Predicting Market Share for a Large Packaged Food Manufacturer

STAT 823: Summer Class Project, 2020

Brenton Summers

Department of Biostatistics and Data Science
University of Kansas, USA
March 13, 2021

Contents

List of Tables

List of Figures

Abstract

Understanding and predicting a product's market share allows a manufacturer to strategically plan for a better future. This knowledge can lead to increased efforts in areas of need, increased revenue and improved budgeting. Finding the right factors allows a manufacturer to understand the market and optimize their revenue.

The objective of this study is to find accurate and valuable predictor variables which influence market share of a specified product. The dataset used for this study is from a large packaged food manufacturer for the three year period of 1999 to 2002.

The data was generated from the National Nielson dataset. The variables collected include the average monthly price of a specified product, whether the product was discounted, whether the product had an active promotion, month, year, and the GNR Points which is an index of the amount of advertising exposure that the product received.

Introduction

Market share can be defined as the percent of total sales in an industry generated by a particualr company or product. The percentage is calculated by taking the product sales over the defined period and dividing it by the total sales across the industry over the same period. This is a metric that shows the size and success of a company in comparison to its competitors.

As the demand for a product increases, those companies who maintain their market share ensure an increase in revenue at the same rate as the market. A company or product who grows its market share will grow its revenues faster than its competitors. This leads to larger profits and more success.

Companies can attempt to increase their market share by lowering prices, using advertising, including promotions. and/or appeal to other audiences.Hence for this analysis we are focused on the impact price, discounts, promotions, advertising and date have on a product's market share. Understanding the cause and effect relationship between the market share and the predictor variables can yield to an improved business plan and increased revenue. *{Hayes}*

Primary Analysis Objectives

To investigate the linear association between the market share percentage for packaged food products, date, price, and advertising amount index as well as to predict the market share using the same variables.

Materials and Methods

Data Sources

The dataset was obtained online via the Nielsen national database. Here is Nielsen's homepage.

Variables in the data set include **price**, **discount**, **promotion**, **month**, **year**, and **gnrpoints**. The price variable is the monthly average price for the specified product. The discount and promotion variables are binary and state whether there is a discount/promotion applied to the price of the product. The month and year help allow a seasonal variable to be factored in. The gnrpoints variable is an index of the advertising amount for the product.

This dataset was chosen based off interest in marketing and a curiousity in the relationship between market share and price as the value fluctuates.

Statistical Analysis

The data is available in .xlsx (excel) format. The data analysis is done using the statistical software R version 3.6.1 (2019-07-10) and the project focuses mainly on multiple linear

regression. Each of the predictor variable is explored individually and the illustrations used are conducted on the entire dataset for the preliminary investigation. No missing values were found in the dataset. The large sample size and absence of missing value is assumed to ensure better predictibility and less sampling variability. Automatic model selection method has been used to arrive at the final model. The model assumptions are assessed and a final model is decided upon.

Model Assumptions

All inferences are conducted using alpha = 0.05 unless stated otherwise. No adjustments for multiplicity are made as this is an exploratory analysis. Discrete variables are summarized with proportions and frequencies. Continuous variables are summarized using the basic statistics.

Primary Objective Analysis

Exploring individual predictors and the response variable is very important before starting data analysis. It helps in detecting skewness, presence of outliers or can also suggest if transformations are necessary to fit a better model. Once the dataset variables are explored, the next step is performed to check the linearity between the predictors and the outcome individually. This helps us fit a better model which can do a better job at explaining the variation of the response.

### Analysis of Market Share Percentage

Preliminary data analysis on the market share percentage shows that the response variable is slightly skewed to the right as there are more data points with lower market shares. Presence of outlying values are not observed. To combat the skewness, the market share response variable was transformed using the logarithmic function. While this transformation does not make a large adjustment, we are searching for the best multiple linear regression model possible. Figure 1 illustrates the distribution before and after the transformation.

Market Share %

Market Share

Log Transformed Market Share %

Log Transformed Market Share
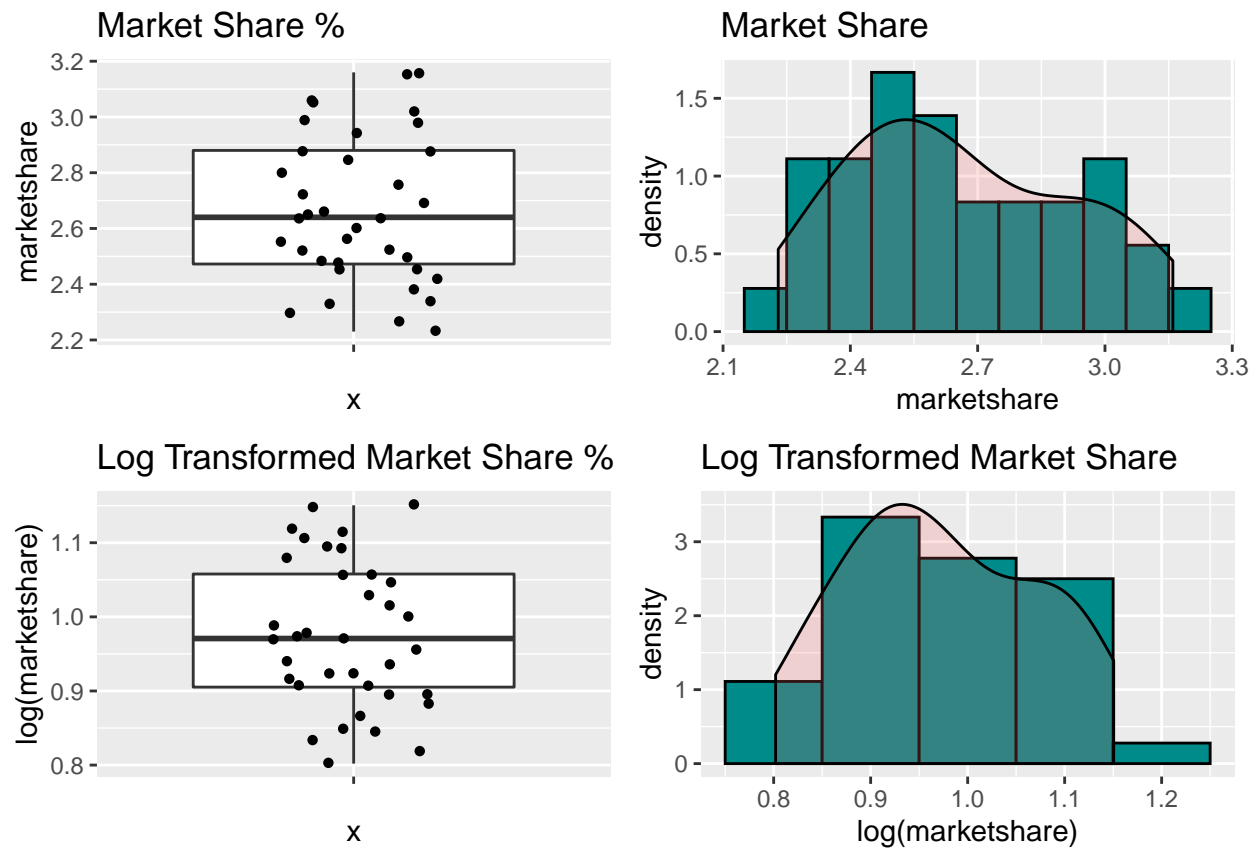
Analysis of Potential Predictors

