Predicting Product Market Share: An Analysis of Price, Advertising, and Promotional Effects

Rafael Pinto

Department of Biostatistics
University of Kansas, US

Table of Contents

List of Tables

List of Figures

Title

Predicting Product Market Share: An Analysis of Price, Advertising, and Promotional Effects

Abstract

This study focuses on the key drivers of a product's market share by analyzing the impact of price, advertising exposure (GNR points), price discounts, and bundle promotions. To this end, we developed a multiple linear regression model using 36 consecutive months of data to quantify the independent effect of each variable on market share. The overall model was statistically significant ($p < 0.001$) and explained approximately 67% of the variance in market share (adjusted $R^2 = 0.6688$), indicating a strong fit. Our analysis revealed a negative effect of price ($p = 0.061$). Furthermore, the presence of a price discount was the most powerful and statistically significant positive predictor ($\beta = 0.40$, $p < 0.001$), followed by promotions ($\beta = 0.12$, $p = 0.039$). The impact of overall advertising was not significant in this model after controlling for the other variables.

Introduction

A product's market share management through strategic planning remains essential for business success in the competitive consumer goods market. Different marketing and sales tools including pricing strategies and advertising campaigns and promotional activities help companies establish and defend their market position. Historical data analysis serves as an essential tool for making data-driven decisions and optimizing resource distribution.

The main goal of this research involves statistical analysis of the "Market Share Dataset" to identify the primary elements that affect market share performance of a particular product across 36 months. This report investigates market share factors through business variable analysis to develop a predictive model which identifies and measures each factor's effect.

Our analysis will address the following key research questions:

1. Considering factors such as price, advertising exposure (GNR points), discounts, and promotions, which are statistically significant predictors of a product's market share?
2. What is the direction (positive or negative) and magnitude of the effect of each significant predictor?
3. To what extent is the overall change in the product's market share over time explained by the combination of these factors?

Data Description

We used the "Market Share Dataset" from a STAT 823 class for this project. The data originally came from a national Nielsen database and tracks 36 months of activity for a single packaged food product, from September 1999 to August 2002.

Our goal was to understand what drives the product's market share. To do this, we focused on a few key predictor variables:

- Response Variable:
  - marketshare (numeric): the average monthly market share for the product.
- Predictor Variables:
  - price (numeric): the average monthly price of the product in dollars.
  - gnrpoints (numeric): a Gross Nielsen index that measures how much advertising the product got.
  - discount (binary, 0/1): a variable showing if a price discount was offered (this happened in 21 of the 36 months).
  - promotion (binary, 0/1): a variable showing whether there was a special package promotion (this was active in 20 of the 36 months).

The raw data also included other variables like month and year, which were helpful for context, but we didn't use them as predictors in our final regression model. Table 1 gives a quick summary of the key numbers we analyzed.

Table 1: Descriptive Statistics

Variable	**Mean**	**Std. Dev.**	**Median**	**Min**	**Max**
marketshare	2.664	0.264	2.640	2.230	3.160
price	2.324	0.163	2.280	2.124	2.781
gnrpoints	388.1	168.493	412.0	72.0	858.0

Exploratory Data Analysis

Before diving into any complex modeling, we wanted to poke around the data to see what initial patterns we could find. We started by getting a feel for the main variable, **marketshare**. Then we looked at how it connected with each of the predictors.

First, we checked the distribution of the marketshare itself. The boxplot in Figure 1 shows it's balanced, with a median of 2.64. The middle half of the data points fall between 2.47 and 2.88, and importantly, there were no outliers to worry about.

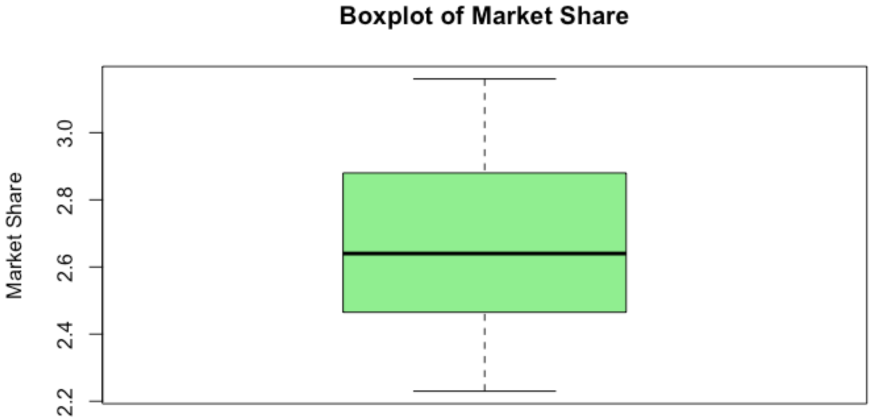


Figure 1: Boxplot of Monthly Market Share

Next, we started looking for relationships. The scatter plot in Figure 2, which compares marketshare to price, shows a clear downward trend. As you might expect, when the price went up, the market share tended to go down. That said, the data points are fairly scattered, which suggests the connection is there but maybe not super strong by itself.

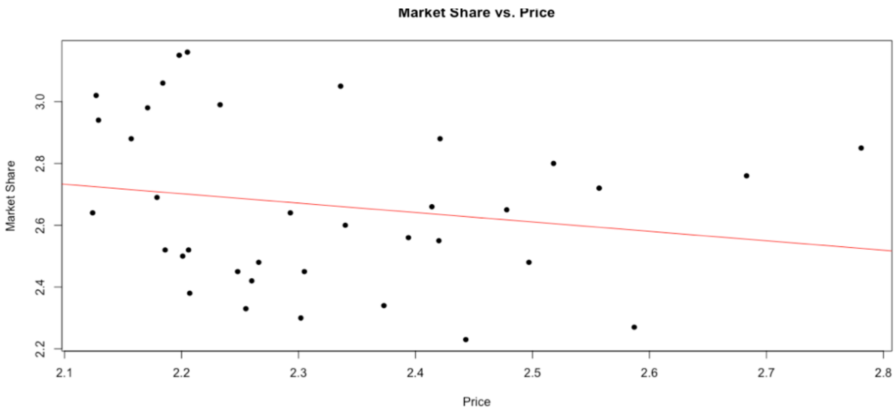


Figure 2: Scatter Plot of Market Share vs. Price

Finally, we checked out the impact of the binary (yes/no) variables. Offering a price discount, as you can see in Figure 3, seems to have a huge impact. The median market share was way higher in months that had a discount. This visual evidence is compelling. Package promotions also seemed to boost market share, but the effect wasn't nearly as dramatic as it was for the discounts.

