Group Assignment on **Advanced Statistics**

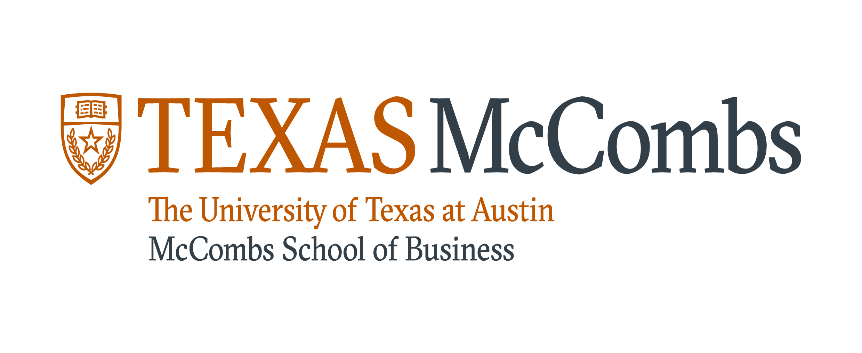
Submitted by, members of Group 6,

**Nivitha**

**Manoj Vairavan R**

**Surya Prakash Raju**

**Sajin Madhavan**



**1st Solution**

**Analyses of consideration behaviour of 12 different branded cereals using Factor Analysis**

1. Initially the file is being loaded and basic validations are performed

data <- read.csv(file.choose())

attach(data)

# First validation

str(data)

summary(data)

head(data)

tail(data)

1. Data are in the range of 1-5, so scaling is not required. Few ratings are out of range 5, so they

#Replacing 6 with 5

# incorrect ratings are handled and here 6th value is replaced by 5

data[data==6]<-5

#Soggy and Boring are negative variables, so the rating order are reversed

data$Soggy<-6-data$Soggy

data$Boring<-6-data$Boring

# Getting matrix of 25 different columns

cereal\_matrix <- as.matrix(data[,2:26])

# install package - nFactor

install.packages("nFactors")

install.packages("corrplot")

library(nFactors)

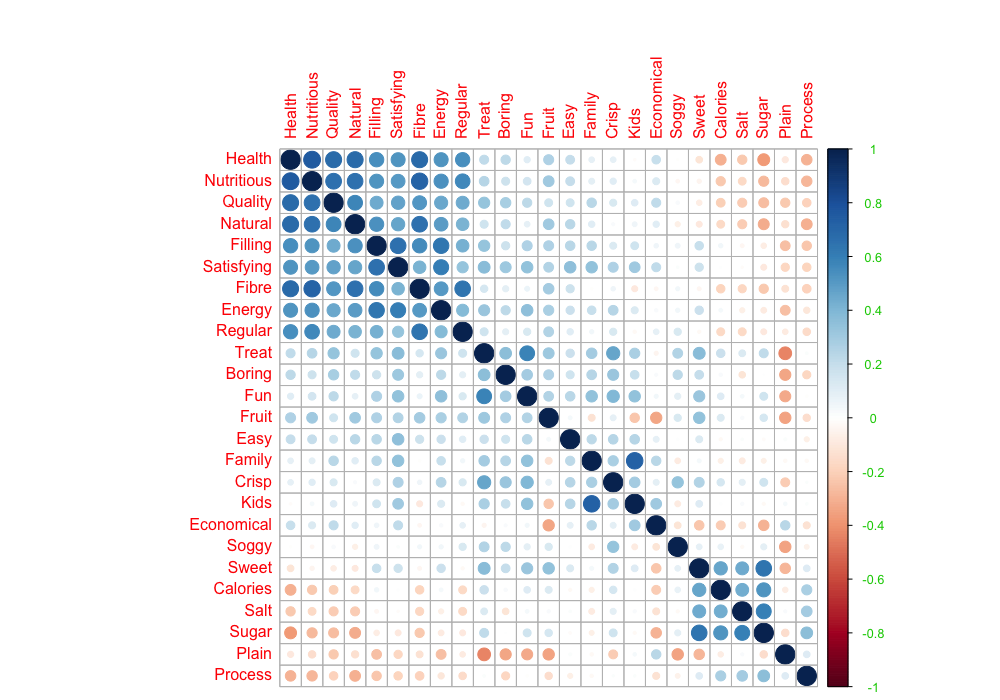
library(corrplot)

