STATISTICAL ANALYSIS FOR DATA SCIENCE (C7081) ASSESSMENT

JABEN PETER DARINYAN BAKO (213221)

2022-11-07

Github Link: [<https://github.com/Jabendah/C7081>]

**BACKGROUND**

Company sales is a dataset comprising two hundred (200) company (Observations) with four (4) different variables showing cost incured on advertisement. These variables include cost on Television, Radio, Newspaper, and Sales for 200 different companies. Often times companies produce products with sole aim to sale them to consumers but, the big task faced by these companies is getting these products or services to the knowledge of the general public and this is where various options of mode of advertisement come into play. Advertisement serves as a major tool in creating product awareness. It also conditions the mind of a potential consumer to take eventual purchase decision (Agbeja *etal,* 2015).

## Objectives

Since the dataset shows cost incurred on the different forms of advertisement, this work is aimed at the following objectives;

1. identifying the most efficient form of advertisement;
2. analyse the relationship between cost of advertisement and sales and;
3. predict Sales

## Hypothesis statement

The Null hypothesis of this study is;

H0: There’s no significant relationship between the cost of advertisement and Sales

**METHODS**

The company sales dataset is a dataset downloaded from kaggle [<https://www.kaggle.com/datasets/ishaanthareja007/company-sales>]. The methods or tools used to analyse this dataset outlined objectives include;

1. Exploratory Data Analysis
2. Regression Analysis (Simple Linear and Multiple Regression )
3. Decision Tree

The simple linear regression model is used to predict a quantitative response Y on the basis of a single predictor variable X. While, Multiple regression measures the response Y on two or more variable of X. Mathematically, it can be written as thus;

Y = B0+B1X Simple Linear Regression.

Y = B0+B1X1+B2X2+….+BnXn Multiple Regression.

where; Y = Sales X = Variables (Tv, radio and Newspaper) B0 = Intercept and; B1 = Regression Coefficients of various Variables

The Linear model helps in determining the relationship between variables and it’s predictive function is one big reason why it was chosen for this set of data.

The Decision tree is widely used because of it flowchart that shows a clear pathway to a decision and the predictive function too.

Results would also be displayed in charts and graphs/ plots (Histogram)

**RESULTS**

## Regression Analysis

library(readxl)  
company <- read\_excel("C:\\Users\\hp\\Desktop\\My R\\company.xlsx")  
head(company)

## # A tibble: 6 × 4  
## TV Radio Newspaper Sales  
## <dbl> <dbl> <dbl> <dbl>  
## 1 230. 37.8 69.2 22.1  
## 2 44.5 39.3 45.1 10.4  
## 3 17.2 45.9 69.3 12   
## 4 152. 41.3 58.5 16.5  
## 5 181. 10.8 58.4 17.9  
## 6 8.7 48.9 75 7.2

summary(company)

## TV Radio Newspaper Sales   
## Min. : 0.70 Min. : 0.000 Min. : 0.30 Min. : 1.60   
## 1st Qu.: 74.38 1st Qu.: 9.975 1st Qu.: 12.75 1st Qu.:11.00   
## Median :149.75 Median :22.900 Median : 25.75 Median :16.00   
## Mean :147.04 Mean :23.264 Mean : 30.55 Mean :15.13   
## 3rd Qu.:218.82 3rd Qu.:36.525 3rd Qu.: 45.10 3rd Qu.:19.05   
## Max. :296.40 Max. :49.600 Max. :114.00 Max. :27.00

According to the data summary, TV advertisement seem to cost more yet, more efficient.

Run a simple linear regression analysis to see the effect Television advertisement has on Sales.

lm.fit <- lm(Sales ~ TV, data = company)  
summary(lm.fit)

##   
## Call:  
## lm(formula = Sales ~ TV, data = company)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -6.4438 -1.4857 0.0218 1.5042 5.6932   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 6.974821 0.322553 21.62 <2e-16 \*\*\*  
## TV 0.055465 0.001896 29.26 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 2.296 on 198 degrees of freedom  
## Multiple R-squared: 0.8122, Adjusted R-squared: 0.8112   
## F-statistic: 856.2 on 1 and 198 DF, p-value: < 2.2e-16

cor(company$Sales, company$TV)

## [1] 0.9012079

From the above summary, we can say the Dist = 6.97 + 0.06\*TV, the R2 of 0.8 1 and adjusted R2 of 81% explains the variance in the data. Degree of Freedom = 1,198. It indicates also that, for every unit increase in TV advert, there’s a 60 unit increase in sales. The correlation 0.90 signifies a strong relationship between TV advert and Sales Ultimately for the companies.

