**Project Name: INSURANCE ANALYSIS**

**(QMB 6304 Regression Project)**

**SUBJECT: ANALYTIC METHODS FOR BUSINESS**

**Team:**

1)AKHIL MADUPU

U number: U01264561

2)MANOHAR REDDY PASHAM

U number: U51001133

3)SOWMYA KOMIRELLY

U number: U28160805

Thanks, professor, for this wonderful opportunity to present our project work.

**PROJECT REPORT**

**Objective:**

We took the data set from an insurance company, the data set was available online and we made analysis of how the charges vary according to age, bmi and gender. We used different methods to predict the charge of the insurance.

**DATA SET REFERENCE**

We have taken the data from an insurance company online source. This data pertains the insurance costs of all individual based on genders,age,bmi. They have published their data online for reporting or to have a record. You can find the dataset in this link

“https://www.kaggle.com/code/annetxu/health-insurance-cost-predicition/data?select=insurance.csv”

**DESCRIPTION OF VARIABLES**

Insurance: It is used to store and load the data set

Insurance$y: It is used to copy the ‘charges’ column of the data set.

Insurance$x1: It is used to copy the ‘age’ column of the data set.

Insurance$x2: It is used to copy the ‘bmi’ column of the data set.

Insurance$x3: It is used to copy the ‘sex’ column of the data set.

data\_insurance: It is used to load a subset of insurance.

df: It is used to load data frame of size 100 from the data set ‘data\_insurance’.

lr1: this variable is used for doing linear regression of y~X1 from data set ‘df’.

lr2: this variable is used for doing linear regression of y~X2 from data set ‘df’.

lr3: this variable is used for doing linear regression of y~X3 from data set ‘df’.

mlr1: this variable is used for doing multiple regression of y~X1+X2 from data set ‘df’.

mlr2: this variable is used for doing multiple regression of y~X1+X3 from data set ‘df’.

mlr3: this variable is used for doing multiple regression of y~X2+X3 from data set ‘df’.

mr1: it is used for doing main effect of multiple regression model y~X1+X2+X3 from data set ‘data\_ insurance’.

Interact\_lm: this variable is used for loading interactions from data set ‘df’.

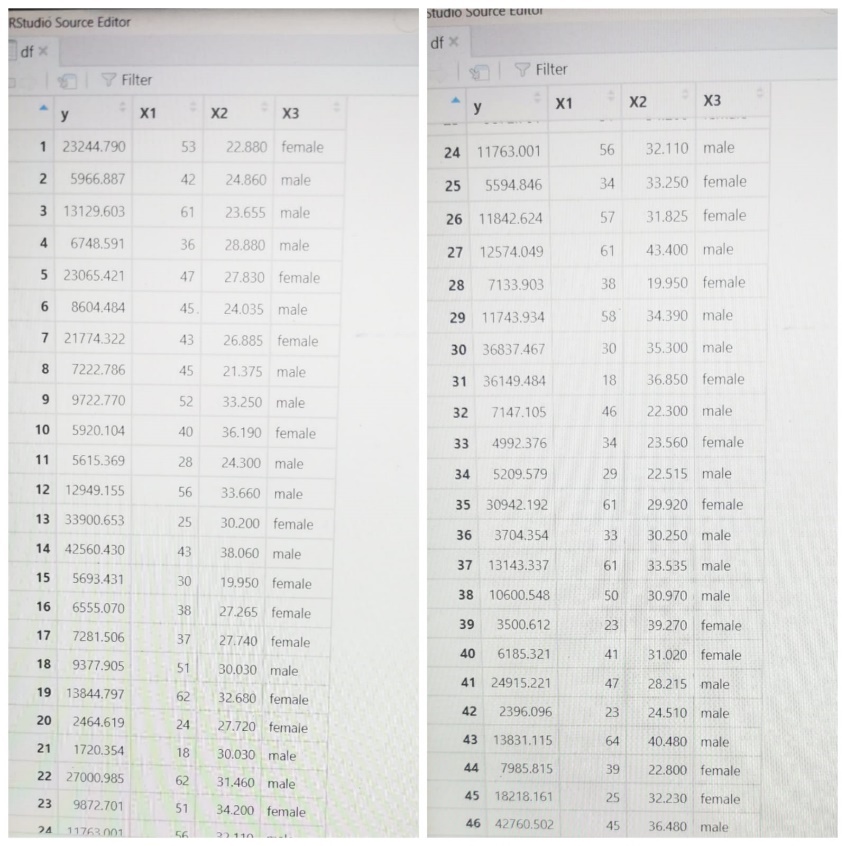
Slm1: this variable is used for doing simple regression 1 on data set ‘df’.