1. Correlation matrix is drawn to check how the variables are related and to find the eigen matrix

# Check for the correlation between 25 factors

cor=cor(cereal\_matrix)

corrplot(cor,method="circle",order="FPC")



# **Eigen value** calculates the sum of squares of correlation - column wise

Eigen <- eigen(cor(cereal\_matrix))

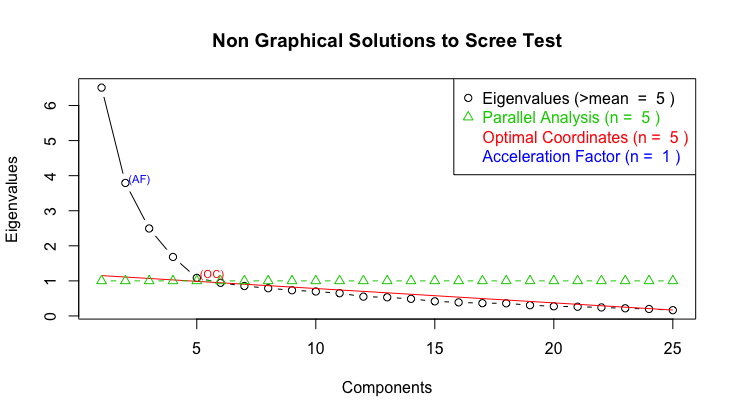
Eigen

# **SCREE PLOT**

# Plot Eigen values to determine the Scree Plot

Scree <- nScree(x=Eigen$values)

plotnScree(Scree)



Generally eigen value of >1 is chosen, so from the graph we are selecting 4. We are not selecting 5th value as it is almost touching 1.

1. Now, let us try to group the vactor using **Factor analyses** with"no rotation"

factor<-factanal(cereal\_matrix,4,n.obs=235,rotation="none")

print(factor, digits=2, cutoff= 0.3, sort= FALSE)

Since we see the variables are not clearly classified into different factors, we will do

**Varimax** Rotation for better classification without change in the variance

rotated\_factor<-factanal(cereal\_matrix,factors=4,n.obs=235,rotation="varimax")

print(rotated\_factor, digits=2, cutoff= 0.3, sort= TRUE)

Rotated factor with 4 value gives reasonable classifications of 25 variables. The model explains around 0.50% of data only and p value is less than alpha(0.05).

In order to get a high cumulative variance, model is further evaluated by removing variables with low loading. The factor remains 4 but the p value increases to 0.0001 from 2.3e-14

rotated\_factor\_v1<-factanal(cereal\_matrix[,c(-20,-21,-16,-5,-11,-12,-23,-24)],factors=4,n.obs=235,rotation="varimax")

print(rotated\_factor\_v1, digits=2, cutoff= 0.3, sort= TRUE)

1. **Naming Factors**

Factor 1 - Healthier Cereal

Factor 2 determines Tasty Cereal

Factor 3 determines Appealing

Factor 4 identifies Family Choice

1. **Visualizing 4 factors**

plot(rotated\_factor$loadings[,-c(3,4)], type= "n",cex=0.7)