Figure 2 and table 1 show the distribution of individual predictors. The first two charts are boxplots of price and gnrpoints. The price variable appears mostly symmetrical but slightly skewed. The gnrpoints variable appears symmetrical.Since discount and promotion are categorical variables, a bar chart displaying frequency was used to show counts. There are more data points with discounted products while the promotion variable appears more even. Since this data is from a consecutive 36 month period, the months are evenly spread out while the year spans the end of 1999 to the middle of 2002.

(a) price

(b) gnrpoints

(c) discount

(d) promotion

(e) month

(e) year
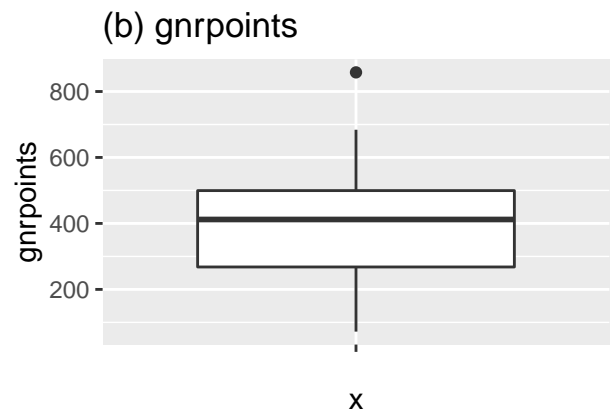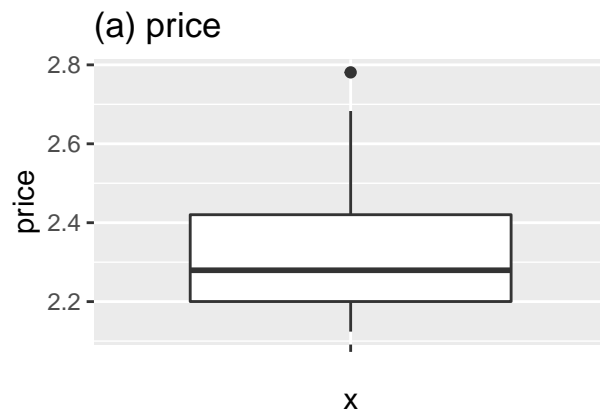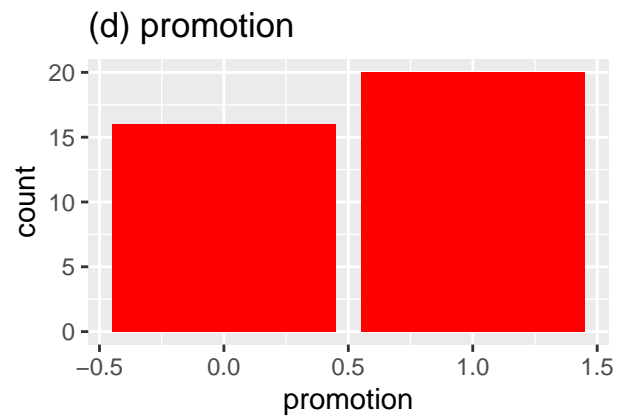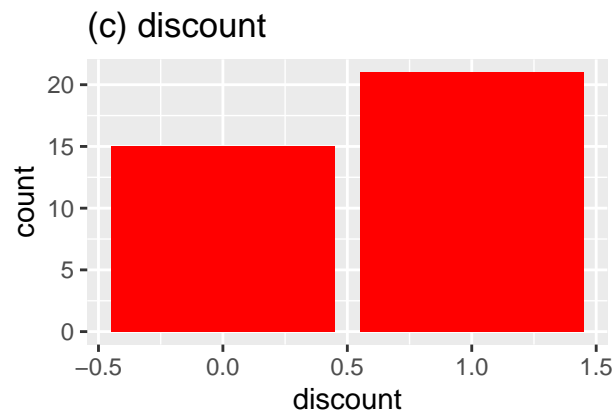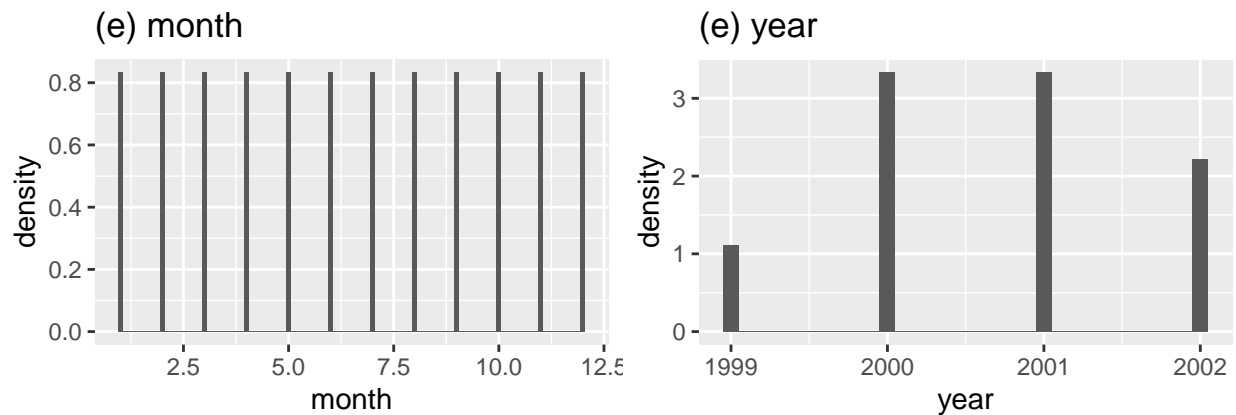
Table 1: Basic Statistics of Predictor Variables

	price	gnrpoints	discount	promotion	month	year
min	2.12400	72.00	0.00000	0.00000	1.000	1999.00000
max	2.78100	858.00	1.00000	1.00000	12.000	2002.00000
mean	2.32439	388.06	0.58333	0.55556	6.500	2000.66667
sd	0.16298	168.49	0.50000	0.50395	3.501	0.95618
var	0.02656	28389.71	0.25000	0.25397	12.257	0.91429
mean	2.32439	388.06	0.58333	0.55556	6.500	2000.66667
IQR	0.22000	231.50	1.00000	1.00000	5.500	1.00000