#Decision: reject H0 and conclude that there’s is significant relationship between the cost of advertisement and Sales.

### Predict Sales as response from TV

predict(lm.fit, data.frame(TV = ( c(5, 10, 15) )),  
 interval = "confidence")

## fit lwr upr  
## 1 7.252145 6.632144 7.872147  
## 2 7.529469 6.925397 8.133542  
## 3 7.806793 7.218487 8.395099

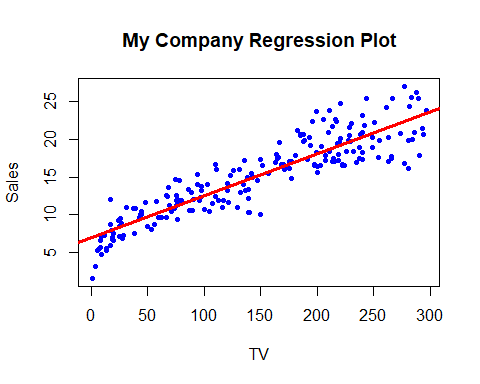
predict(lm.fit, data.frame(TV = ( c(5, 10, 15) )),  
 interval = "prediction")

## fit lwr upr  
## 1 7.252145 2.682638 11.82165  
## 2 7.529469 2.962096 12.09684  
## 3 7.806793 3.241478 12.37211

The 95% confidence interval associated with TV value of 10 is (6.93, 8.13) and the 95% prediction interval associated with TV value of 10 is (2.96, 12.1) for every 1 increase in Sales.

### plot

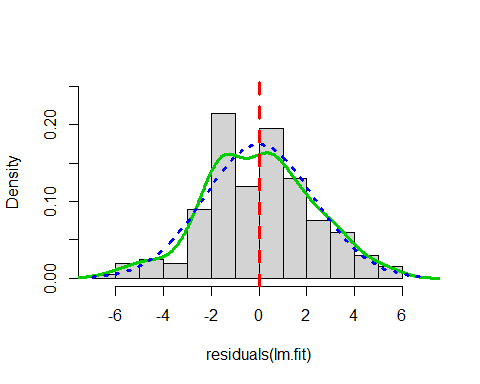
plot(y = company$Sales, x = company$TV,  
 ylab = "Sales", xlab = "TV",  
 main = "My Company Regression Plot",  
 pch = 20, col = "blue", cex = 1)  
abline(lm.fit, lwd = 3, col = "red",  
 text(x = 3,  
 y = 11,  
 labels = "y = 6.98 + (0.56)"), )



There is linearity in sales response to TV.

hist(residuals(lm.fit),  
 xlim = c(-7, 7), ylim = c(0,.25),  
 main = "",  
 prob = T)  
lines(  
 density(residuals(lm.fit)),  
 col = "green3", lty = 1, lwd = 3)  
x <- seq(-1, +1, by=0.02)  
curve(dnorm(x, mean = mean(residuals(lm.fit)),  
 sd = sd(residuals(lm.fit))),  
 add = T,  
 col = "blue", lty = 3, lwd = 3)  
abline(v = 0,   
 Freq = F,  
 col = "red", lty = 2, lwd = 3)

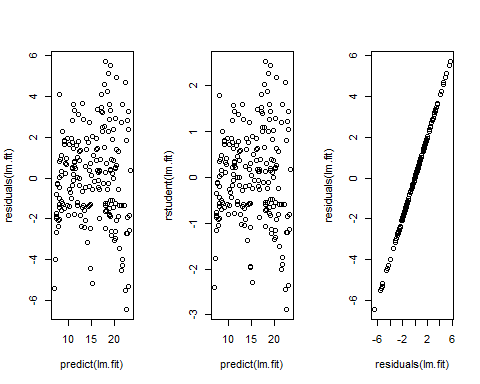
## Warning in int\_abline(a = a, b = b, h = h, v = v, untf = untf, ...): "Freq" is  
## not a graphical parameter



Near the mean, our residual density dropped and raised. The distribution is mostly symmetrical around the mean. Between -2.4 and -4.5 and also between 2 and 3, our residual density is lower than the expected under theoretical Gaussian.