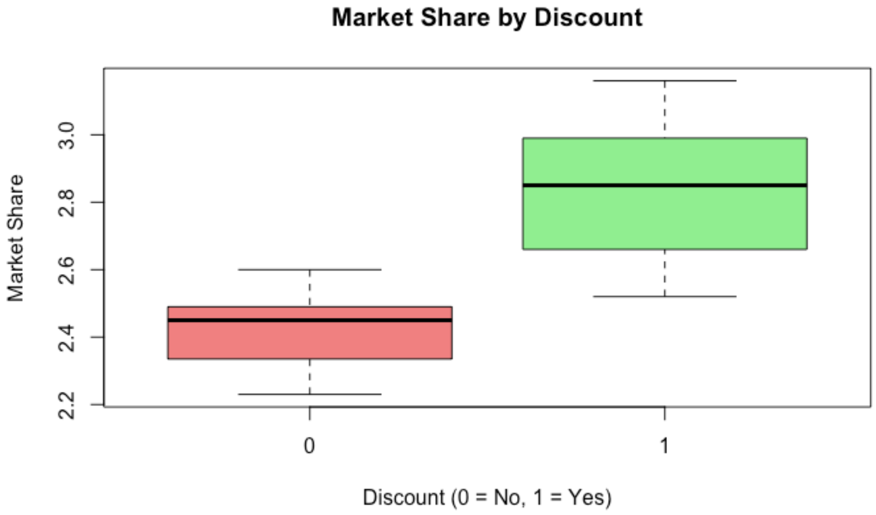


Figure 3: Market Share by Discount Status

Methodology

While a simple one-on-one comparison can give us a first glimpse of relationships and meanings in our data, this wouldn't be enough to confidently understand in-depth what drives market share. It could happen, for example, that a price drop occurred during a promotion. Therefore, we needed a tool that could unravel these effects and simultaneously analyze all potential factors (price, advertising, discounts, and promotions). That's why we chose multiple linear regression as the central axis of our analysis. The specific model we built sought to predict market share based on this equation:

$$\textbf{marketshare} = \beta_0 + \beta_1 * (\textbf{price}) + \beta_2 * (\textbf{gnrpoints}) + \beta_3 * (\textbf{discount}) + \beta_4 * (\textbf{promotion}) + \varepsilon$$

In this setup, each coefficient (the "beta" values) shows the unique impact of a single factor, holding the others constant. This allowed us to see, for example, the unique effect of a discount, independent of any advertising that might be occurring simultaneously.

We used the R programming language for all calculations. Our determination of whether a factor's effect was statistically significant or simply random was based on a standard significance threshold of 0.05. Understanding that the quality of a model's results depends largely on its basis, we performed diagnostic checks to ensure that the basic assumptions of linear regression were met, thus ensuring the reliability of the conclusions obtained.

Results

The first step after building the multiple linear regression model involved checking its predictive capability. The model demonstrated strong predictive power since the $F(4, 31) = 18.67$ value was significant at $p < 0.001$. The selected factors together demonstrated superior predictive power for market share than random chance because the model results were highly significant. The four factors together explained about 67% of market share variations based on the adjusted R^2 value of 0.669. The individual effects of each factor became apparent through the results. Table 2 shows these specific effects.

Table 2: Multiple Regression Model Coefficients for Predicting Market Share

Predictor	**Estimate (β)**	**Std. Error**	**t value**	**p-value**
(Intercept)	3.158	0.441	7.168	< .001 ***
price	-0.344	0.177	-1.946	0.061 .
gnrpoints	0.00002	0.00017	0.116	0.908
discount	0.400	0.052	7.623	< .001 ***
promotion	0.117	0.054	2.160	0.039 *

*Note: Signif. codes: 0.001*** 0.01*** 0.05***

The data presented in the results section provides a clear story. The model demonstrated a straightforward order of influence which showed that offering discounts had the greatest impact on market share. The positive effect of this factor was both significant and substantial ($\beta = 0.40$, $p < 0.001$). The market share increases by about 0.40 points through discount strategies after controlling for all other variables. The positive impact of bundle promotions on market share was statistically significant (β = 0.117, p = 0.039) but smaller than the effect of direct price discounts.

The relationship between market share and price demonstrated a negative pattern. Although the p-value of 0.061 failed to meet the conventional 0.05 threshold the data showed a clear pattern that higher prices resulted in decreased market share.

The most unexpected discovery emerged from the advertising data because advertising (GNR points) failed to show a statistically significant relationship with market share (p = 0.908).

Discussions

The results from multiple linear regression show that the product market share depends mainly on price discounts and the presence of package promotions. The selected marketing and sales variables are highly relevant since the overall model explains 67% of the monthly variation in market share.

The strongest variable in the model is price discount since its effect is highly significant and positive. The model shows that the presence of discounts leads to a market share increase of 0.40 points while all other factors are constant. The findings from the exploratory analysis are consistent with this result and confirm that consumers of this product type are highly responsive to price discounts. Package promotions have a positive impact on market share (0.12 points) yet it is less than one-third of the impact of price discounts. The data shows that discount promotions are more effective at increasing market share than other promotional methods.

Price has a marginally significant negative relationship with market share according to the model. The result is not statistically significant at the 0.05 level (p = 0.061) although it follows economic theory that higher prices lead to lower demand and subsequently lower market share. The marginal significance could be due to the small sample size (n=36) and more data could reveal a stronger statistical relationship.

The final model did not show any significant relationship between advertising (GNR points) and market share. The initial exploratory plot showed a weak positive relationship but the multiple regression analysis which controls for other variables found the effect to be statistically insignificant. It is likely that advertising is less influential than discounts and promotions because they have stronger and faster effects on the market.