Effect of Price on Transformed Market Share Percent

Figure 3 illustrates the scatter plot computed by the simple linear regression model fit to data for the market share percent and average price of the product for the month. The entire dataset was used for the preliminary investigation. The model states that for every dollar increase in price, the market share decreases 0.109 percent. Note that since this is a log transformed response variable, the impact the predictor variables has on the response variable is not a unit for unit factor, but rather a log-scaled percentage.
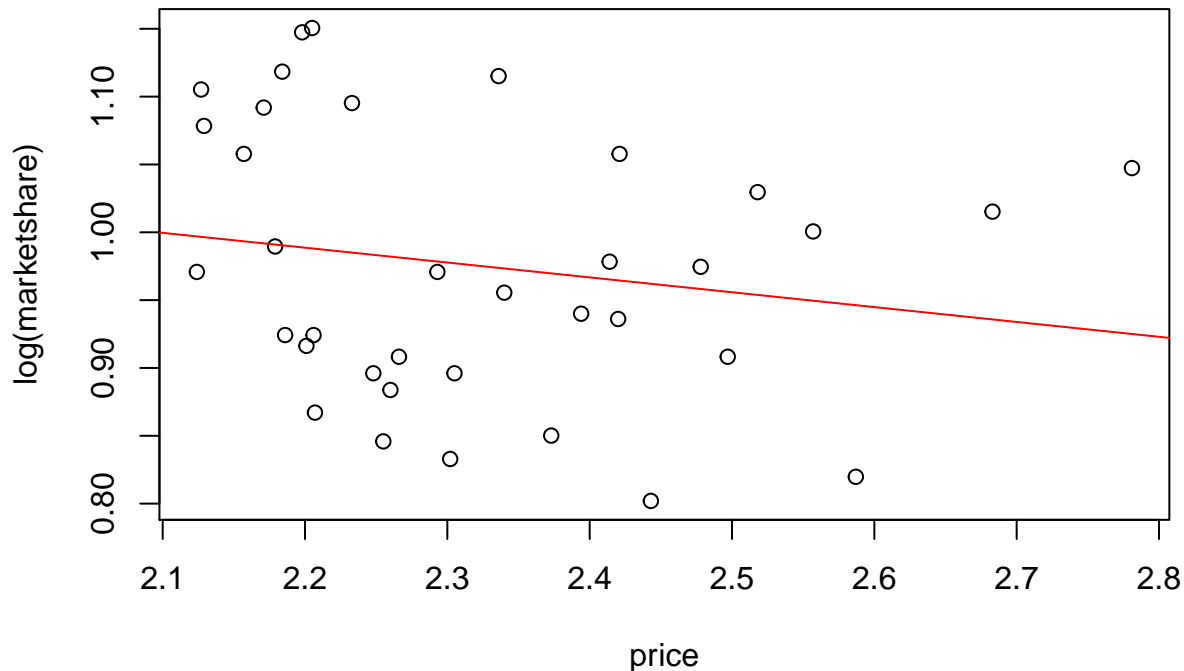
```
## NULL
```

Figure 3: Log Market Share v Price

Effect of GNR Points on Market Share Percent

Figure 4 illustrates the scatter plot of the simple linear regression model fit to the data for the market share percent and gnrpoints, which is an advertisng exposure index. This model states that for every dollar increase in price, the market share increases 0.0000453 percent. Note that since this is a log transformed response variable, the impact the predictor variables has on the response variable is not a unit for unit factor, but rather a log-scaled percentage. With a coefficient of determination of 0.6 percent, this is not a good model on it's own.
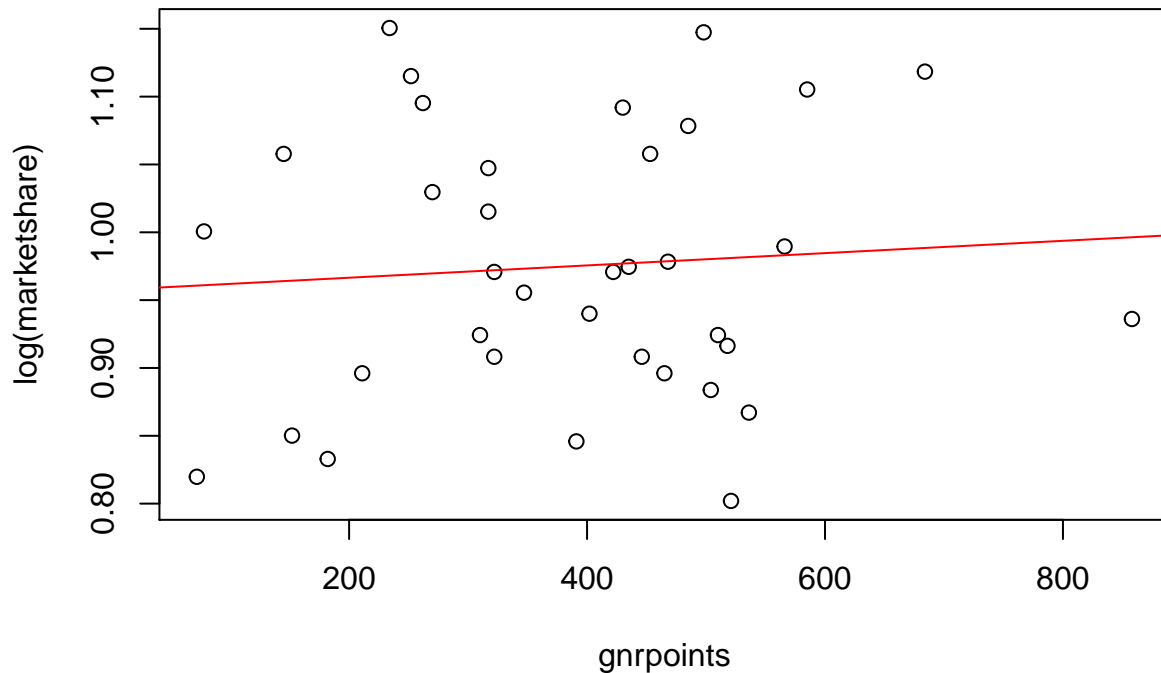
```
## NULL
```