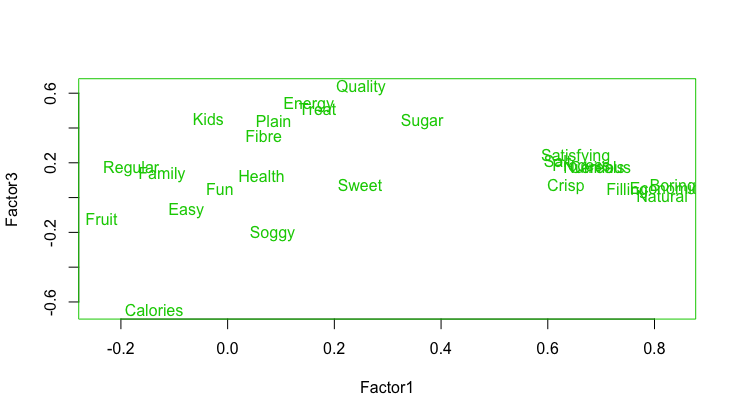
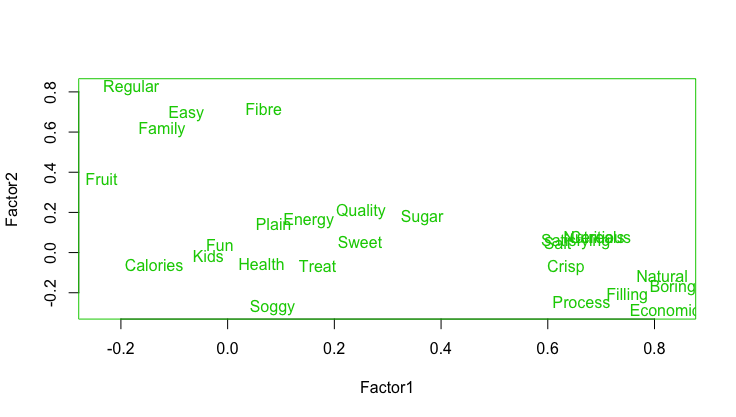
text(rotated\_factor$loadings[,-c(3,4)], labels=names(data))

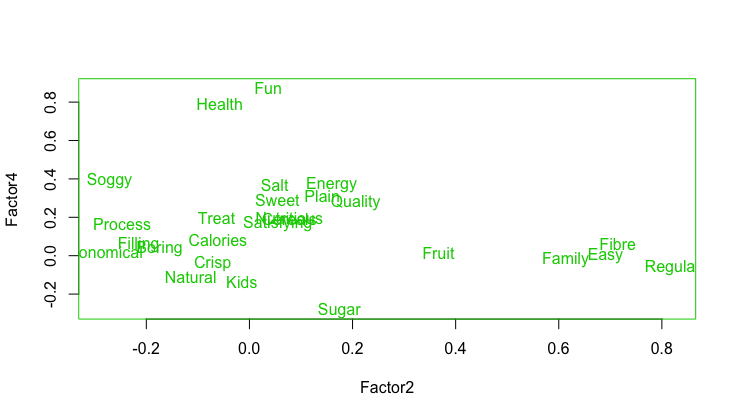
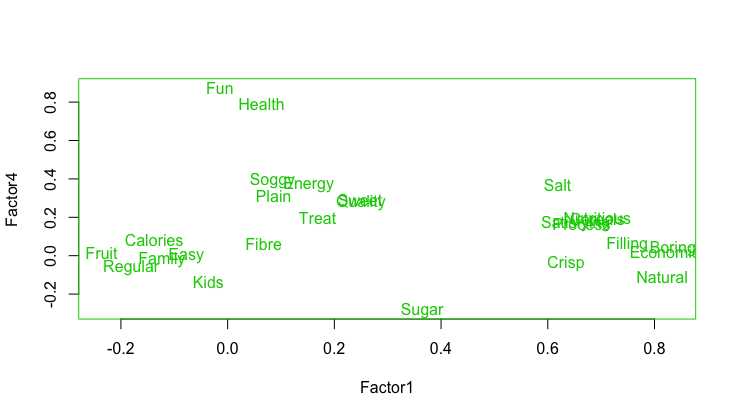
plot(rotated\_factor$loadings[,-c(2,4)], type= "n",cex=1.0)

text(rotated\_factor$loadings[,-c(2,4)], labels=names(data))

plot(rotated\_factor$loadings[,-c(1,3)], type= "n",cex=1.0)

text(rotated\_factor$loadings[,-c(1,3)], labels=names(data))





1. Characterization of variables is performed using loading retrieved from factors

and the factor scores are generated from loadings like below.