The purchasing behavior of this product's customer base is more likely to be determined by point-of-sale incentives than by general advertising exposure. The GNR points metric might not effectively measure the creative quality or effectiveness of advertising methods.

The analysis shows that the product's market share can be best influenced by price-based sales tactics.

Conclusions and Limitations

Conclusion

The statistical analysis used in this study proved successful in determining the main drivers of market share for the studied product. The study results show that direct sales methods are effective. Price discounts prove to be the most powerful positive factor for market share growth and promotional packaging comes in second. Market share decreases when prices rise. The model results show that general advertising exposure through GNR points does not significantly affect market share after price and promotional effects are considered.

Limitations

Several study limitations exist in the model results which need consideration for proper interpretation of the findings:

1. **Limited Sample Size:** The analysis of 36 months of data restricts the model from detecting small price changes because of its limited sample size. A study that uses data from multiple years would produce more reliable and robust findings.
2. **Potential for Omitted Variable Bias:** The model's limited variable set might result in omitted variable bias because it does not account for competitor pricing actions and economic changes and seasonal consumer patterns. The model estimates might contain effects from unobserved variables because of this limitation.
3. **Mild Heteroscedasticity:** The model produced heteroscedasticity during diagnostic testing because it generated larger prediction errors when market share

reached high levels. The results remain valid although some coefficient estimates experience a slight decrease in precision.

4. **Generalizability:** The findings are limited to a single product throughout the period from 1999 to 2002. The effectiveness of marketing levers shows different behavior across different products and industries under various market conditions. The results need to be used with caution because they remain specific to their original context.

Appendix: R-code

Setup

```
echo=TRUE
library(readxl)
library(ggplot2)
library(plotly)

market_share <- readxl::read_excel("market_share.xlsx")
```

Initial Look

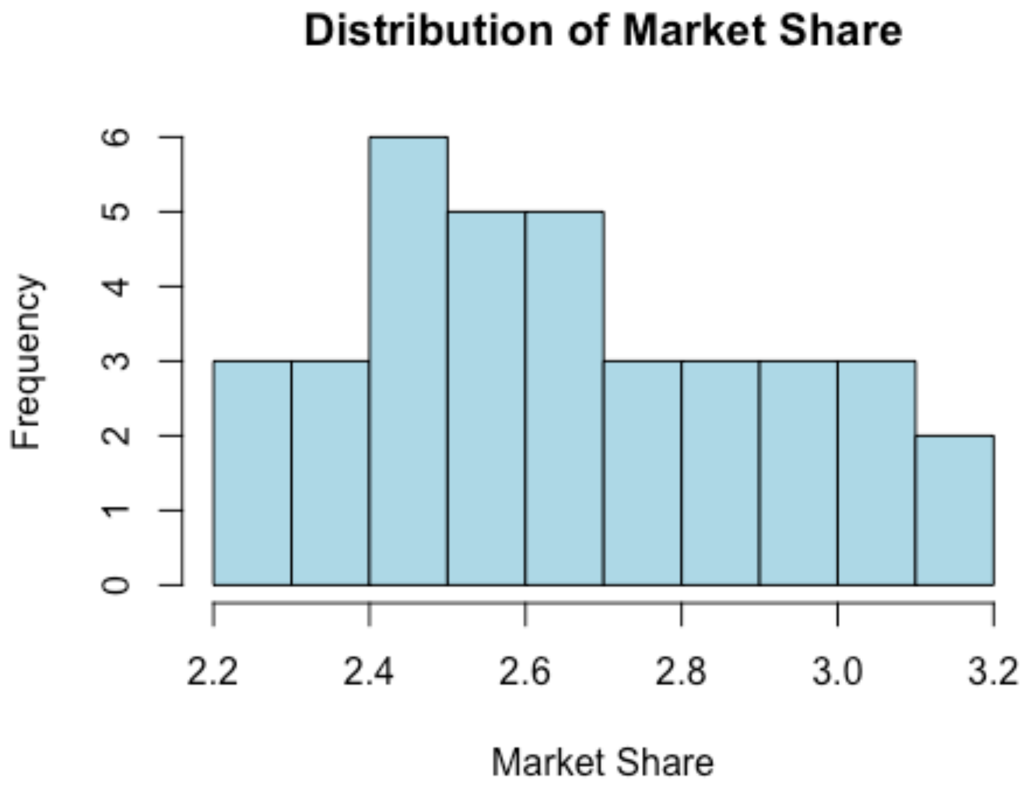
```
echo=TRUE
head(market_share)

## # A tibble: 6 × 8
##   idnum marketshare price gnrpoints discount promotion month  year
##   <dbl>       <dbl> <dbl>     <dbl>    <dbl>     <dbl> <chr> <dbl>
## 1     1        3.15  2.20       498        1         1 Sep    1999
## 2     2        2.52  2.19       510        0         0 Oct    1999
## 3     3        2.64  2.29       422        1         1 Nov    1999
## 4     4        2.55  2.42       858        0         1 Dec    1999
## 5     5        2.69  2.18       566        1         0 Jan    2000
## 6     6        2.38  2.21       536        0         0 Feb    2000

str(market_share)

## tibble [36 × 8] (S3: tbl_df/tbl/data.frame)
##  $ idnum      : num [1:36] 1 2 3 4 5 6 7 8 9 10 ...
##  $ marketshare: num [1:36] 3.15 2.52 2.64 2.55 2.69 2.38 3.02 2.52 2.45 2.
42 ...
##  $ price      : num [1:36] 2.2 2.19 2.29 2.42 2.18 ...
##  $ gnrpoints  : num [1:36] 498 510 422 858 566 536 585 310 211 504 ...
##  $ discount   : num [1:36] 1 0 1 0 1 0 1 1 0 0 ...
##  $ promotion  : num [1:36] 1 0 1 1 0 0 1 0 0 1 ...
##  $ month      : chr [1:36] "Sep" "Oct" "Nov" "Dec" ...
##  $ year       : num [1:36] 1999 1999 1999 1999 2000 ...

summary(market_share)

##      idnum        marketshare        price          gnrpoints
##  Min.   : 1.00   Min.   :2.230   Min.   :2.124   Min.   : 72.0
##  1st Qu.: 9.75   1st Qu.:2.473   1st Qu.:2.200   1st Qu.:268.0
##  Median :18.50   Median :2.640   Median :2.280   Median :412.0
##  Mean   :18.50   Mean   :2.664   Mean   :2.324   Mean   :388.1
##  3rd Qu.:27.25   3rd Qu.:2.880   3rd Qu.:2.420   3rd Qu.:499.5
##  Max.   :36.00   Max.   :3.160   Max.   :2.781   Max.   :858.0
##     discount        promotion         month                year
##  Min.   :0.0000   Min.   :0.0000   Length:36          Min.   :1999
##  1st Qu.:0.0000   1st Qu.:0.0000   Class :character   1st Qu.:2000
##  Median :1.0000   Median :1.0000   Mode  :character   Median :2001
##  Mean   :0.5833   Mean   :0.5556                      Mean   :2001  
##  3rd Qu.:1.0000   3rd Qu.:1.0000                      3rd Qu.:2001  
##  Max.   :1.0000   Max.   :1.0000                      Max.   :2002

names(market_share)

## [1] "idnum"       "marketshare" "price"       "gnrpoints"   "discount"   
## [6] "promotion"   "month"       "year"

dim(market_share)

## [1] 36  8
```