Figure 4: Log Maket Share v GNRPoints

Effect of Discount on Market Share Percent

Figure 5 illustrates the scatter plot of the simple linear regression model fit to the data for the market share percent and whether a discount was applied to the product. This model states that when the product is on discount, the market share increases 0.158 percent. Note that since this is a log transformed response variable, the impact the predictor variables has on the response variable is not a unit for unit factor, but rather a log-scaled percentage. The impact the discount predictor variable has on the market share response variable is difficult to visualize as it is a binary variable, but a discount (variable = 1) does seem to correlate with a higher market share.
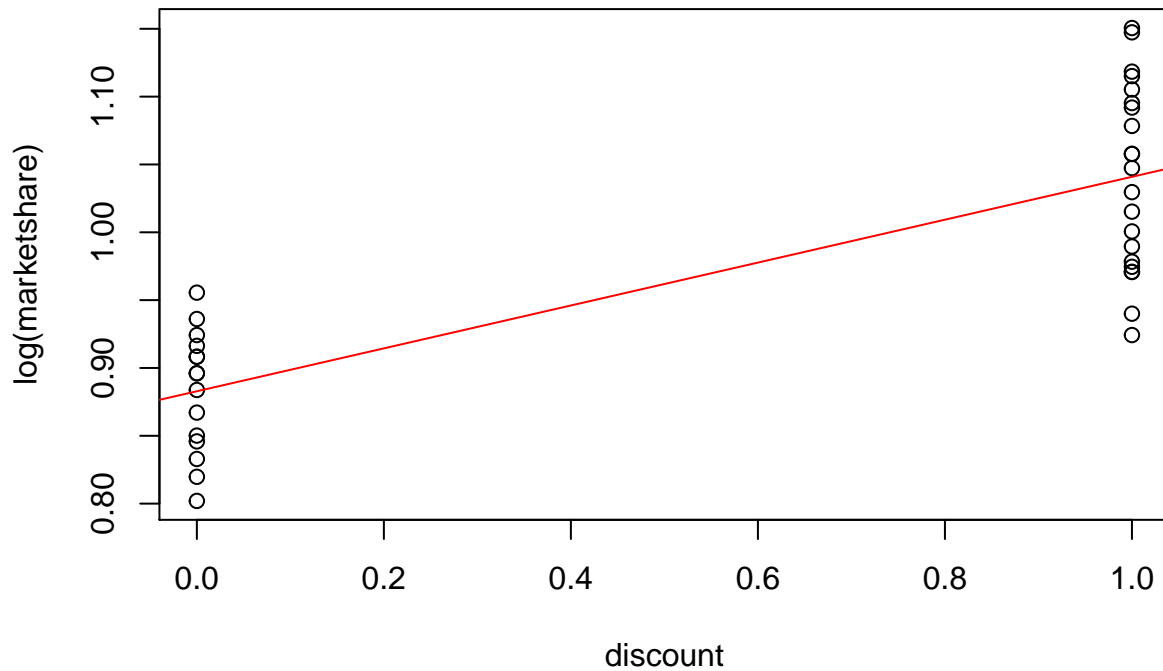
```
## NULL
```

Figure 5: Log Market Share v Discount

Effect of Promotion on Market Share Percent

Figure 6 illustrates the scatter plot of the simple linear regression model fit to the data for the market share percent and whether a promotion was applied to the product. This model states that during a promotion, the market share increases 0.059 percent. Note that since this is a log transformed response variable, the impact the predictor variables has on the response variable is not a unit for unit factor, but rather a log-scaled percentage.

```
## NULL
```

Figure 6: Log Market Share v Promotion

log(marketshare) vs promotion scatter plot with red regression line

Effect of Month on Market Share Percent

Figure 7 illustrates the scatter plot of the simple linear regression model fit to the data for the market share percent and what month the product was sold in. This model states that for every change in month starting with January as value 1, the market share increases 0.0011 percent. Note that since this is a log transformed response variable, the impact the predictor variables has on the response variable is not a unit for unit factor, but rather a log-scaled percentage.

```
## NULL
```
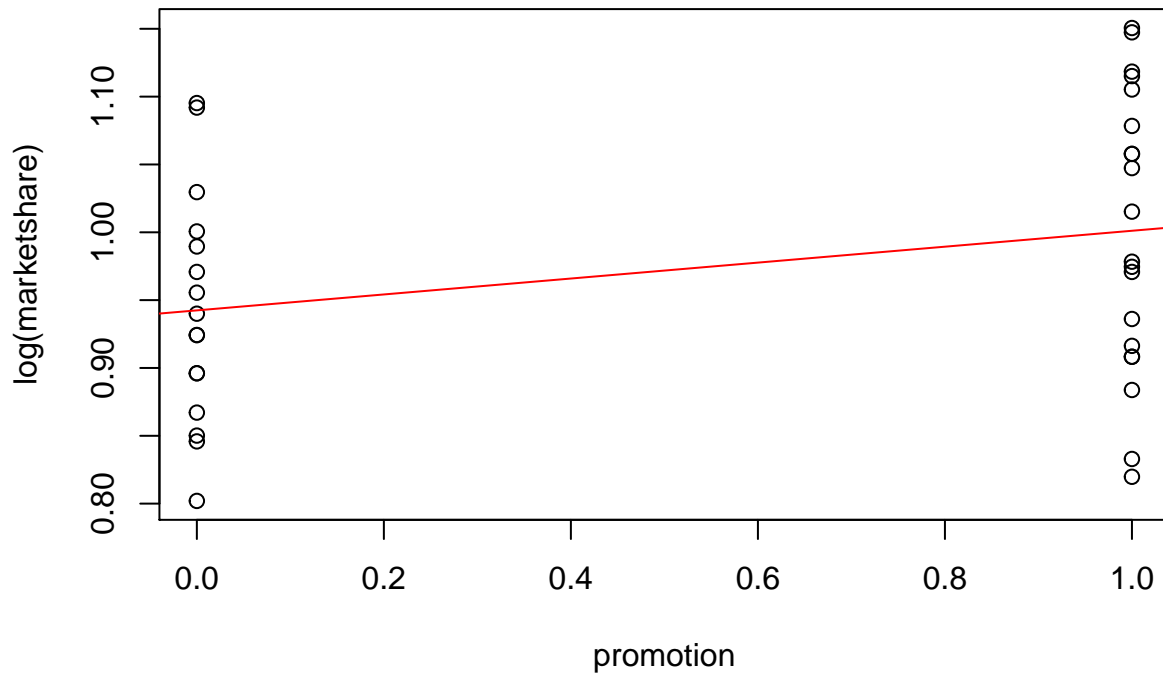
Figure 7: Log Market Share v Month