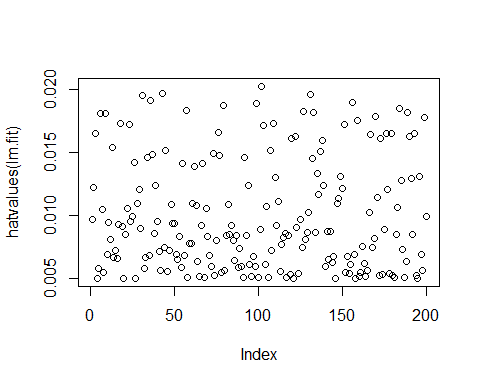
\*Diagnostic Plot

par(mfrow = c(1, 3))  
plot(predict(lm.fit), residuals(lm.fit))  
plot(predict(lm.fit), rstudent(lm.fit))  
plot(residuals(lm.fit), residuals(lm.fit))



par(mfrow = c(1, 1))

plot(hatvalues(lm.fit))



which.max(hatvalues(lm.fit))

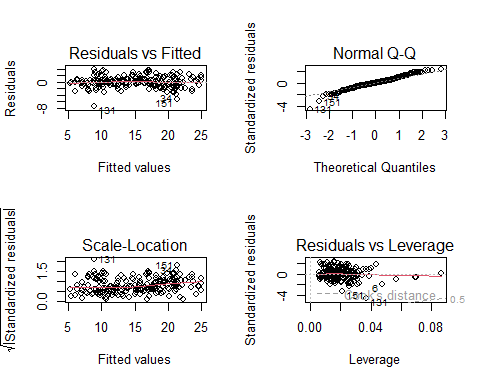
## 102   
## 102

## Multiple Regression

lm.fit1 <- lm(Sales ~ TV + Radio + Newspaper, data = company)  
summary(lm.fit1)

##   
## Call:  
## lm(formula = Sales ~ TV + Radio + Newspaper, data = company)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -7.3034 -0.8244 -0.0008 0.8976 3.7473   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 4.6251241 0.3075012 15.041 <2e-16 \*\*\*  
## TV 0.0544458 0.0013752 39.592 <2e-16 \*\*\*  
## Radio 0.1070012 0.0084896 12.604 <2e-16 \*\*\*  
## Newspaper 0.0003357 0.0057881 0.058 0.954   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 1.662 on 196 degrees of freedom  
## Multiple R-squared: 0.9026, Adjusted R-squared: 0.9011   
## F-statistic: 605.4 on 3 and 196 DF, p-value: < 2.2e-16

par(mfrow = c(2, 2))  
plot(lm.fit1)



### Prediction

intercept <- coef(summary(lm.fit1))["(Intercept)", "Estimate"]  
TV <- coef(summary(lm.fit1))["TV", "Estimate"]  
Radio <- coef(summary(lm.fit1))["Radio", "Estimate"]  
Newspaper <- coef(summary(lm.fit1))["Newspaper", "Estimate"]  
intercept + TV\*0.054 + Radio\*0.107 + Newspaper\*0.000

## [1] 4.639513

The model predicts that the companies would have a Sales of 4.639513.

### Summary

1. Based on the Coefficient estimate, It shows a 4.6 increase in cost to make sales
2. For every 1 increase in cost, the sales up by 0.05 for TV, 0.11 for Radio.
3. TV and Radio are statistically significant but the Newspaper isn’t according to the P-value
4. The t-value which measures our standard deviations estimates from 0 shows both TV & Radio to be relevant to sales.
5. The R-squared of 0.09026 (90%) of the observed response explained the variance in the variable.
6. A Sales of 4.639513 is predicted for the companies
7. Reject H0 and conclude that there’s is significant relationship between the cost of advertisement and Sales.