- General Notes
  - Response variable: marketshare
  - Potential predictors: price, gnrpoints, discount, promotion.
    - `discount`: Mean 0.5833 (a discount was active in approximately 58.33% of the months)
    - `promotion`: Mean 0.5556. (promotion was active in approximately 55.56% of the months)

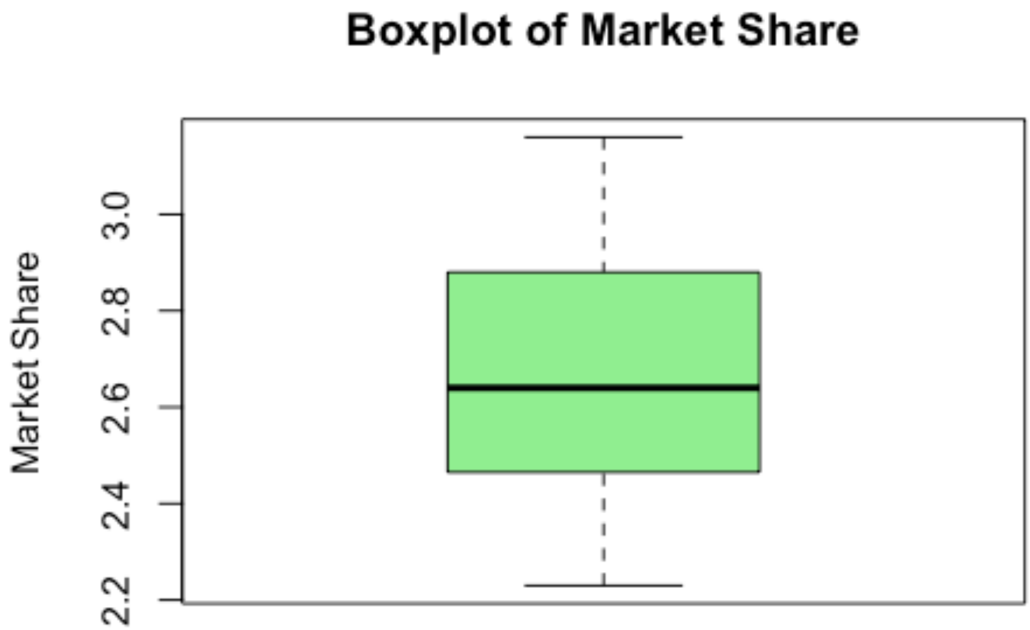
Visualizing the Data

Distribution of `marketshare`

```
echo=TRUE
hist(market_share$marketshare,
     main = "Distribution of Market Share",
     xlab = "Market Share",
     col = "lightblue")
```



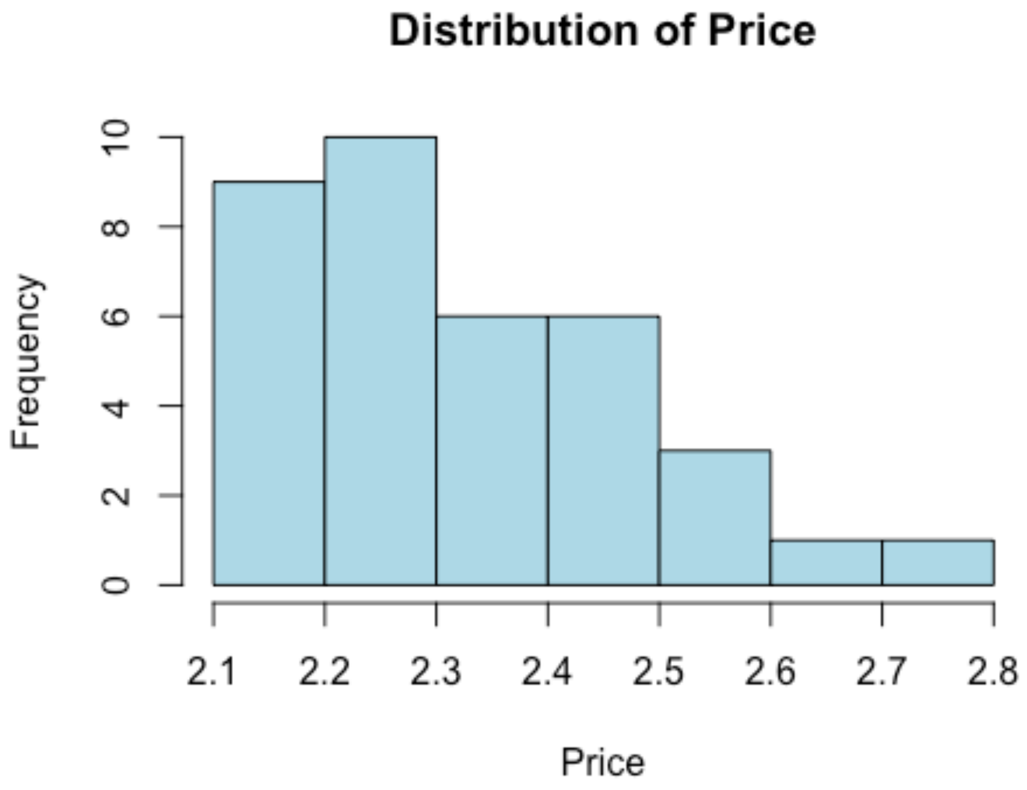
```
boxplot(market_share$marketshare,
        main = "Boxplot of Market Share",
        ylab = "Market Share",
        col = "lightgreen")
```