Effect of Year on Market Share Percent

Figure 8 illustrates the scatter plot of the simple linear regression model fit to the data for the market share percent and which year the product was sold in. This model states that as the years pass by, the market share decreases 0.0051 percent. Note that since this is a log transformed response variable, the impact the predictor variables has on the response variable is not a unit for unit factor, but rather a log-scaled percentage. Table 2 contains a breakdown of the individual models.
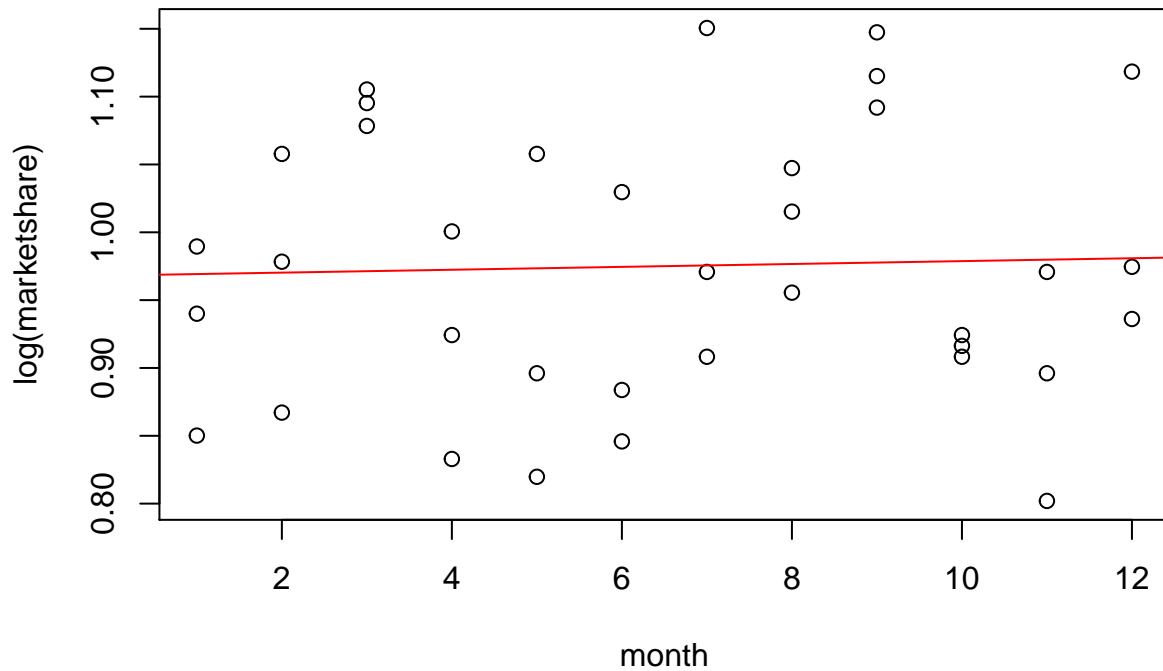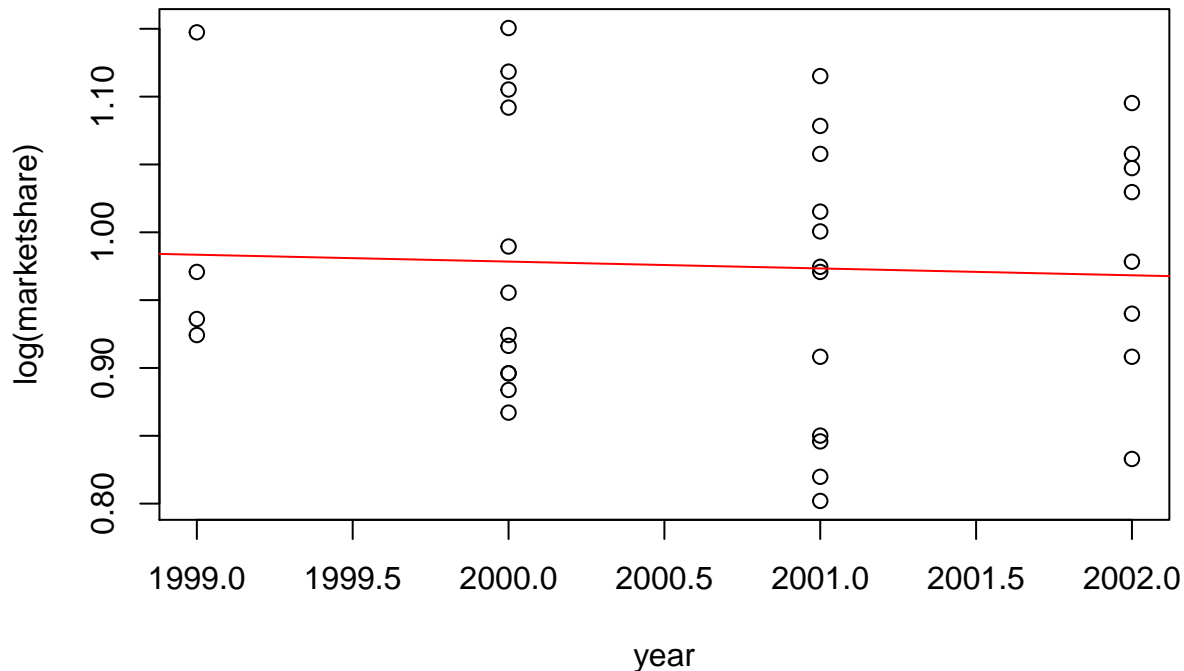
```
## NULL
```

Figure 8: Log Market Share v Year

Table 2: Analysis of Inidividual Predictor Models

Predictor	Beta1	R.2	SE	F.Stat
Price	-0.19000	0.0327	0.0980	1.150
GNRPoints	0.00045	0.0060	0.0998	0.210
Discount	0.15810	0.6420	0.0600	61.000
Promotion	0.05870	0.0890	0.0960	3.360
Month	0.00110	0.0014	0.1000	0.048
Year	-0.00500	0.0024	0.1000	0.082

Multicollinearity between Predictor variables :

Table 3 shows the correlation coefficients r between all the variables present in the dataset. From these values the response variable, market share, is highly correlated to discount and midly correlated to promotion. The rest of the variables do not seem that correlated to market share. Amongst the predictor variables, year and price appear mildly correlated but that may be triggered by natural price inflation. Although the correlation coefficients between predictors were not high, the Variance Inflation Factor (VIF) was checked. The VIF values remained below 2 which are significantly less than 10, thus concluding multicollinearity is certainly absent in this dataset.