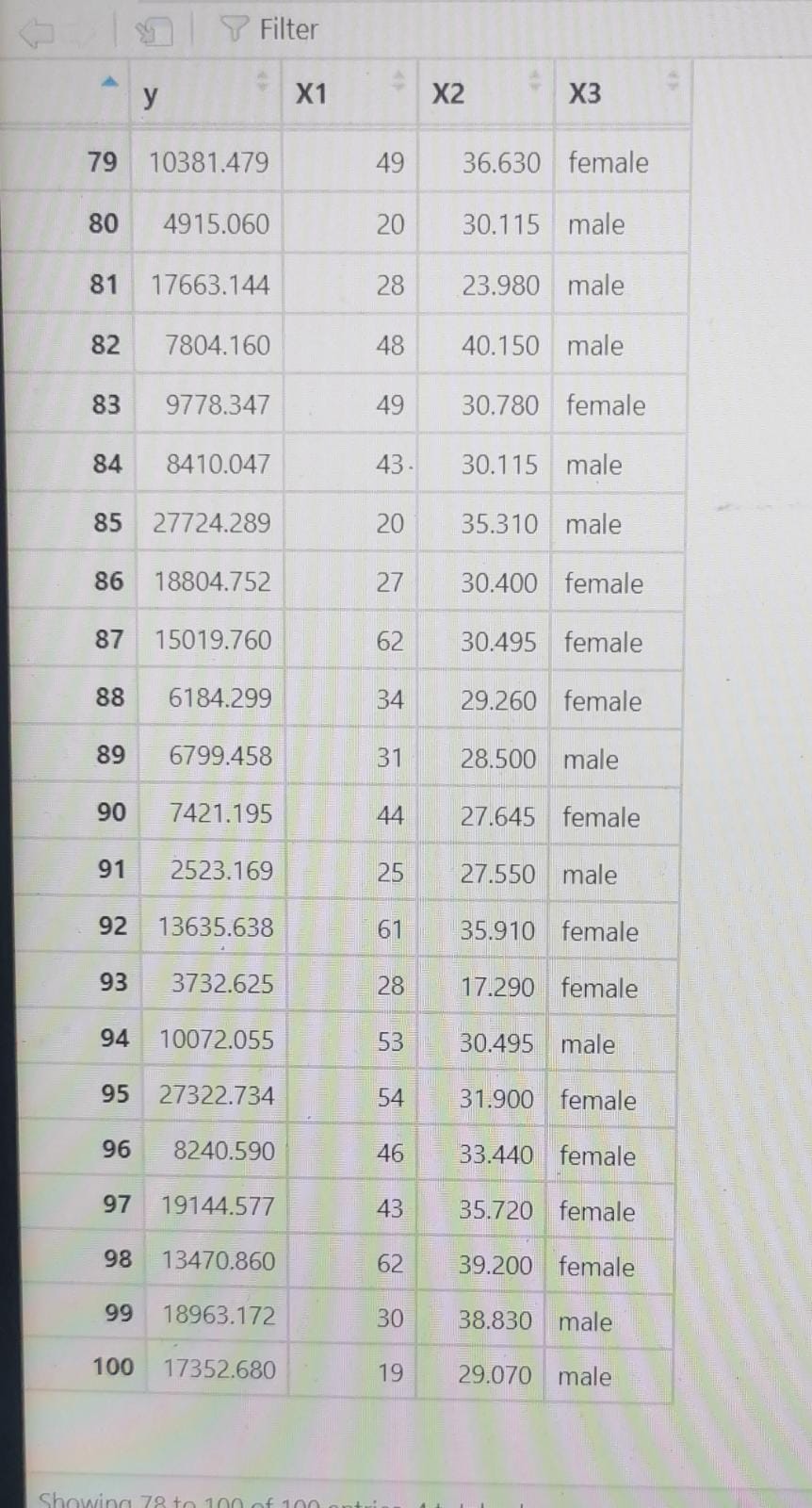
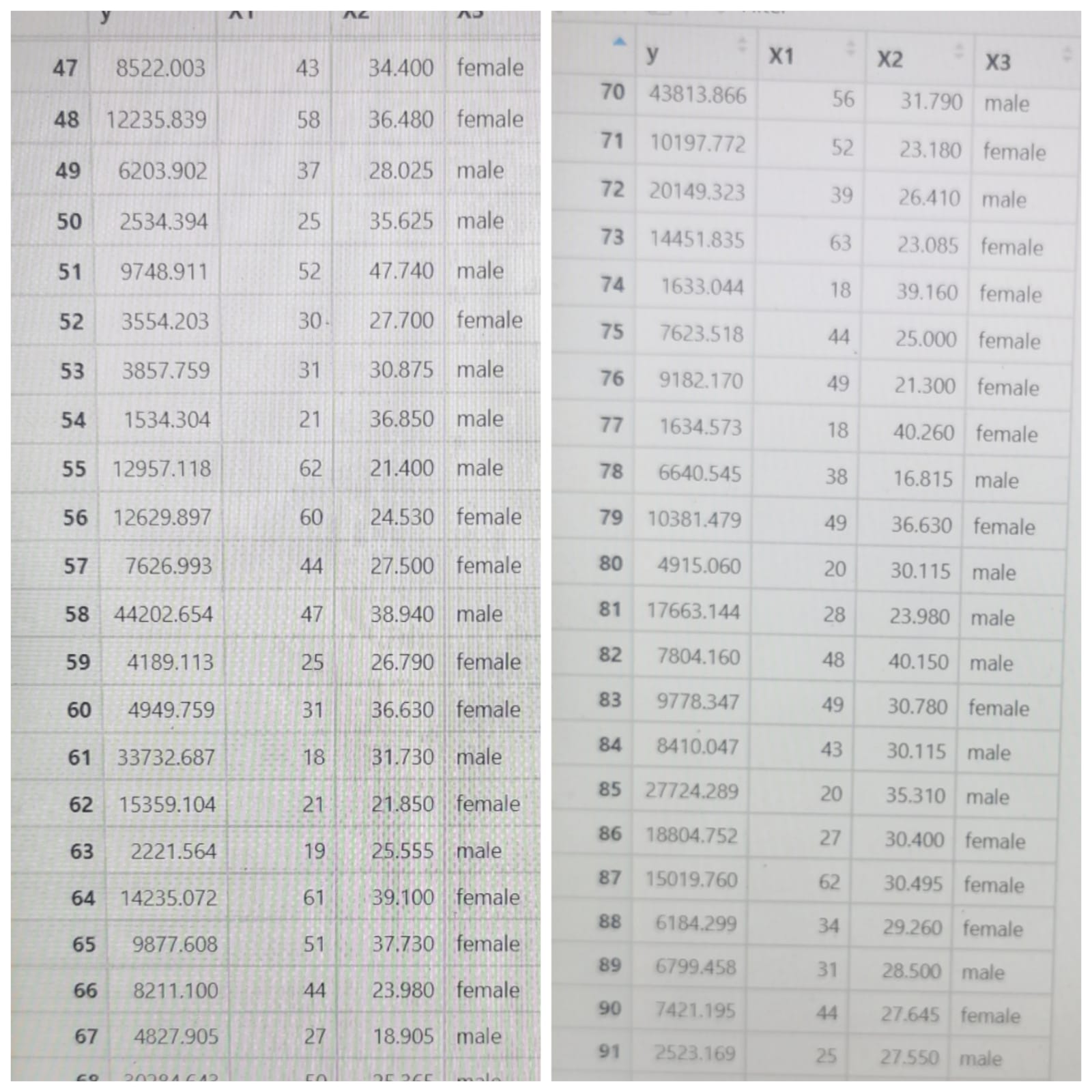
Slm2: this variable is used for doing simple regression 2 ‘df’.