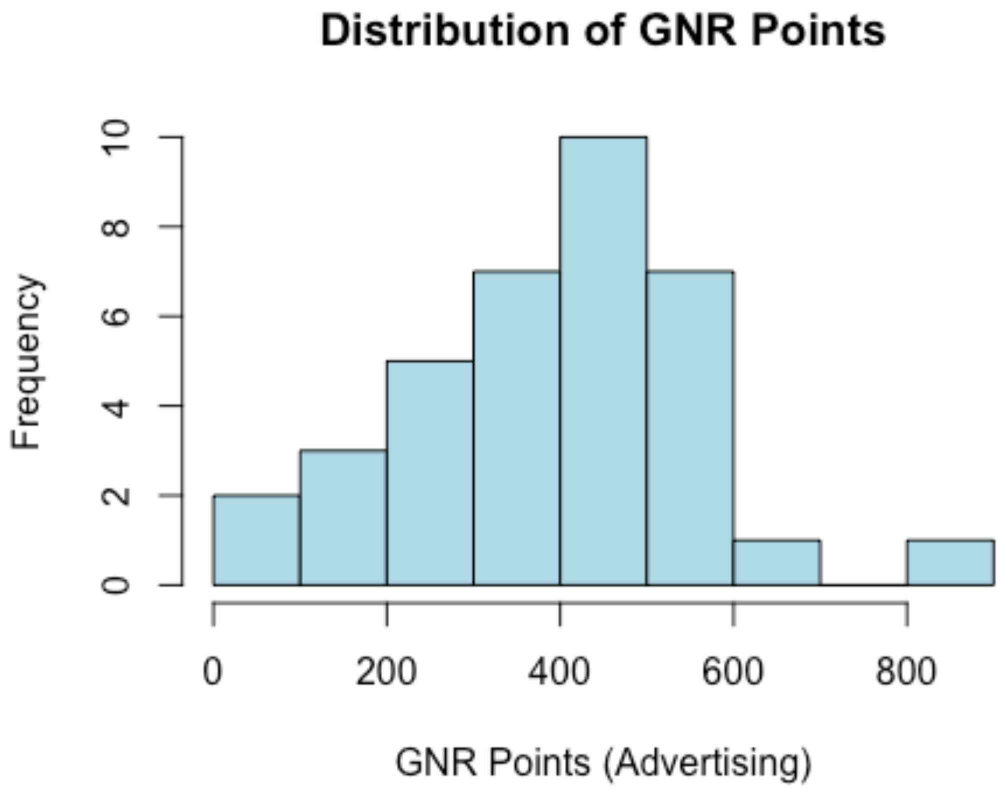
- General Notes:
  - Histogram:
    - the histogram is right-skewed
      - I our summary we could confirm this as the mean (2.664) is slightly greater than the median (2.640)
    - the majority of the values fall between 2.4 and 2,8
    - the most frequent market share values are in the 2.4-2.5 range
  - Boxplot
    - there are no obvious outliers

Distribution of `price` and `gnrpoints` (continuous predictors)

```
echo=TRUE
hist(market_share$price,
     main = "Distribution of Price",
     xlab = "Price",
     col = "lightblue")
```



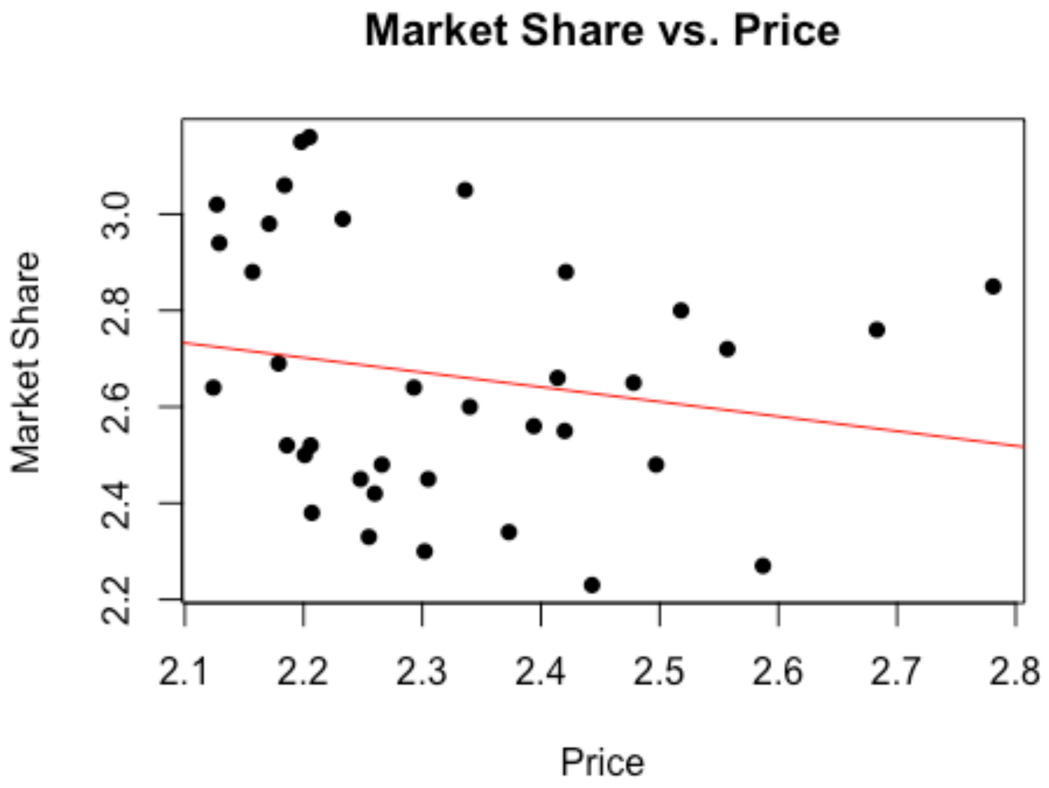
```r
hist(market_share$gnrpoints,
     main = "Distribution of GNR Points",
     xlab = "GNR Points (Advertising)",
     col = "lightblue")
```



- General Notes:
  - Histogram of price
    - the most common price range is between 2.2 and 2.3
    - right skewed (mean: 2.324; median: 2.280)
      - it show a decreasing frequency towards higher prices
  - Histogram of gnrpoints (advertising)
    - it looks like hish levels of advertising (above 600) are less common
    - there is a clear tendency for advertising from 300 to to 600 point range with a peak between 400-500.