# Loading factors into loading variable

loadings <- rotated\_factor$loadings

#find the factor scores to characterize the variables from factors identified

factor\_scores<-factor.scores(cereal\_matrix,loadings)

scores<-factor\_scores$scores

temp=cbind(data["Cereals"],scores)

names(temp)<-c("Cereals","Healthy","Tastier","Appealingess","Family Choice")

cereal\_factor\_scores=aggregate(temp[c(2,3,4,5)], by=list(Category=temp$Cereals), FUN=sum)

cereal\_factor\_scores

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Category** | **Healthy** | **Tastier** | **Appealing** | **Family Choice** |
| AllBran | 5.8566423 | -3.947096 | -7.831832 | -14.11968 |
| CMuesli | 6.4905757 | 6.72477 | 9.622101 | -3.362169 |
| CornFlakes | -14.979576 | 2.933266 | 1.765407 | 14.601523 |
| JustRight | -0.4295564 | 3.510803 | 8.245765 | -6.906638 |
| Komplete | 7.5390967 | 3.035992 | 8.977481 | -15.745752 |
| NutriGrain | -10.481539 | 20.527754 | 8.058152 | 14.036137 |
| PMuesli | 11.6004691 | 12.983277 | 10.772268 | -9.309564 |
| RiceBubbles | -25.579789 | -10.647209 | 3.409976 | 13.472102 |
| SpecialK | -7.5089385 | -5.301682 | -1.480165 | 3.682074 |
| Sustain | 7.7274152 | -4.793838 | 11.788929 | -4.970729 |
| Vitabrit | 10.0367283 | -14.989367 | -25.799144 | 7.137948 |
| Weetabix | 9.7284719 | -10.03667 | -27.528939 | 1.484748 |

**Conclusion**

* Cereals PMuseli, Vitabrit, Weetabix, Sustain, Komplete are more healthier
* NutriGrain, Pmuseli are more tastier.
* Pmuseli, CMuseli, JustRight and Komplete are more appealing cereal.
* CornFlakes, Nutrigrain and RiceBubbles are most preferred by family even when those are not healthier
* SpecialK has very less ratings comparatively.

**2nd Solution**

**Regression analysis of Price determination of Leslie salt company for Sale**

leslie<-read.csv(file.choose())

leslie

attach(leslie)

summary(leslie)

# First Validation

head(leslie)

tail(leslie)

leslie\_orginal<-leslie

leslie<-leslie\_orginal

**Nature of the variables**

* Looking at the data, the variables are continuous and categorical/factor.
* Price of the leslie salt company area has to be determined which will be the dependent variable and

Size, Elevation, Sewer, Distance, Sewer, Date, County and Flood are independent variables.

leslie$County <- as.factor(leslie$County)

leslie$Flood <- as.factor(leslie$Flood)

leslie$Date<-Date\*sign(Date)

**Transformation of Variables**

* County and Flood indicates TRUE/FASE and hence those have to be converted into factor.
* Date is calculated from past to present and hence it is negative. But for modelling purpose, we will change it to positive.

**Multiple Linear Regression**

For the linear regression to be a better model, we have to check for

**1. Outliers**

**2. Normality of data**

**3. Multicollinearity of data**

**4. Autocorrelation of data**

**Durbin Watson test is used, which assumes the H0 as there is no autocorrelation between residuals and H1 assumes there is auto correlation**

**5. Homoscedasticity**

1. **Outliers & Scaling**

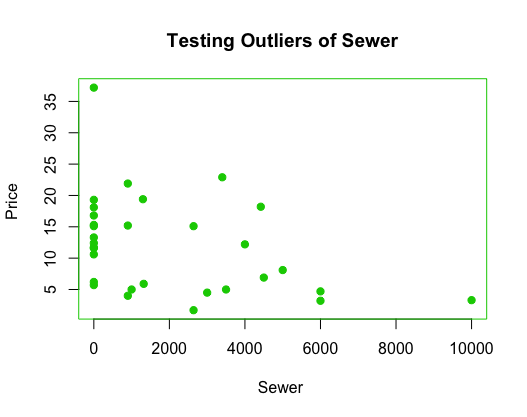
Scaling of data is required for Sewer as it is in the range of (0,10000) and for Size as it s in the range of (0,1695).If scaling is not done then the model will be biased towards unscaled data. We can verify using checking the outliers. Outliers existence can be identified by box plot or scatter plot like below