pred: this variable is used for doing predictions on Slm1 and df with confidence intervals.

r: we have used this variable for doing correlation on data set ‘df’.

**DATA SET of 100 sample observations**





**###PREPROCESSING**

**CODE:**

# Akhil Madupu

# Sowmya Komirelli

# Manohar reddy

rm(list=ls())

library(rio)

library(car)

library(moments)

insurance<- read.csv("Downloads/insurance.csv")

insurance

**ANALYSIS:**

We have loaded libraries rio,car,moments. We have also imported the data set into our Rstudio.

**CODE:**

insurance$y<-insurance$charges

insurance$X1<-insurance$age

insurance$X2<-insurance$bmi

insurance$X3<-insurance$sex

df<-insurance[,c(8:11)]

set.seed(1133)

df<-df[sample(nrow(df),100),]

**ANALYSIS:**

We are copying charges column to a new variable y , age column to X1, bmi column to X2 and sex column to X3.

**#ANALYSING THE DATA SET USING REGRESSION MODELS**

**CODE:**

#simple Linear Regression

lr1<-lm(y~X1,data = df)

summary(lr1)

**OUTPUT:**

> #simple Linear Regression

> lr1<-lm(y~X1,data = df)

> summary(lr1)

Call:

lm(formula = y ~ X1, data = df)

Residuals:

Min 1Q Median 3Q Max

-8059 -6671 -5939 5440 47829

Coefficients:

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

(Intercept) 3165.9 937.1 3.378 0.000751 \*\*\*

X1 257.7 22.5 11.453 < 2e-16 \*\*\*

---

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

Residual standard error: 11560 on 1336 degrees of freedom

Multiple R-squared: 0.08941, Adjusted R-squared: 0.08872

F-statistic: 131.2 on 1 and 1336 DF, p-value: < 2.2e-16

**ANALYSIS:**

* P value is less than 5%, hence we reject null hypothesis.
* Adjusted r square value is 0.08872. As it is quite low it does not have any significant correlation between the independent and dependent variables.
* The value of F stat is 131.2.
* Alternate hypothesis is accepted.

**CODE:**

lr2<-lm(y~X2,data = df)

summary(lr2)

> lr2<-lm(y~X2,data = df)

> summary(lr2)

Call:

lm(formula = y ~ X2, data = df)

Residuals:

Min 1Q Median 3Q Max

-20956 -8118 -3757 4722 49442

Coefficients:

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

(Intercept) 1192.94 1664.80 0.717 0.474

X2 393.87 53.25 7.397 2.46e-13 \*\*\*

---

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

Residual standard error: 11870 on 1336 degrees of freedom

Multiple R-squared: 0.03934, Adjusted R-squared: 0.03862

F-statistic: 54.71 on 1 and 1336 DF, p-value: 2.459e-13

**ANALYSIS:**

* P value is less than 5%, hence we reject null hypothesis.
* Adjusted r square value is 0.03862. As it is quite low it does not have any significant correlation between the independent and dependent variables.
* The value of F stat is 54.71.
* Alternate hypothesis is accepted.

**CODE:**

lr3<-lm(y~X3,data = df)

summary(lr3)

**OUTPUT:**

> lr3<-lm(y~X3,data = df)

> summary(lr3)

Call:

lm(formula = y ~ X3, data = df)

Residuals:

Min 1Q Median 3Q Max

-12835 -8435 -3980 3476 51201

Coefficients:

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

(Intercept) 12569.6 470.1 26.740 <2e-16 \*\*\*

X3male 1387.2 661.3 2.098 0.0361 \*

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

Residual standard error: 12090 on 1336 degrees of freedom

Multiple R -squared: 0.003282, Adjusted R-squared: 0.002536

F-statistic: 5.12 on 1 and 1336 DF, p-value: 0.03613

**ANALYSIS:**

* P value is less than 5%, hence we reject null hypothesis.
* Adjusted r square value is 0.002536. As it is quite low it does not have any significant correlation between the independent and dependent variables.
* The value of F stat is 5.12
* Alternate hypothesis is accepted.