How marketshare might relate to price

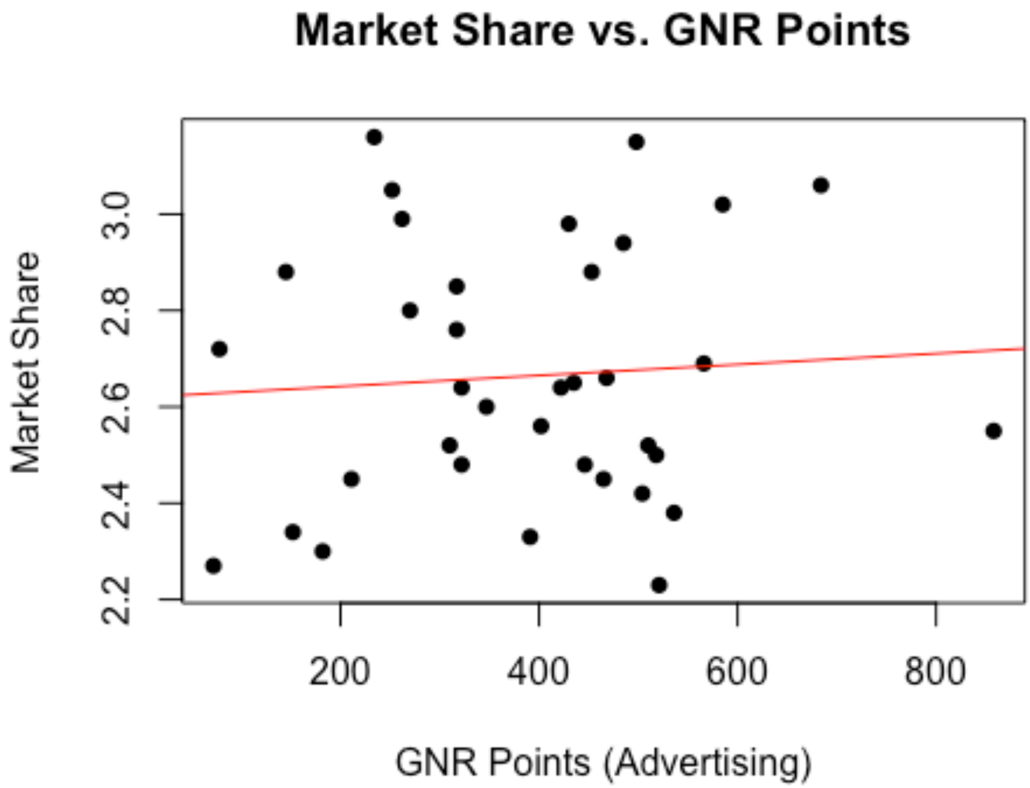
```
echo=TRUE
plot(market_share$price, market_share$marketshare,
     main = "Market Share vs. Price",
     xlab = "Price",
     ylab = "Market Share",
     pch = 16) # pch = 16 gives solid circles

# Adding a simple linear regression line to the scatter plot
abline(lm(marketshare ~ price, data = market_share), col = "red")
```



- General Notes:
    - there is a negative relationships
        - it looks like as prices increase, there are fewer points in the plot and the line goes low.
        - ==so, if you raise the price of a product, you might sell less==
    - black dots are not tightly culstered==
    - ==becuase of the dispertion of black points around the line, it looks like price along cannot predict marketshare in a perfect way==

How marketshare might relate to gnrpoints

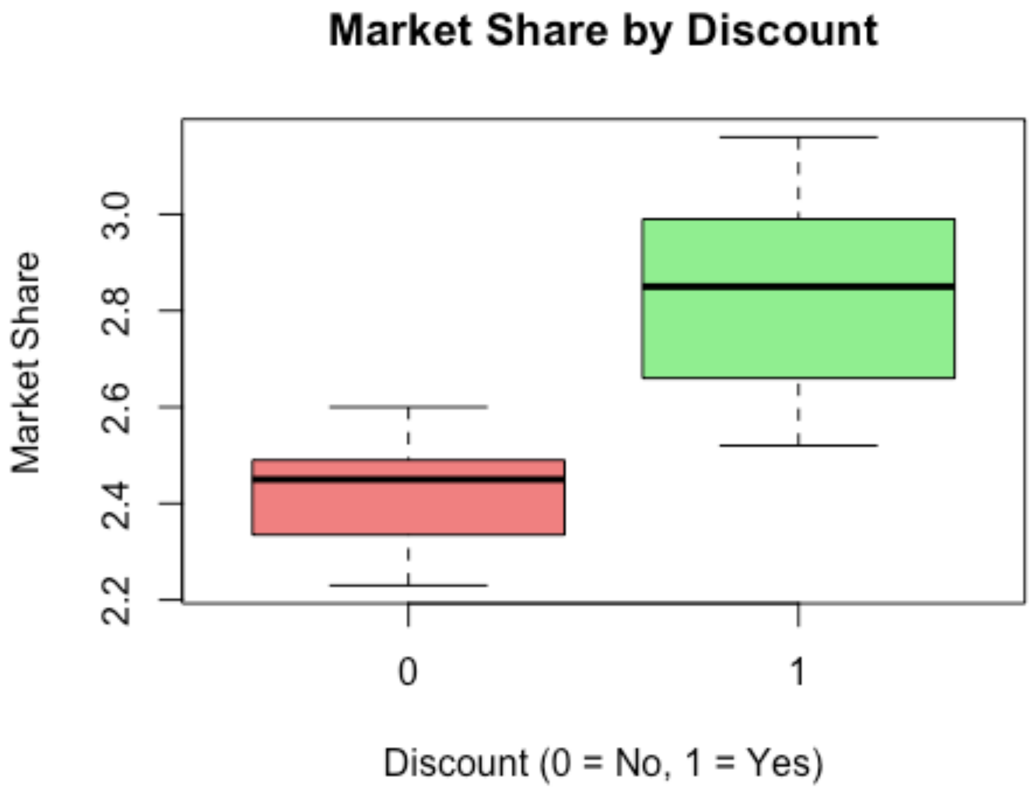
```
echo=TRUE
plot(market_share$gnrpoints, market_share$marketshare,
     main = "Market Share vs. GNR Points",
     xlab = "GNR Points (Advertising)",
     ylab = "Market Share",
     pch = 16)

# Adding a simple linear regression line
abline(lm(marketshare ~ gnrpoints, data = market_share), col = "red")
```