plot(leslie$Sewer, leslie$Price, main="Testing Outliers of Sewer",

xlab="Sewer ", ylab="Price ", pch=19)

plot(leslie$Size, leslie$Size, main="Testing Outliers of Size",

xlab="Size ", ylab="Price", pch=19)

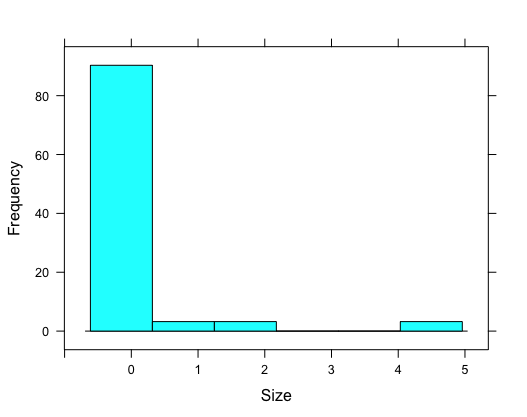
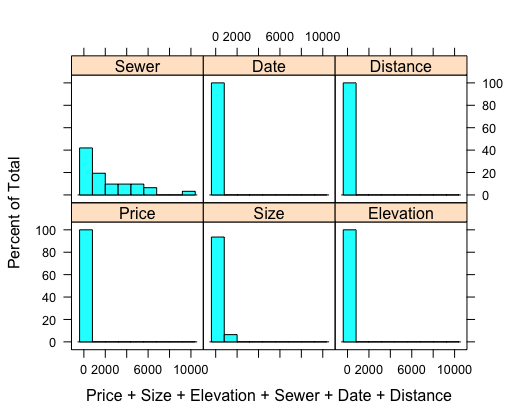


Size has clear outliers in the range of 900 and one around 1700

And sewer has a outlier at 10000 and from the histogram, we can see that Sewer and Size has the values distributed and whereas other variables are having values fitted into single ranged bin.

histogram(~Price+Size+Elevation+Sewer+Date+Distance,data=leslie)

histogram(Size,ylab="Frequency",xlab="Size")



So, scaling of size and sewer varibles are performed.

leslie$Size<-scale(leslie$Size)

leslie$Sewer<-scale(leslie$Sewer)

1. **Normality Test of the variables**

with(leslie, shapiro.test(Size)) # Needs scaling as p value is less than alpha 0.05

**Shapiro-Wilk normality test**

**data: Size**

**W = 0.40531, p-value = 4.108e-10**

with(leslie, shapiro.test(Sewer)) # Needs scaling as alpha is less than 0.05

**Shapiro-Wilk normality test**

**data: Sewer**

**W = 0.80027, p-value = 5.221e-05**

with(leslie, shapiro.test(Elevation))

**Shapiro-Wilk normality test**

**data: Elevation**

**W = 0.85914, p-value = 0.000798**

with(leslie, shapiro.test(Date))

**Shapiro-Wilk normality test**

**data: Date**

**W = 0.91472, p-value = 0.01714**

with(leslie, shapiro.test(Distance))

**Shapiro-Wilk normality test**

**data: Distance**

**W = 0.87773, p-value = 0.002096**

1. **Regression Model**

library(car)

names(leslie)

model<-lm(Price~County+Size+Elevation+Sewer+Date+Flood+Distance,data=leslie)

From the model developed, the fit of the model can be estimated based on plotting fitted and residual values. Values should be around the mean, there should be any outliers and no U or cone shaped pattern in the graph.

fitted <- fitted(model)

residual <- residuals(model)