**SUMMARY OF SIMPLE LINEAR REGRESSION:**

* We have taken the y which has the values of charges as dependent variable and X1,X2,X3 as independent variables.
* P values are less than 5% for X1,X2,X3. So, they have a statistical significance.The coefficients of X1,X2,X3 are 0.08872,0.03862,0.002536. So it is not the best fit.
* It is having very low significance due to low coefficient values.

**CODE:**

#Multiple Linear Regression

mlr1<-lm(y~X1+X2,data = df)

summary(mlr1)

**OUTPUT:**

> #Multiple Linear Regression

> mlr1<-lm(y~X1+X2,data = df)

> summary(mlr1)

Call:

lm(formula = y ~ X1 + X2, data = df)

Residuals:

Min 1Q Median 3Q Max

-14457 -7045 -5136 7211 48022

Coefficients:

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

(Intercept) -6424.80 1744.09 -3.684 0.000239 \*\*\*

X1 241.93 22.30 10.850 < 2e-16 \*\*\*

X2 332.97 51.37 6.481 1.28e-10 \*\*\*

---

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

Residual standard error: 11390 on 1335 degrees of freedom

Multiple R-squared: 0.1172, Adjusted R-squared: 0.1159

F-statistic: 88.6 on 2 and 1335 DF, p-value: < 2.2e-16

**ANALYSIS**

**CODE**

mlr2<-lm(y~X1+X3,data = df)

summary(mlr2)

**OUTPUT**

> mlr2<-lm(y~X1+X3,data = df)

> summary(mlr2)

Call:

lm(formula = y ~ X1 + X3, data = df)

Residuals:

Min 1Q Median 3Q Max

-8821 -6947 -5511 5443 48203

Coefficients:

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

(Intercept) 2343.62 994.35 2.357 0.0186 \*

X1 258.87 22.47 11.523 <2e-16 \*\*\*

X3male 1538.83 631.08 2.438 0.0149 \*

---

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

Residual standard error: 11540 on 1335 degrees of freedom

Multiple R-squared: 0.09344, Adjusted R-squared: 0.09209

F-statistic: 68.8 on 2 and 1335 DF, p-value: < 2.2e-16

**ANALYSIS**

* P value is less than 5%, hence we reject null hypothesis.
* Adjusted r square value is 0.09209. As it is quite low it does not have any significant correlation between the independent and dependent variables.
* The value of F stat is 68.8
* Alternate hypothesis is accepted.
* Additional value is not affecting much on the model

**CODE**

mlr3<-lm(y~X2+X3,data = df)

summary(mlr3)

**OUTPUT**

> mlr3<-lm(y~X2+X3,data = df)

> summary(mlr3)

Call:

lm(formula = y ~ X2 + X3, data = df)

Residuals:

Min 1Q Median 3Q Max

-21434 -8172 -3827 4658 48863

Coefficients:

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

(Intercept) 739.43 1682.45 0.439 0.6604

X2 389.43 53.26 7.311 4.54e-13 \*\*\*

X3male 1166.99 649.42 1.797 0.0726 .

---

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

Residual standard error: 11860 on 1335 degrees of freedom

Multiple R-squared: 0.04166, Adjusted R-squared: 0.04022

F-statistic: 29.01 on 2 and 1335 DF, p-value: 4.626e-13

**ANALYSIS**

* P value is less than 5%, hence we reject null hypothesis.
* Adjusted r square value is 0.04022. As it is quite low it does not have any significant correlation between the independent and dependent variables.
* The value of F stat is 29.01
* Alternate hypothesis is accepted.
* Additional value is not affecting much on the model

**CODE:**

# one full main effects multiple regression model

mr1<-lm(y~X1+X2+X3,data = df)

summary(mr1)

**OUTPUT:**

> mr1<-lm(y~X1+X2+X3,data = df)

> summary(mr1)

Call:

lm(formula = y ~ X1 + X2 + X3, data = df)

Residuals:

Min 1Q Median 3Q Max

-14974 -7073 -5072 6953 47348

Coefficients:

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

(Intercept) -6986.82 1761.04 -3.967 7.65e-05 \*\*\*

X1 243.19 22.28 10.917 < 2e-16 \*\*\*

X2 327.54 51.37 6.377 2.49e-10 \*\*\*

X3male 1344.46 622.66 2.159 0.031 \*

---

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

Residual standard error: 11370 on 1334 degrees of freedom

Multiple R-squared: 0.1203, Adjusted R-squared: 0.1183

F-statistic: 60.78 on 3 and 1334 DF, p-value: < 2.2e-16

**ANALYSIS:**

* P value is less than 5%, hence we reject null hypothesis.
* Adjusted r square value is 0.1183. It comparatively has a better r squared value.
* The value of F stat is 60.78
* Alternate hypothesis is accepted.

**SUMMARY FOR MULTIPLE LINEAR REGRESSION**

* We have taken the y which has the values of charges as dependent variable and X1,X2,X3 as independent variables.
* P values are less than 5% for X1,X2,X3. So, they have a statistical significance.The coefficients of X1,X2,X3 are 0.09209,0.04022,0.1183, .
* From all the values multiple linear regression is the best fit. This is based on r square error and residual standard error.

**CODE:**

#interactions

interact\_lm<-lm(y~X1+X2+X1\*X2,data = df)

summary(interact\_lm)

**OUTPUT:**

> interact\_lm<-lm(y~X1+X2+X1\*X2,data = df)

> summary(interact\_lm)

Call:

lm(formula = y ~ X1 + X2 + X1 \* X2, data = df)

Residuals:

Min 1Q Median 3Q Max

-15534 -7030 -5118 7234 47992

Coefficients:

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

(Intercept) -9162.555 4633.918 -1.977 0.04822 \*

X1 313.657 114.663 2.735 0.00631 \*\*

X2 422.247 149.134 2.831 0.00471 \*\*

X1:X2 -2.321 3.640 -0.638 0.52376

---

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

Residual standard error: 11390 on 1334 degrees of freedom

Multiple R-squared: 0.1175, Adjusted R-squared: 0.1155

F-statistic: 59.18 on 3 and 1334 DF, p-value: < 2.2e-16

**ANALYSIS:**

* We find the charge of insurance according to age ,bmi and gender.
* From the results of interaction, we can clearly say that age has more impact on charges of insurance.
* F static is 59.18 , R squared value is 0.1155 and p value is 2.2e-16. The coefficient of determination is 0.1175 which mean 11.75% of data in regression model is not fit. So the age bmi and gender does not affect much on insurance charge but age has a significant effect.

**CODE:**

#Two Simple regression models

Slm1<-lm(y~X1+X1^2,data = df)

summary(Slm1)

**OUTPUT:**

> Slm1<-lm(y~X1+X1^2,data = df)

> summary(Slm1)

Call:

lm(formula = y ~ X1 + X1^2, data = df)

Residuals:

Min 1Q Median 3Q Max

-8059 -6671 -5939 5440 47829

Coefficients:

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

(Intercept) 3165.9 937.1 3.378 0.000751 \*\*\*

X1 257.7 22.5 11.453 < 2e-16 \*\*\*

---

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

Residual standard error: 11560 on 1336 degrees of freedom

Multiple R-squared: 0.08941, Adjusted R-squared: 0.08872

F-statistic: 131.2 on 1 and 1336 DF, p-value: < 2.2e-16

**ANALYSIS**

* P value is less than 5%, hence we reject null hypothesis.
* Adjusted r square value is 0.08872.
* The value of F stat is 131.2.
* Alternate hypothesis is accepted.
* We can say that charge can be calculated by the age of the person. But using bmi is also significant.
* As r square value is low it does not have a significance on dependent and independent variables.
* Residual standard error is 11560 on 1336 degrees of freedom. It is big number. So, the value is not significant.