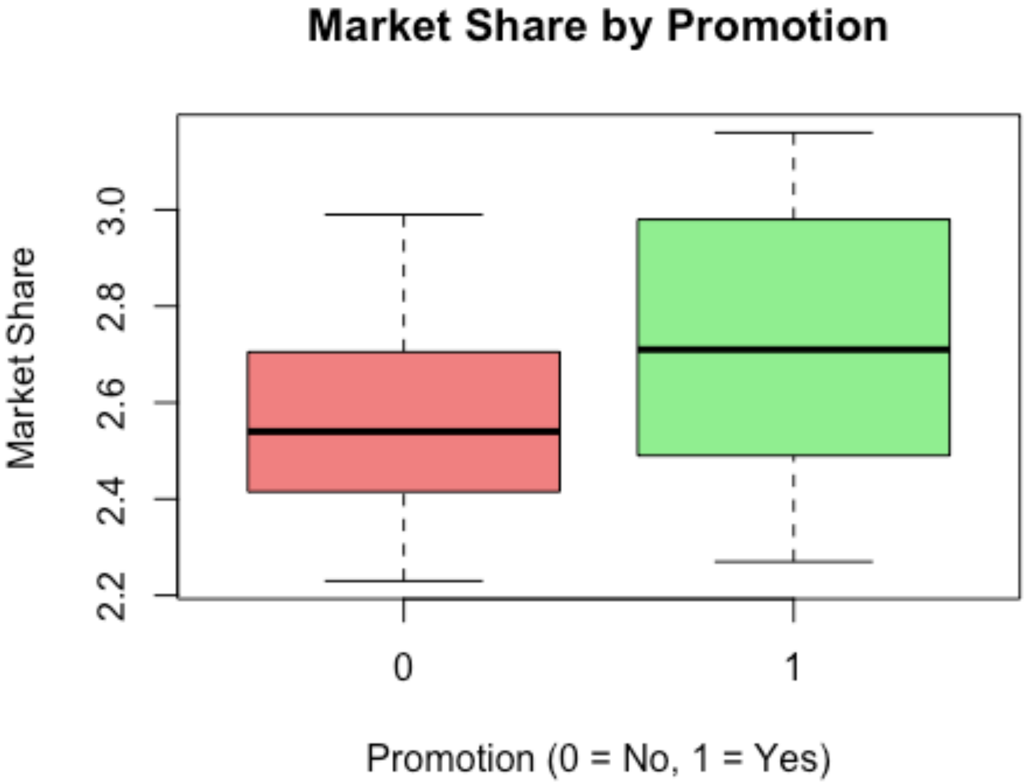
- General Notes:
  - positive relationships or correlation
    - as gnrpoints increases, ‘marketshare’ tends to increase
    - it looks like this relationship is very weak (the data points are widely spread out)
      - ==advertising alone doesn;t seem to be a strong driver of market share==

How discount and promotion affect marketshare

```
echo=TRUE
boxplot(marketshare ~ discount, data = market_share,
        main = "Market Share by Discount",
        xlab = "Discount (0 = No, 1 = Yes)",
        ylab = "Market Share",
        col = c("lightcoral", "lightgreen"))
```



```r
boxplot(marketshare ~ promotion, data = market_share,
        main = "Market Share by Promotion",
        xlab = "Promotion (0 = No, 1 = Yes)",
        ylab = "Market Share",
        col = c("lightcoral", "lightgreen"))
```



- General Notes:
  - Boxplot marketshare ~ discount
    - the median market share seems higher when there’s a discount
      - ==it suggests that in months when a discount was offered, the market share was higher==
      - as the two medians loojs large, it may suggest this could be an important factor
      - ==as the discount box (green) is taller than the no discount box (red), this indicate more variability in market share during months with a discount.==
  - Boxplot market ~ promotion
    - the median market share is higher when there’s a promotion
      - ==this may suggest that months with a package promotion tend to have a higher market share==
    - the promotion box (green) is taller than the no promotion box (red). this suggests that there is more variability in market share when a promotion is active

Multiple Linear Regression Model

- RECAP:
    - `price`: it looks like it has a negative effect on market share (higher price -> lower share).
    - `gnrpoints`: it looks like it have a weak positive effect (more advertising -> slightly higher share).
    - `discount`: it looks like it has a strong positive effect (discount -> higher share).
    - `promotion`: it looks like it has a positive effect, but it looks weaker than the effect of a discount.
- RESEARCH QUESTION:
    - What are the effect of `price`, `gnrpoints`, `discount`, and `promotion` on the market share? Which of these factors are significant predictors of market share?

