1, 1))  
#No outliers for the response variable  
  
#Multicollinearity  
pairs.panels(excel\_data[,-1], ellipses = FALSE, density = FALSE)  
temp <- c(1, 7, 8)  
cor(excel\_data[,-temp])  
#there is some concern with multicollinearity with discount and market share  
  
#Check to see what are the best three subsets  
mr <- regsubsets(marketshare~price + gnrpoints + discount + promotion + year,  
 data=excel\_data)  
sum\_of\_mr <- summary(mr)  
  
#Check Cp, R^2, and BIC  
sum\_of\_mr$cp  
sum\_of\_mr$rsq  
sum\_of\_mr$bic  
  
#Model 3 is our best model  
sum\_of\_mr$which[3,]  
#Thus, we will include price, discount and promotion  
  
#Create a model with those three variables in it  
final\_model <- lm(marketshare~price + discount + promotion, data = excel\_data)  
  
#Check for outliers  
#Check studentized residuals against the predicted values in the model  
plot(fitted(final\_model), rstudent(final\_model),  
 col = "lightgreen", pch = 16,  
 xlab = expression(hat(Y)),  
 ylab = "Studentized Residuals",  
 ylim = c(-5, 5), main = "(a)",  
 cex.lab = 0.7, cex = 0.5)  
abline(h = 0, lty = 2, lwd = 3,  
 col = "darkgray")  
abline(h = 3, lty = 3, lwd = 3,  
 col = "darkgray")  
abline(h = -3, lty = 3, lwd = 3,  
 col = "darkgray")  
  
#Cook's Distance  
print("Value to use for Cook's Distance:")  
pf(0.95, df1 = 3, df2 = 33)  
cat("\n")  
cat("\n")  
  
cooksd <- cooks.distance(final\_model)  
plot(cooksd, pch = "\*", cex.main = 0.7, main = "(b)",  
 ylab = "Cooks distance", cex.lab = 0.7)  
text(x = 1:length(cooksd) + 1, y = cooksd, labels = ifelse(cooksd > 0.5722362, names(cooksd), ""),col = "yellow")  
  
  
#Homorskedasticity  
plot(fitted(final\_model), residuals(final\_model), main = "(c)",  
 xlab = expression(hat(Y)), ylab = "Residuals")  
abline(h = 0, col = "gray", lwd = 2,lty = 2)  
  
#Normality  
qqnorm(residuals(final\_model), main = "(d)")  
qqline(residuals(final\_model), col = "blue", lwd = 2)  
hist(residuals(final\_model), main = "(e)", xlab = "Residuals")  
  
#Independence of error terms  
plot(residuals(final\_model), type = "l", col = "blue", main = "(f)", ylab = "Residuals")  
  
#Validate the model  
predict(final\_model, data.frame(price = 2.179, discount = 1, promotion = 0),  
 level = 0.95, interval = "prediction")  
print("Actual value: 2.69")  
cat("\n")  
  
predict(final\_model, data.frame(price = 2.260, discount = 0, promotion = 1),  
 level = 0.95, interval = "prediction")  
print("Actual value: 2.42")  
cat("\n")  
  
predict(final\_model, data.frame(price = 2.248, discount = 0, promotion = 0),  
 level = 0.95, interval = "prediction")  
print("Actual value: 2.45")  
cat("\n")  
  
predict(final\_model, data.frame(price = 2.557, discount = 1, promotion = 0),  
 level = 0.95, interval = "prediction")  
print("Actual value: 2.72")  
cat("\n")  
  
predict(final\_model, data.frame(price = 2.336, discount = 1, promotion = 1),  
 level = 0.95, interval = "prediction")  
print("Actual value: 3.05")  
cat("\n")  
  
#Check p-score  
p\_score <- coeftest(final\_model)  
  
abs(p\_score[2,4]) <= 0.05  
abs(p\_score[3,4]) <= 0.05  
abs(p\_score[4,4]) <= 0.05

# Reference

[1] What is a gross rating point (GRP)? Available online. <https://www.adjust.com/glossary/gross-rating-point-grp/>