**CODE**

Slm2<-lm(y~X2+X2^2,data = df)

summary(Slm2)

**OUTPUT**

> Slm2<-lm(y~X2+X2^2,data = df)

> summary(Slm2)

Call:

lm(formula = y ~ X2 + X2^2, data = df)

Residuals:

Min 1Q Median 3Q Max

-20956 -8118 -3757 4722 49442

Coefficients:

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

(Intercept) 1192.94 1664.80 0.717 0.474

X2 393.87 53.25 7.397 2.46e-13 \*\*\*

---

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

Residual standard error: 11870 on 1336 degrees of freedom

Multiple R-squared: 0.03934, Adjusted R-squared: 0.03862

F-statistic: 54.71 on 1 and 1336 DF, p-value: 2.459e-13

**ANALYSIS:**

* P value is less than 5%, hence we reject null hypothesis.
* Adjusted r square value is 0.03862.
* The value of F stat is 54.71.
* Alternate hypothesis is accepted.
* We can say that charge can be calculated by the age of the person. But using gender is not significant.
* As r square value is low it does not have a significance on dependent and independent variables.
* Residual standard error is 11870 on 1336 degrees of freedom. It is big number. So, the value is not significant.

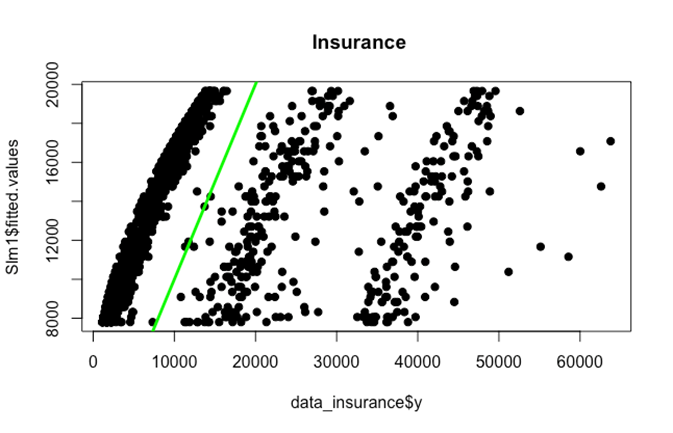
**CODE:**

##linearity

plot(df$y, Slm1$fitted.values,pch=19, main = ' Insurance')

abline(0,1,lwd=3, col='green')

**OUTPUT:**

****

**CODE:**

##normality

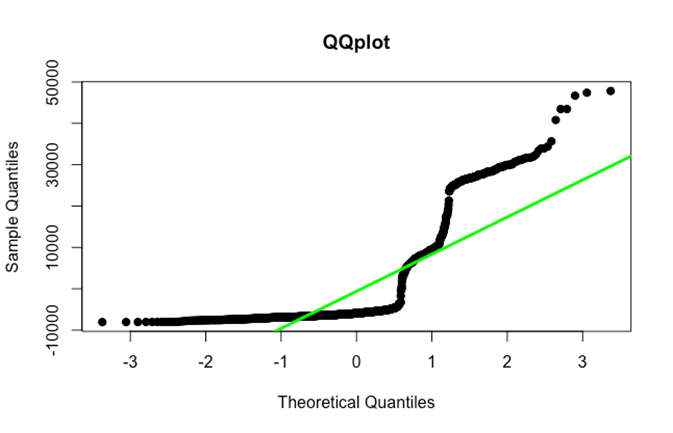
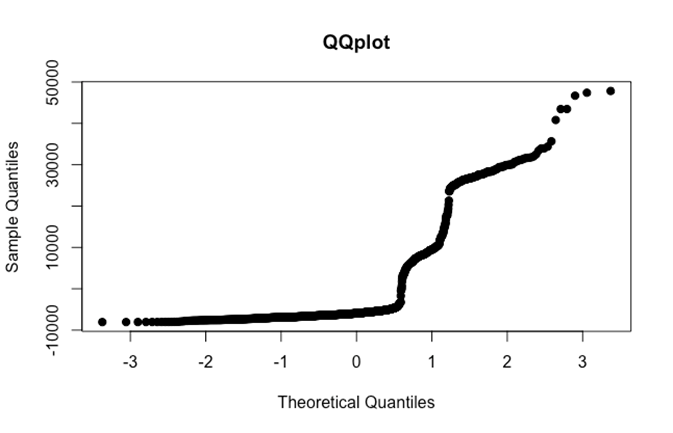
qqnorm(Slm1$residuals,pch=19,main="QQplot")

qqline(Slm1$residuals,lwd=3,col="green")

hist(Slm1$residuals,col="green", main = "Residual values", probability = TRUE)

curve(dnorm(x,0,sd(Slm1$residuals)), from = min(Slm1$residuals),to= max(Slm1$residuals), lwd=3, col="green", add=TRUE)