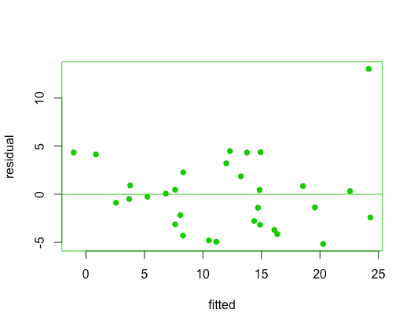
#Merge the fitted and residual values

leslieFR <- cbind(leslie, fitted, residual)

#Plot the actual versus fitted values in a plot

with(leslieFR, plot(fitted,residual, pch=19, cex=0.9))

abline(a=0,b=0)

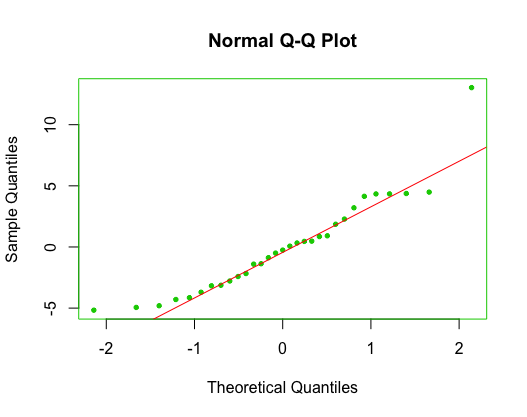


There are no pattern found in residual plotting and hence we can say that the model fits linear regression.

#Residuals on a QQplot

with(leslieFR, qqnorm(residual, pch=19, cex=0.6))

with(leslieFR, qqline(residual, col='red'))



1. **Multicollinearity**

To check the **variance inflation factor(VIF**). VIF > 10 indicates that multicollinearity exists.

**vif(model)**

**County Size Elevation Sewer Date Flood Distance**

**4.995597 2.003925 1.649759 1.635122 2.174889 1.907942 3.623612**

All variables are less than 10 and hence the model doesn't have multicollinearity.

1. **Autocorrelation**

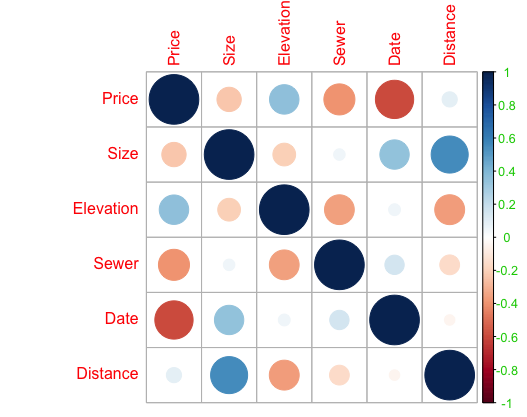
**Correlation**

Correlation should be less than or nearer to zero for any variable to be independent of other.

corr=cor(leslie[-c(2,7)])

library(corrplot)

corrplot(corr,method="circle")



From the graph, we can see that no variables are highly correlated but there are positive and negative correlation.

**Durbin-Watson test**

dwtest(model)

DW value from 0 to 4 indicates there is no autocorrelation. The model has value of 2.4 and so, we can conclude that there is no autocorrelation.

**Durbin-Watson test**

**data: model**

**DW = 2.4267, p-value = 0.7618**

**alternative hypothesis: true autocorrelation is greater than 0**

1. **Homoscedasticity – To test whether the values are normally distributed from the mean**

We can use Goldfeld Quandt test method to test the homoscedasticity, where HO represent data satisfies homoscedasticity whereas H1 represents that data is not suitable for homoscedasticity. P value is greater than 0.05 and hence we accept the null hypothesis that data is having homoscedasticity.

gqtest(model)

