DATA1204 - Assignment6

Chinedu Onyeka

12/18/2021

**Load Libraries**

# load libraries  
library(tidyverse)

**Read the file**

# read the file  
url <- "https://raw.githubusercontent.com/chinedu2301/DC\_Analytics/main/MultiRegDataset.csv"  
hughes\_dataset <- read\_csv(url)

**Check the head**

# check the head  
head(hughes\_dataset)

## # A tibble: 6 x 7  
## age sex bmi children smoker region expenses  
## <dbl> <chr> <dbl> <dbl> <chr> <chr> <dbl>  
## 1 19 female 27.9 0 yes southwest 16885.  
## 2 18 male 33.8 1 no southeast 1726.  
## 3 28 male 33 3 no southeast 4449.  
## 4 33 male 22.7 0 no northwest 21984.  
## 5 32 male 28.9 0 no northwest 3867.  
## 6 31 female 25.7 0 no southeast 3757.

**Check the structure**

# Check the datatype and structure of the dataset  
glimpse(hughes\_dataset)

## Rows: 1,338  
## Columns: 7  
## $ age <dbl> 19, 18, 28, 33, 32, 31, 46, 37, 37, 60, 25, 62, 23, 56, 27, 1~  
## $ sex <chr> "female", "male", "male", "male", "male", "female", "female",~  
## $ bmi <dbl> 27.9, 33.8, 33.0, 22.7, 28.9, 25.7, 33.4, 27.7, 29.8, 25.8, 2~  
## $ children <dbl> 0, 1, 3, 0, 0, 0, 1, 3, 2, 0, 0, 0, 0, 0, 0, 1, 1, 0, 0, 0, 0~  
## $ smoker <chr> "yes", "no", "no", "no", "no", "no", "no", "no", "no", "no", ~  
## $ region <chr> "southwest", "southeast", "southeast", "northwest", "northwes~  
## $ expenses <dbl> 16884.92, 1725.55, 4449.46, 21984.47, 3866.86, 3756.62, 8240.~

*There are 7 columns (variables) and 1,338 rows (observations) in the dataset.*

**Descriptive Statistics**

summary(hughes\_dataset)

## age sex bmi children   
## Min. :18.00 Length:1338 Min. :16.00 Min. :0.000   
## 1st Qu.:27.00 Class :character 1st Qu.:26.30 1st Qu.:0.000   
## Median :39.00 Mode :character Median :30.40 Median :1.000   
## Mean :39.21 Mean :30.67 Mean :1.095   
## 3rd Qu.:51.00 3rd Qu.:34.70 3rd Qu.:2.000   
## Max. :64.00 Max. :53.10 Max. :5.000   
## smoker region expenses   
## Length:1338 Length:1338 Min. : 1122   
## Class :character Class :character 1st Qu.: 4740   
## Mode :character Mode :character Median : 9382   
## Mean :13270   
## 3rd Qu.:16640   
## Max. :63770

*Age*

sd(hughes\_dataset$age)

## [1] 14.04996

The minimum age is 18 years while the maximum age is 64. The mean age is about 39 years and the standard deviation of age is about 14 years.

*bmi*

sd(hughes\_dataset$bmi)

## [1] 6.098382

The minimum bmi is 16 while the maximum bmi is 53.10. The mean bmi is about 30.67 and the standard deviation of bmi is about 6.1.

*children*

sd(hughes\_dataset$children)

## [1] 1.205493

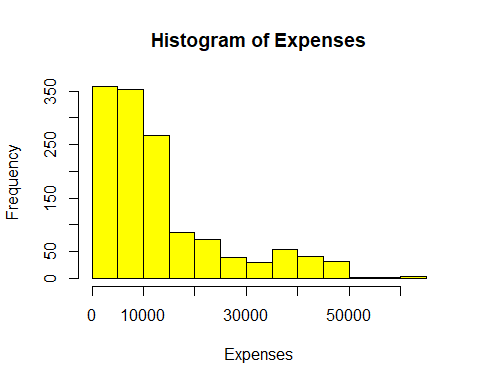
The minimum number of children is 0 while the maximum number of children is 5. The mean number of children is 1 and the standard deviation of number of children is 1.

*sex*  
There are 676 Males and 662 females in the dataset.

*smoker*  
There are 274 who are smokers and 1064 who are not in the dataset.

**Histogram - expenses**

# histogram  
hist(hughes\_dataset$expenses, main = "Histogram of Expenses", xlab = "Expenses", col = "yellow")



**T-test**

*Hypothesis Testing*  
Null hypothesis   
Alternative hypothesis

t-test

# t-test  
t.test(hughes\_dataset$expenses, mu = 10000)

##   
## One Sample t-test  
##   
## data: hughes\_dataset$expenses  
## t = 9.8784, df = 1337, p-value < 2.2e-16  
## alternative hypothesis: true mean is not equal to 10000  
## 95 percent confidence interval:  
## 12620.95 13919.89  
## sample estimates:  
## mean of x   
## 13270.42

*The p-value is less than 0.05, so we reject the null hypothesis that the mean of expenses is equal to 10,000*