## Decision Tree

library(tree)  
attach(company)

## The following objects are masked \_by\_ .GlobalEnv:  
##   
## Newspaper, Radio, TV

High <- factor(ifelse(Sales <= 8, "No", "Yes"))  
company <- data.frame(company, High)  
tree.company <- tree(High ~ . -Sales, company)  
summary(tree.company)

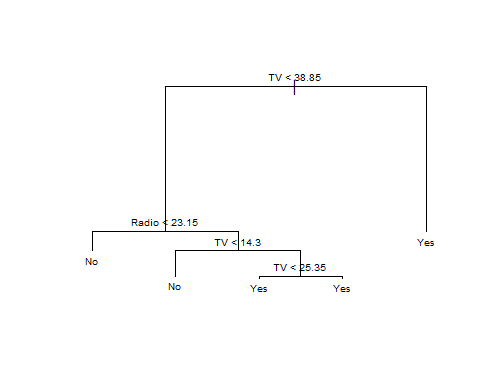
##   
## Classification tree:  
## tree(formula = High ~ . - Sales, data = company)  
## Variables actually used in tree construction:  
## [1] "TV" "Radio"  
## Number of terminal nodes: 5   
## Residual mean deviance: 0.02566 = 5.004 / 195   
## Misclassification error rate: 0.005 = 1 / 200

The tree summary has five (5) terminal nodes with a misclassification rate of 0.005 and a Residual mean = 0.03.

tree.company

## node), split, n, deviance, yval, (yprob)  
## \* denotes terminal node  
##   
## 1) root 200 134.400 Yes ( 0.1050 0.8950 )   
## 2) TV < 38.85 30 36.650 No ( 0.7000 0.3000 )   
## 4) Radio < 23.15 13 0.000 No ( 1.0000 0.0000 ) \*  
## 5) Radio > 23.15 17 23.510 Yes ( 0.4706 0.5294 )   
## 10) TV < 14.3 7 0.000 No ( 1.0000 0.0000 ) \*  
## 11) TV > 14.3 10 6.502 Yes ( 0.1000 0.9000 )   
## 22) TV < 25.35 5 5.004 Yes ( 0.2000 0.8000 ) \*  
## 23) TV > 25.35 5 0.000 Yes ( 0.0000 1.0000 ) \*  
## 3) TV > 38.85 170 0.000 Yes ( 0.0000 1.0000 ) \*

par(mfrow = c(1, 1))  
plot(tree.company)  
text(tree.company, pretty = 0, cex = .6)



From the above tree, the most important indicator of Sales is TV and stands as the parent nodes.

### Prediction

set.seed(2)  
train <- sample(1:nrow(company), 100)  
company.test <- company[-train, ]  
High.test <- High[-train]  
tree.company <- tree(High ~ . -Sales, company, subset = train)  
tree.pred <- predict(tree.company, company.test, type = "class")  
table(tree.pred, High.test)

## High.test  
## tree.pred No Yes  
## No 5 0  
## Yes 1 94

### % correct

(5 + 94) / 100

## [1] 0.99

#### Interpretation

The predication leads to 99% of the test.

Television is the root node of the tree and as represented on the tree table to be the most important and result yielding form of advertisement. This can be viewed practically as people spend more time watching TV as compared to listening to radio or reading newspaper.

**CONCLUSION**

The assessment used a dataset containing 200 observations and 4 variables with sole aim to identifying the most efficient form of advertisement; analysing the relationship between cost of advertisement and sales and; predict Sales. The best tool that best explains the observation is the regression model which showed that Television is the best mode of advertisement with positive statistical significance. we also reject the Null hypothesis and conclude that there’s a significant relationship between advertisement and sales. Also based on our analysis, using the prediction function in Regression and Decision Tree, we can comfortably conclude that in every increase in unit of cost in advertisement, there’s a complementary increase in sales.

**LITERATURE CITED**

Agbeja,O., Adelakun, O.J., & Akinyemi, D.(2015). *Analysis of the effect of advertising on Sales and Profitability of company.* International Journal of Novel Research in Marketing Management and Economics.

Ed Harris, (2022). *C7081 Statistical Analysis for Data Science* [<http://c7081-2022.github.io/website>]

James *et al.* (2021). *Introduction to statistical learning.* 2nd edition

Ott, R.L, Michael.L (2001), *An introduction to statistical methods and data analysis.* 5th ed. Texas A&M University.