Table 3: Correlation Coefficient

	price	gnrpoints	discount	promotion	month	year
price	1.00000	-0.38778	-0.00847	0.15696	0.09924	0.48138
gnrpoints	-0.38778	1.00000	-0.07263	0.14667	0.34878	-0.52162
discount	-0.00847	-0.07263	1.00000	0.15119	-0.17138	0.17928
promotion	0.15696	0.14667	0.15119	1.00000	0.24291	0.03953
month	0.09924	0.34878	-0.17138	0.24291	1.00000	-0.40967
year	0.48138	-0.52162	0.17928	0.03953	-0.40967	1.00000

Model Selection

Automatic Variable Selection Method

Automatic variable selection method is a starting point to eliminate the redundant variables from the model. The "regsubsets" function from "leap" package in R is used to for this purpose. The best model is selected on the basis of Mallow's Cp, BIC and adjusted coefficient of determination. The selection method selects the model which has least value for Mallow's Cp and BIC; while the model with the largest value of the adjusted coefficient of determination is considered best. In this case, the best model includes price, discount, promotion, and month as the 4 predictor variables. Refer to figure 10 for a brief summary of automatic selection method. Thus, the final model can be expressed as:
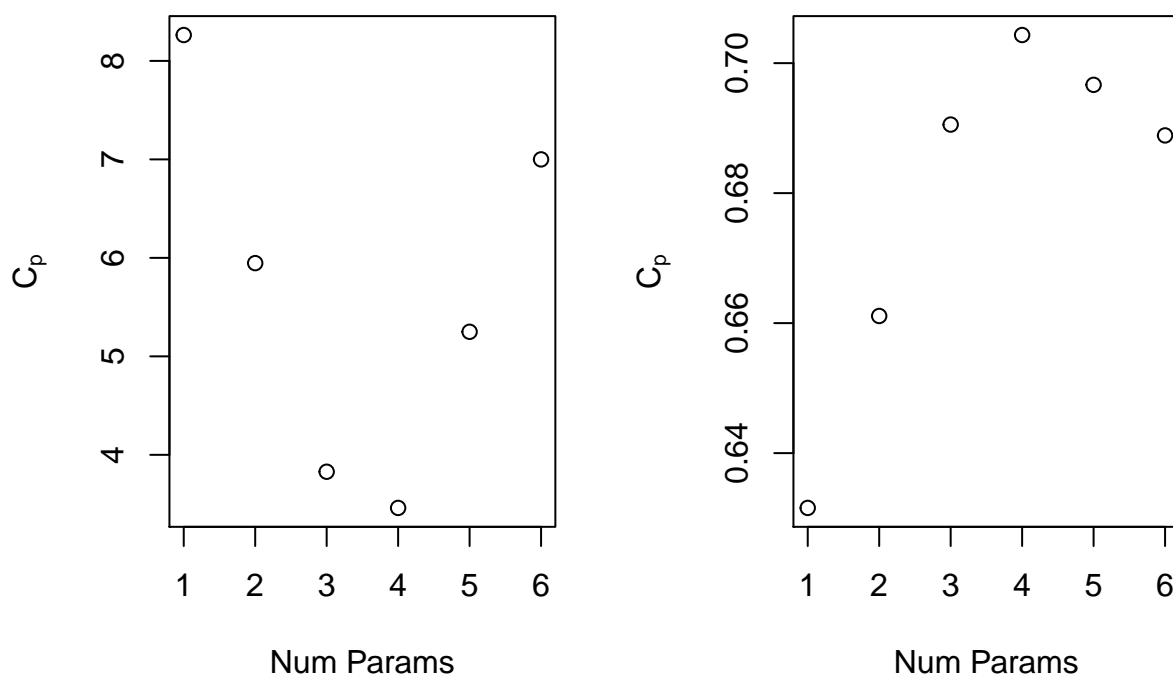
Residual Diagnostics

Figure 10 (a) shows the standardized residuals which confirms equal variance. No residuals fall more than 3 standard deviations away. Figure 10 (b) confirms the residuals follow normality. The independence of error terms is also quite evident in the sequence plot (figure 10(c)).
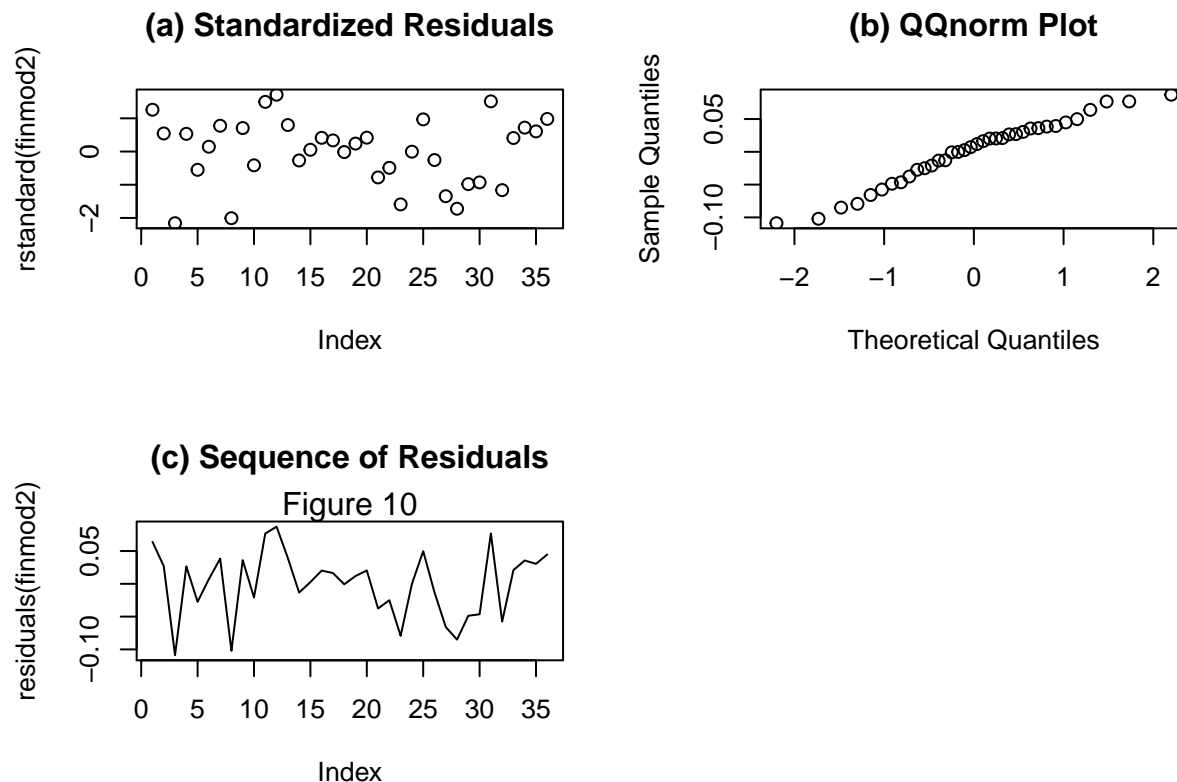
Goodness of Fit Test

To fulfill the objective of this study, scatterplots, tables, linear regression models, t-test, p-value, coefficient of determenation, adjusted R^2, Mallows Cp and BIC has been used.