**Simple Linear Regression**  
y = Expenses, x = Smoker  
Expenses = Intercept + slope \* Smoker

Ho: β=0, co-efficient β of the predictor is zero and not statistically significant  
Ha: β ≠0, co-efficient β of the predictor is not equal to zero and is statistically significant

*Train the model*

# train the model   
simple\_model <- lm(expenses ~ smoker, data = hughes\_dataset)   
  
# get the summary  
summary(simple\_model)

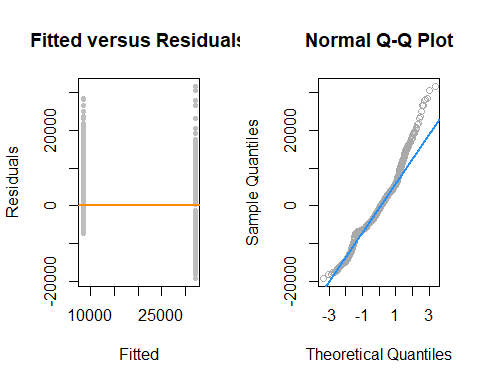
##   
## Call:  
## lm(formula = expenses ~ smoker, data = hughes\_dataset)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -19221 -5042 -919 3705 31720   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 8434.3 229.0 36.83 <2e-16 \*\*\*  
## smokeryes 23616.0 506.1 46.66 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 7470 on 1336 degrees of freedom  
## Multiple R-squared: 0.6198, Adjusted R-squared: 0.6195   
## F-statistic: 2178 on 1 and 1336 DF, p-value: < 2.2e-16

The “smoker-yes”is significant, and the R-squared is 61.98%  
The simple linear regression model is

*Since the p-value is not zero, we reject the null hypothesis and accept the alternative hypothesis that the co-efficient of the predictor variable is not zero*

*Evaluate the model*

par(mfrow = c(1, 2))  
# plot the fitted vs residual  
plot(fitted(simple\_model), resid(simple\_model), col = "grey", pch = 20,  
 xlab = "Fitted", ylab = "Residuals", main = "Fitted versus Residuals")  
abline(h = 0, col = "darkorange", lwd = 2)  
  
# Q-Q plot  
qqnorm(resid(simple\_model), main = "Normal Q-Q Plot", col = "darkgrey")  
qqline(resid(simple\_model), col = "dodgerblue", lwd = 2)



**Multiple Linear Regression**  
 where i = 1,2, …, n

*Hypothesis Statement*  
 at least one

*Train the model*

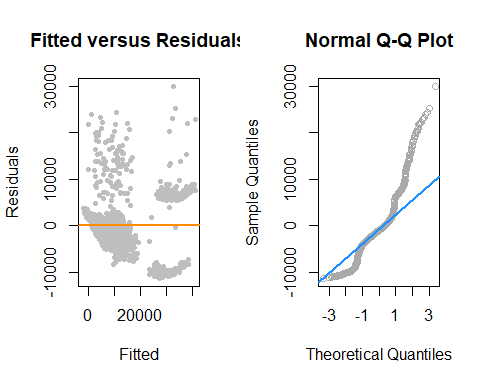
# train the model   
multiple\_model <- lm(expenses ~ ., data = hughes\_dataset)   
  
# get the summary  
summary(multiple\_model)

##   
## Call:  
## lm(formula = expenses ~ ., data = hughes\_dataset)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -11302.7 -2850.9 -979.6 1383.9 29981.7   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -11941.6 987.8 -12.089 < 2e-16 \*\*\*  
## age 256.8 11.9 21.586 < 2e-16 \*\*\*  
## sexmale -131.3 332.9 -0.395 0.693255   
## bmi 339.3 28.6 11.864 < 2e-16 \*\*\*  
## children 475.7 137.8 3.452 0.000574 \*\*\*  
## smokeryes 23847.5 413.1 57.723 < 2e-16 \*\*\*  
## regionnorthwest -352.8 476.3 -0.741 0.458976   
## regionsoutheast -1035.6 478.7 -2.163 0.030685 \*   
## regionsouthwest -959.3 477.9 -2.007 0.044921 \*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 6062 on 1329 degrees of freedom  
## Multiple R-squared: 0.7509, Adjusted R-squared: 0.7494   
## F-statistic: 500.9 on 8 and 1329 DF, p-value: < 2.2e-16

*The R-squared is 75.09% for the multi-linear regression*

*Evaluate the model*

par(mfrow = c(1, 2))  
# plot the fitted vs residual  
plot(fitted(multiple\_model), resid(multiple\_model), col = "grey", pch = 20,  
 xlab = "Fitted", ylab = "Residuals", main = "Fitted versus Residuals")  
abline(h = 0, col = "darkorange", lwd = 2)  
  
# Q-Q plot  
qqnorm(resid(multiple\_model), main = "Normal Q-Q Plot", col = "darkgrey")  
qqline(resid(multiple\_model), col = "dodgerblue", lwd = 2)



**Conclusion**  
Smoking has a positive effect on the expenses for the simple linear regression. Also, other input variables helped in increasing the R-squared values for predicting the expenses.