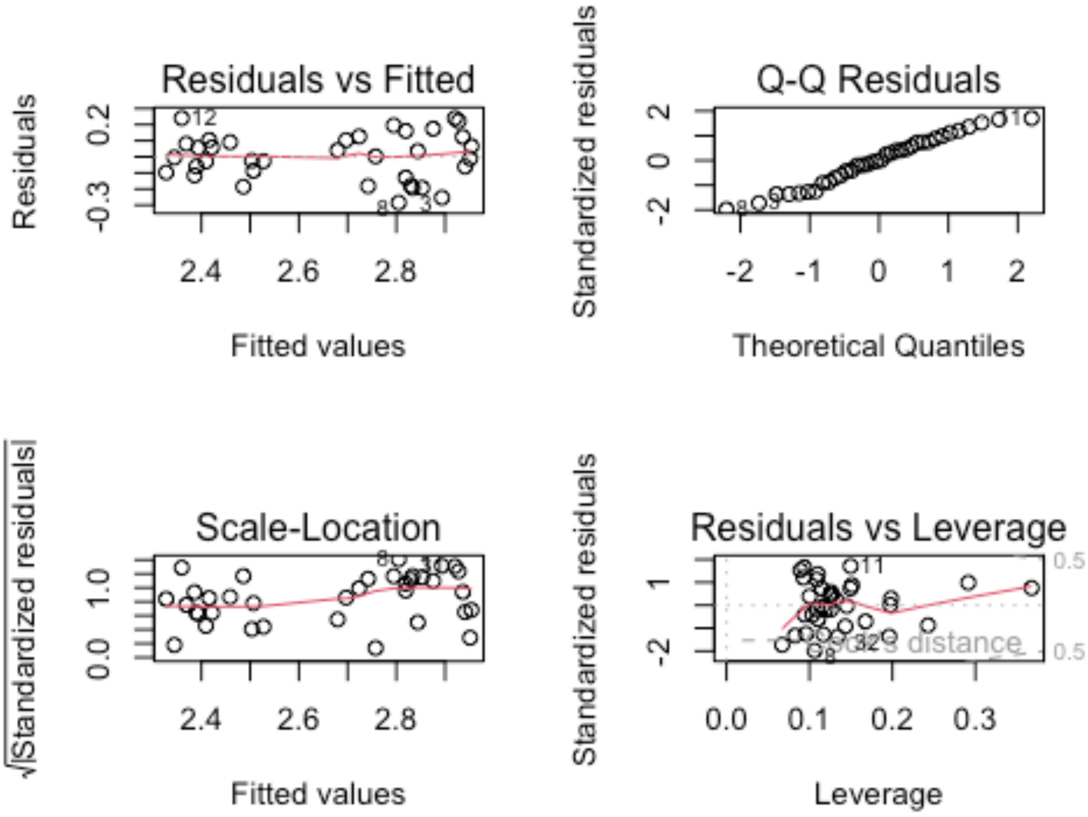
```
echo=TRUE
fit_model <- lm(marketshare ~ price + gnrpoints + discount + promotion, data
= market_share)
summary(fit_model)

##
## Call:
## lm(formula = marketshare ~ price + gnrpoints + discount + promotion,
##     data = market_share)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.284946 -0.102265 -0.001004  0.103386  0.240284
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.158e+00  4.405e-01   7.168 4.67e-08 ***
## price       -3.439e-01  1.767e-01  -1.946   0.0607 .
## gnrpoints    1.993e-05  1.714e-04   0.116   0.9081
## discount     3.999e-01  5.246e-02   7.623 1.35e-08 ***
## promotion    1.165e-01  5.394e-02   2.160   0.0386 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1522 on 31 degrees of freedom
## Multiple R-squared:  0.7066, Adjusted R-squared:  0.6688
## F-statistic: 18.67 on 4 and 31 DF,  p-value: 6.641e-08
```

- General Notes:
    - F-statistic and p-value: The overall regression model was statistically significant, F(4, 31) = 18.67, p < .001

- Multiple R-squared & Adjusted R-squared: my model explains about 67% of the variability in market share.
- Predictors: Estimate and Pr(>|t|)
    - `price`:
        - Estimate and Pr(>|t|): `price` has a marginally significant negative effect on market share (β = -0.34, p = 0.061).
    - `gnrpoints`:
        - Estimate and Pr(>|t|): `gnrpoints` is not a statistically significant predictor (p= 0.908).
        - not a statistically significant predictor (p= 0.908).
    - `discount`:
        - Estimate and Pr(>|t|): `discount` is a highly significant positive predictor (β = 0.40, p < .001).
        - Offering a discount is associated with a 0.40-point increase in market share
    - `promotion`:
        - Estimate (1.165e-01) and Pr(>|t|) (0.0386): `promotion` is a significant positive predictor (β = 0.12, p = 0.039)
        - it associates with a 0.12-point increase in market share
- Notably, when visualizing the data, almost all of these outputs came across as well: the strong positive effect of discount, the weaker positive effect of promotion, the negative effect of price, and the very weak effect of gnrpoints.

Checking Model Assumptions

```
echo=TRUE
par(mfrow = c(2, 2))
plot(fit_model)
```



```
par(mfrow = c(1, 1))
```

- General notes:
    - Residuals vs. Fitted:
        - Does the relationship between the predictors and the response variable is linear (linearity)?
            - as the points look randomly scattered around the zero line, I understand that the linearity assumption is met.
        - Do the residuals have a constant variance across all levels of the predicted values (homoscedasticity)?
            - as the vertical spread of the points seems consistent, I understand that the equal variance assumption is also met.
    - Normal Q-Q
        - Does the residuals follow a normal distribution?
            - there are some points that deviate (like 8 and 3), but the vast majority are falling on the dotted line. Thus, the normality of residuals assumption is met.
    - Scale-Location

  - Do the residuals have a constant variance across all levels of the predicted values (homoscedasticity)?
    - There is an upward trend in the red line, which might suggest heteroscedasticity.
    - While the ‘Residuals vs. Fitted’ plot showed no major signs of non-linearity or heteroscedasticity, the ‘Scale-Location’ plot did suggest a slight tendency for the error variance to increase with larger fitted values.
      - Noting that the sample size is small, this might not be a violation but a limitation that could be further explores with more data.
- Residuals vs Leverage
  - Are there influential outliers?
    - there are no highly influential outliers
      - point 8 is a significant outlier, with standardized residual below -2. Also, there are points that show a high leverage, even one point exceeding 0.3. But the fact that none of these outliers or high-leverage points are influential strengthens is good for the confidence on the model.

References

STAT 823 Course Materials. (2025). "Datasets and Instructions." University of Kansas, Department of Biostatistics.

R Core Team (2023). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.