All of the statistical analyses in this document will be performed using R version 3.6.1 (2019-07-05) R (R Core Team, 2018). R packages used will be maintained using the packrat dependency management system.

Results

For each of the variables in the dataset, we test if there exists a linear assocation with market share using two-tailed t-test individually. The t test considers the following hypothesis:

Null Hypothesis H_o : Beta_1 = 0

Alternative Hypothesis H_a : Beta_1 != 0

The decision is taken considering t* = b_1-Beta_1/SE(b_1)

where:

- t* is the test statistic for the t test
- b_1 is the observed slope coefficient
- Beta_1 is the expected slope coefficient of the fitted regression model
- SE(b_1) is the sampling variability of b_1

The t* statistic is tested agains t(1-alpha/2, df) where: - alpha is the level of significance = 0.05 - df is the degrees of freedom

If t* > t(1-alpha/2, df), H_0 is rejected else the decision is taken in the favour of H_a. The decision rule also considers the p-value and the R^2. If p-value <= alpha the dcision is to reject H_o else we fail to reject H_o. While considering the coefficient of determination, if the value is closer to 1 the association is considered strong, but if the value is closer to 0 then the model will not be considered a good fit.

Effect of Price on Market Share Percent

At significance level (alpha) = 0.05, the decison reached by the t test is to fail to reject H_o, thus, concluding that there does not exist evidence of a linear association between price and market share The t test explains that the model with just price as a predictor is able to explain only 3.3% of the unexplained variation in market share while the other 96.7% variation remains unexplained.

Effect of GNR Points on Market Share Percent

At significance level (alpha) = 0.05, the decison reached by the t test is to fail to reject H_o, thus, concluding that there does not exist evidence of a linear association between gnrpoints and market share. The t test explains that the model with just gnrpoints as a predictor is able to explain only 0.6% of the unexplained variation in market share while the other 99.4% variation remains unexplained.

Effect of Discount on Market Share Percent

At significance level (alpha) = 0.05, the decison reached by the t test is to reject H_o, thus, concluding that there exists evidence of a linear association between discount and market share. The t test explains that the model with just discount as a predictor is able to explain only 64.2% of the unexplained variation in market share while the other 35.8% variation remains unexplained.

Effect of Promotion on Market Share Percent

At significance level (alpha) = 0.05, the decison reached by the t test is to fail to reject H_o, thus, concluding that there does not exist evidence of a linear association between

promotion and market share. The t test explains that the model with just promotion as a predictor is able to explain only 9% of the unexplained variation in market share while the other 91% variation remains unexplained.

Effect of Month on Market Share Percent

At significance level (alpha) = 0.05, the decison reached by the t test is to fail to reject H_o, thus, concluding that there does not exist evidence of a linear association between month and market share. The t test explains that the model with just month as a predictor is able to explain only 0.2% of the unexplained variation in market share while the other 99.8% variation remains unexplained.

Effect of Year on Market Share Percent

At significance level (alpha) = 0.05, the decison reached by the t test is to fail to reject H_o, thus, concluding that there does not exist evidence of a linear association between year and market share. The t test explains that the model with just year as a predictor is able to explain only 0.3% of the unexplained variation in market share while the other 99.7% variation remains unexplained.

Primary Objective Results

The results from the individual t tests show that a linear association only exists between discount and market share. However, since the simple linear regression model with discount only explained 62.5% of variation, the model selection process began with considering all predictor variables in the dataset. Discount is clearly the strongest predictor as indicated by the R^2 value of 64.2% while the remaining variables generated R^2 values less than 10%. To confirm that the selected predictor variables of price, discount, promotion, and month can actually predict the market share percent, the predicted values are compared to the actual observed values from the dataset. At 95% prediction interval, it is found that the model does a decent job in estimating the market share.

Table 4: Head of Predicted v Observed

Predicted	Observed
1.08342	1.14740
1.06681	1.10526
0.90465	0.88377
0.86002	0.89609

Discussion and Conclusion

The estimated regression function from this data analysis would be:

Y-hat = 1.141 - 0.131X1 + 0.158X2 + 0.034X3 + 0.004X4

where, Y-hat is the logarithmic market share X1 is the price variable X2 is the discount categorical variable X3 is the promotion categorical variable X4 is the numerical month variable

The sigmaˆ2 = MSE = 0.0029. The other competing model for this dataset would be the same four predictor variables with the non-transformed market share percentage. This non-transformed model yielded a lower coefficient of determination (72.7% vs 73.8%) and a higher MSE (0.022 vs 0.0029). The decision to select the log-transformed model was based off of these values along with the slight reduction in market share skewness. This study shows that the predictor variables of price, discount, promotion, and month can predict the market share of a packaged food product. All statistical analysis is conducted at 95% confidence interval and at 0.05 significance level. All potential predictors in this data set were explored individually and the two-tailed t test was conducted on each of the fitted linear regression model. Although only one predictor yielded a linear association, all predictors were initially considered to generate the best model possible. Table 4 shows the estimated regression coefficient, the standard error, t value, and p-value associated with each of hte predictors. The coefficient of determination, MSE, and F statistic of the final model is also included. Table 5 shows the ANOVA table to for the final model having SSE, SSR, MSE, MSR, F values and the corresponding p-values. Exploring the test statistics, it was found that the strongest association among variables is best explained by this final model, which was able to explain about 74% of the unexplained variation in the logarithmic market share percentage. This study shows that the price of the packaged food product exhibits a negative association with the response variable market share percentage. All other predictors have a positive association with discount having the largest impact. With a fairly high coefficient of determination of 74% and well behaved residuals, it can be concluded that the model does a fairly good job at predicting the market share. At 95% prediction interval, the market share from the predicted model was compared with the observed values and it confirmed the model can be used to predict the market share percent. No presence of outliers were detected.