**OUTPUT:**

****

****

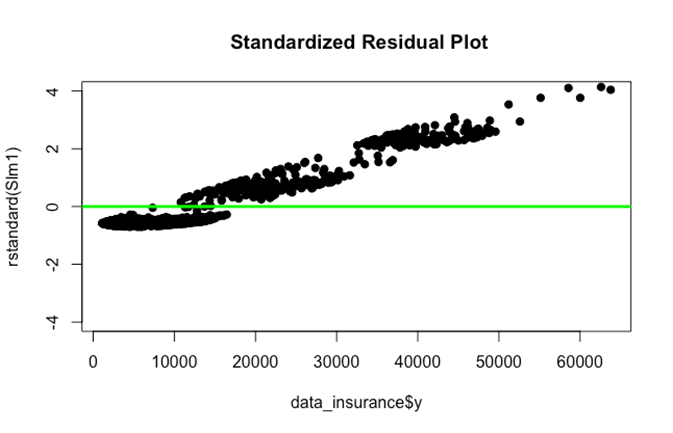
**CODE:**

#EQUALITY

plot(df$y,rstandard(Slm1), pch=19, main=" Standardized Residual Plot", ylim=c(-4,4))

abline(0,0,col="green", lwd=3)

**OUTPUT:**

****

**ANALYSIS:**

In linearity we can observe that we have many outliers. In qqplot we can see that data points are normally distributed with few outliers above the line. In equality plot we can clearly see that we have some outliers and extreme values.

From the plots hence we can say there is a linear relationship between dependent and independent variables and also it following normality. But is not following equality

**CODE:**

pred=predict(Slm1,df,interval = 'confidence')

pred

**OUTPUT:**

pred=predict(Slm1,df,interval = 'confidence')

> pred

fit lwr upr

1 8062.615 6976.301 9148.929

**ANALYSIS:**

**CODE:**

library(corrplot)

r=cor(df[,c(1:3)])

r

corrplot(r)

**OUTPUT:**

> library(corrplot)

> r=cor(df[,c(1:3)])

> r

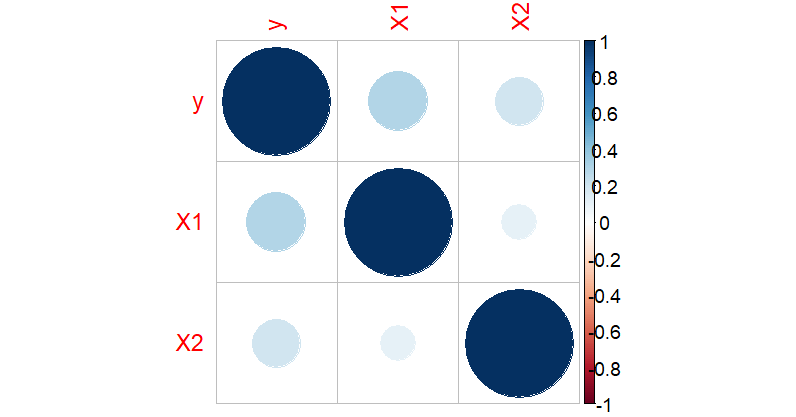
y X1 X2

y 1.0000000 0.2990082 0.1983410

X1 0.2990082 1.0000000 0.1092719

X2 0.1983410 0.1092719 1.0000000

> corrplot(r)



**ANALYSIS:**

* We used the predict function to predict the best fit of the test data against multiple linear regression.
* The upper bound is 9148.929
* The lower bound is 6976.301
* So, we can say that it has a range of 6976-9148 and the best fit is 8062.625