**Goldfeld-Quandt test**

**data: model**

**GQ = 3.3534, df1 = 8, df2 = 7, p-value = 0.06421**

**alternative hypothesis: variance increases from segment 1 to 2**

1. **Model Summary**

**summary(model)**

**Call:**

**lm(formula = Price ~ County + Size + Elevation + Sewer + Date +**

**Flood + Distance, data = leslie)**

**Residuals:**

**Min 1Q Median 3Q Max**

**-5.169 -2.957 -0.256 2.070 13.031**

**Coefficients:**

**Estimate Std. Error t value Pr(>|t|)**

**(Intercept) 2.364e+01 3.829e+00 6.174 2.68e-06 \*\*\***

**County1 -8.789e+00 3.652e+00 -2.407 0.024532 \***

**Size -6.043e-03 3.501e-03 -1.726 0.097702 .**

**Elevation 5.193e-01 2.386e-01 2.177 0.040030 \***

**Sewer -9.573e-04 4.169e-04 -2.296 0.031126 \***

**Date -8.508e-02 4.865e-02 -1.749 0.093646 .**

**Flood1 -1.202e+01 2.989e+00 -4.020 0.000536 \*\*\***

**Distance 1.858e-01 3.395e-01 0.547 0.589386**

**---**

**Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1**

**Residual standard error: 4.431 on 23 degrees of freedom**

**Multiple R-squared: 0.747, Adjusted R-squared: 0.67**

**F-statistic: 9.703 on 7 and 23 DF, p-value: 1.351e-05**

**Let Y be the Price and X1 is County, X2 is Size, X3 is Elevation, X4 is Sewer, X5 is Date, X6 is Flood and X7 is Distance.**

* **Y= 23.64-8.789X1-0.00604X2+0.519x3-0.00096X4-0.08508X5-12.0200X6+0.18580X7**
* **Here b0 is 23.64 and the coefficients of County is -8.789, -0.00604 for Size, 0.519 for Elevation, -0.00096 for Sewer, -12.02 for flood, 0.1858 for Distance.**
* **From the model derived, we can say that when County alone considered, there will 8.789 decrease in Price.**
* **Likewise when Size alone considered keeping other variables constant, there will be 0.006043 decrease in Price.**
* **Sewer also negative impact on Price by -0.00096**
* **When sold date of nearby property is considered, there will 0.0850 decrease in Price**
* **If Elevation increases, then the price seems to increase by 0.519 and Flood also have negative impact on Price by 0.185.**
* **Multiple R squared is 0.747, which explains 74.7% of the price is determined by the independent variables.**
* **p value is less than 0.05 and hence null hypothesis is rejected stating at least one Beta is different. Hence we conclude stating that there exists robustness in the model.**

**3rd Solution**

**All Greens Franchise**

Explain the importance of X2, X3, X4, X5, X6 on Annual Net Sales, X1.

The data (X1, X2, X3, X4, X5, X6) are for each franchise store.

X1 = annual net sales/$1000

X2 = number sq. ft./1000

X3 = inventory/$1000

X4 = amount spent on advertising/$1000

X5 = size of sales district/1000 families

X6 = number of competing stores in district

R codes

setwd("C:/Users/win10/Desktop/Advanced statistics")

mydata=read.csv("Dataset\_All Greens Franchise-1.csv",header=TRUE)

attach(mydata)

names(mydata)

mydata

model=lm(X1~X2+X3+X4+X5+X6)

summary(model)

X1 X2 X3 X4 X5 X6

1 231 3.0 294 8.2 8.2 11

2 156 2.2 232 6.9 4.1 12

3 10 0.5 149 3.0 4.3 15

4 519 5.5 600 12.0 16.1 1

5 437 4.4 567 10.6 14.1 5

6 487 4.8 571 11.8 12.7 4

> str(model)

'data.frame': 27 obs. of 6 variables:

$ X1: num 231 156 10 519 437 487 299 195 20 68 ...

$ X2: num 3 2.2 0.5 5.5 4.4 ...

$ X3: int 294 232 149 600 567 571 512 347 212 102 ...

$ X4: num 8.2 6.9 3 12 10.6 ...

$ X5: num 8.2 4.1 4.3 16.1 14.1 ...

$ X6: int 11 12 15 1 5 4 10 12 15 8 ...