Table 5: Statistics Table for Final Model

Variable	Estimate	Std..Error	t.value	Pr...t..
Intercept	1.141	0.131	8.71	0.000
Price	-0.131	0.056	-2.33	0.027
Discount	0.158	0.019	8.39	0.000
Promotion	0.034	0.019	1.80	0.082
Month	0.004	0.003	1.58	0.125

Table 6: ANOVA for Final Regression Model

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
price	1	0.01115	0.01115	3.8749	0.05801
discount	1	0.21783	0.21783	75.7219	0.00000

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
promotion	1	0.01520	0.01520	5.2854	0.02841
month	1	0.00717	0.00717	2.4911	0.12465
Residuals	31	0.08918	0.00288	NA	NA

Appendix: R-code

```r
# libraries used
library(knitr)
library(readxl)
library(formatR)
library(caret)
library(Hmisc)
library(stargazer)
library(xtable)
library(leaps)
library(ggplot2)
library(gridExtra)

# load marketshare dataset
market_share <- read_excel("C:/Users/Brenton/Desktop/KUMC/STAT823/final/market_share.xls
attach(market_share)

# change month variable to numeric values (jan = 1, feb = 2,
# etc)
market_share$month <- match(market_share$month, month.abb)

# evaluate response variable Y, marketshare
plot1 <- ggplot(market_share, aes(x = "", y = marketshare)) +
    geom_boxplot() + geom_jitter(shape = 16, position = position_jitter(0.2)) +
    ggtitle("Market Share %")

plot2 <- ggplot(market_share, aes(x = marketshare)) + geom_histogram(aes(y = ..density.
    binwidth = 0.1, color = "black", fill = "darkcyan") + geom_density(alpha = 0.2,
    fill = "#FF6666") + ggtitle("Market Share")

# execute log transformation due to slight skew
market_share$transformedshare <- log(market_share$marketshare)

plot3 <- ggplot(market_share, aes(x = "", y = log(marketshare))) +
    geom_boxplot() + geom_jitter(shape = 16, position = position_jitter(0.2)) +
    ggtitle("Log Transformed Market Share %")
```

```
plot4 <- ggplot(market_share, aes(x = log(marketshare))) + geom_histogram(aes(y = ..den
    binwidth = 0.1, color = "black", fill = "darkcyan") + geom_density(alpha = 0.2,
    fill = "#FF6666") + ggtitle("Log Transformed Market Share")

grid.arrange(plot1, plot2, plot3, plot4, ncol = 2, nrow = 2)

# visualize potential predictors
box1 <- ggplot(market_share, aes(x = "", y = price)) + geom_boxplot() +
    ggtitle("(a) price")

box2 <- ggplot(market_share, aes(x = "", y = gnrpoints)) + geom_boxplot() +
    ggtitle("(b) gnrpoints")

scat1 <- ggplot(market_share, aes(discount)) + geom_bar(fill = "red") +
    ggtitle("(c) discount")

scat2 <- ggplot(market_share, aes(promotion)) + geom_bar(fill = "red") +
    ggtitle("(d) promotion")

box5 <- ggplot(market_share, aes(x = month)) + geom_histogram(aes(y = ..density..),
    binwidth = 0.1, ) + ggtitle("(e) month")

box6 <- ggplot(market_share, aes(x = year)) + geom_histogram(aes(y = ..density..),
    binwidth = 0.1, ) + ggtitle("(f) year")

grid.arrange(box1, box2, ncol = 2, nrow = 2)
grid.arrange(scat1, scat2, ncol = 2, nrow = 2)
grid.arrange(box5, box6, ncol = 2, nrow = 2)


# marketshare ~ price
pricemod <- lm(log(marketshare) ~ price, market_share)
plot(log(marketshare) ~ price, market_share) + theme_classic()
abline(pricemod, col = "red")
summary(pricemod)

# marketshare ~ gnrpoints (advertising index)
gnrmod <- lm(log(marketshare) ~ gnrpoints, market_share)
plot(log(marketshare) ~ gnrpoints, market_share) + theme_classic()
abline(gnrmod, col = "red")
summary(gnrmod)

# marketshare ~ discount
dismod <- lm(log(marketshare) ~ discount, market_share)
```

```r
plot(log(marketshare) ~ discount, market_share) + theme_classic()
abline(dismod, col = "red")
summary(dismod)

# marketshare ~ promotion
promomod <- lm(log(marketshare) ~ promotion, market_share)
plot(log(marketshare) ~ promotion, market_share) + theme_classic()
abline(promomod, col = "red")
summary(promotion)

# marketshare ~ month
monmod <- lm(log(marketshare) ~ month, market_share)
plot(log(marketshare) ~ month, market_share) + theme_classic()
abline(monmod, col = "red")
summary(month)

# marketshare ~ year
yearmod <- lm(log(marketshare) ~ year, market_share)
plot(log(marketshare) ~ year, market_share) + theme_classic()
abline(yearmod, col = "red")
summary(yearmod)

# Multicollinearity check
cor(market_share)
VIF(marketshare ~ .)

# Subsetting Results for Model Selection
subsetted <- regsubsets(log(marketshare) ~ price + gnrpoints +
    discount + promotion + month + year, market_share)
sumsubset <- summary(subsetted)
par(mfrow = c(1, 2))
plot(1:6, sumsubset$cp, xlab = "Num Params", ylab = expression(C[p]))
plot(1:6, sumsubset$adjr2, xlab = "Num Params", ylab = "Adj R^2")
mtext("Figure 9: Model Selection Criteria")

finmod2 <- lm(log(marketshare) ~ price + discount + promotion +
    month, market_share)

# Residual Diagnostics
par(mfrow = c(2, 2))
plot(rstandard(finmod2), main = "(a) Standardized Residuals")
qqnorm(residuals(finmod2), main = "(b) QQnorm Plot")
plot(residuals(finmod2), type = "l", main = "(c) Sequence of Residuals")
plot(fitted(finmod2), residuals(finmod2))
```

```r
abline(h = 0)

# Predictive Power
predict(finmod2, data.frame(price = 2.186, discount = 0, promotion = 0,
    month = 10), level = 0.95, interval = "prediction")

predict(finmod2, data.frame(price = 2.305, discount = 0, promotion = 0,
    month = 5), level = 0.95, interval = "prediction")

predict(finmod2, data.frame(price = 2.781, discount = 1, promotion = 1,
    month = 8), level = 0.95, interval = "prediction")

predict(finmod2, data.frame(price = 2.184, discount = 1, promotion = 1,
    month = 12), level = 0.95, interval = "prediction")
```

References

Hayes. (2020), Market Share. https://www.investopedia.com/terms/m/marketshare.asp

R Core Team (2018), R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria, URL https://www.R-project.org/.

Dean, A., Voss, D., and Draguljic, D. (2017), Design and Analysis of Experiments, 2nd ed., Springer.

Kuehl, R. O. (2000), Design of Experiments: Statistical Principles of Research Design and Analysis, 2nd ed., Brooks/Cole, Cengage Learning.

Kutner, M. H., Nachtsheim, C. J., Neter, J., and Li, W. (2014), Applied Linear Statistical Models, 5th ed., McGraw-Hill Irwin.

Bibliography

R Core Team (2018), *R: A Language and Environment for Statistical Computing*, R
Foundation for Statistical Computing, Vienna, Austria, URL
https://www.R-project.org/.