> summary(model)

X1 X2 X3 X4 X5

Min. : 0.5 Min. :0.500 Min. :102.0 Min. : 2.50 Min. : 1.600

1st Qu.: 98.5 1st Qu.:1.400 1st Qu.:204.0 1st Qu.: 4.80 1st Qu.: 4.500

Median :341.0 Median :3.500 Median :382.0 Median : 8.10 Median :11.300

Mean :286.6 Mean :3.326 Mean :387.5 Mean : 8.10 Mean : 9.693

3rd Qu.:450.5 3rd Qu.:4.750 3rd Qu.:551.0 3rd Qu.:10.95 3rd Qu.:14.050

Max. :570.0 Max. :8.600 Max. :788.0 Max. :17.40 Max. :16.300

X6

Min. : 0.000

1st Qu.: 4.000

Median : 8.000

Mean : 7.741

3rd Qu.:12.000

Max. :15.000

Regression as follows

> reg1 = lm(X1~. , data = model)

> summary(reg1)

Call:

lm(formula = X1 ~ ., data = allgreen)

Residuals:

Min 1Q Median 3Q Max

-26.338 -9.699 -4.496 4.040 41.139

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) -18.85941 30.15023 -0.626 0.538372

X2 16.20157 3.54444 4.571 0.000166 \*\*\*

X3 0.17464 0.05761 3.032 0.006347 \*\*

X4 11.52627 2.53210 4.552 0.000174 \*\*\*

X5 13.58031 1.77046 7.671 1.61e-07 \*\*\*

X6 -5.31097 1.70543 -3.114 0.005249 \*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 17.65 on 21 degrees of freedom

Multiple R-squared: 0.9932, Adjusted R-squared: 0.9916

F-statistic: 611.6 on 5 and 21 DF, p-value: < 2.2e-16

Objective

The objective of the exercise is to analyse how each of the variations in each of the variables X 2, X3, X4, X5, X5, X6 cause variations in X1 (Annual Net Sales )

The above outputs give us the following regression equation

X1=16.20157X2+0+17464X3+11.52627X4+13.58031X5-5.31097X6-18.85941

On increase of all factors by 1 unit,

* If X2 increases by 1 unit then X1will increase by $16,202.
* If X3 increases by 1 unit then X1 will increase by $1746.
* If X4 increases by 1 unit then X1 will increase by $11,562.
* If X5 increases by 1 unit then X1 will increase by $13,580.
* If X6 increases by 1 unit then X1 will decrease by $5,311.

**R Square interpretation**

**Multiple R-squared: 0.9932** implies that 99.32% of the variation in the annual net sales is explained by the independent variables

**Overall significance of the model**

* Regression has 5 degrees of freedsom (df). Total df is 26. Hence error or residual has 26-5=21 df.
* Probability (F>611.59) = 2.2e-16 which is much smaller than the significance level – Alpha of 5%
* Hence reject the null hypothesis (H0) that all Betas are zero.
* Conclusion is atleast 1 Beta is non zero and hence accept the alternative hypothesis (H1)
* Overall there is overwhelming evidence that regression model exists in the population meaning the linear model of annual net sales (X1) depending on number of square feet, inventory, amount spent on advertising, size of sales district and number of competing stores in the district is robust and statistically valid.
* The individual coefficients of all the independent variables ***except the intercept*** are highly significant as evidenced by the t stats which have extremely low P values which are significantly much less than alpha of 5%.
* The Beta coefficient of the intercept has a p value of 0.538 which is greater than the alpha of 5%.

**Confidence interval**

Confidence interval for each of the independent variables is given below. Interpretation is given below

**Conclusion**

* Probability that the population slope of number of square feet will lie between 8.831 and 23.573 is 95%
* Probability that the population slope of inventory will lie between 0.055 and 0.294 is 95%
* Probability that the population slope of advertising spend will lie between 6.260 and 16.792 is 95%
* Probability that the population slope of size of district will lie between 9.898 and 17.262 is 95%
* Probability that the population slope of competing stores in district will lie between -8.858 and -1.764 is 95%

The above model explains the importance of all the independent variable on X1.