Project 9

Project Name : **Healthcare cost analysis**

|  |  |
| --- | --- |
| **Attribute** | **Description** |
| Age | Age of the patient discharged |
| Female | A binary variable that indicates if the patient is female |
| Los | Length of stay in days |
| Race | Race of the patient (specified numerically) |
| Totchg | Hospital discharge costs |
| Aprdrg | All Patient Refined Diagnosis Related Groups |

**Dataset File name :** 1555054100\_hospitalcosts.csv

**Business Problem :** To Predict the Hospital Discharge Cost **Business Solution :** To build the Multiple Linear Regression

# STEP 1 : UPLOAD THE DATASET IN R STUDIO AND IMPORT THE DATA

|  |
| --- |
| > getwd()  [1] "/home/kuntalc1\_gmail" |
|  |
| |  | | --- | | > | |

data1 <- read.csv("/home/kuntalc1\_gmail/1555054100\_hospitalcosts.csv")

View(data1)

# STEP 2 : CONDUCT DATA EXPLORATION

> str(data1)

'data.frame': 500 obs. of 6 variables:

$ AGE : int 17 17 17 17 17 17 17 16 16 17 ...

$ FEMALE: int 1 0 1 1 1 0 1 1 1 1 ...

$ LOS : int 2 2 7 1 1 0 4 2 1 2 ...

$ RACE : int 1 1 1 1 1 1 1 1 1 1 ...

$ TOTCHG: int 2660 1689 20060 736 1194 3305 2205 1167 532 1363 ...

$ APRDRG: int 560 753 930 758 754 347 754 754 753 758 ...

> head(data1)

AGE FEMALE LOS RACE TOTCHG APRDRG

1 17 1 2 1 2660 560

2 17 0 2 1 1689 753

3 17 1 7 1 20060 930

4 17 1 1 1 736 758

5 17 1 1 1 1194 754

6 17 0 0 1 3305 347

> tail(data1)

AGE FEMALE LOS RACE TOTCHG APRDRG

495 0 1 3 1 1886 640

496 0 1 6 1 5881 636

497 0 1 2 1 1171 640

498 0 1 2 1 1171 640

499 0 1 2 1 1086 640

500 0 0 4 1 4931 640

> summary(data1)

AGE FEMALE LOS

Min. : 0.000 Min. :0.000 Min. : 0.000

1st Qu.: 0.000 1st Qu.:0.000 1st Qu.: 2.000

Median : 0.000 Median :1.000 Median : 2.000

Mean : 5.086 Mean :0.512 Mean : 2.828

3rd Qu.:13.000 3rd Qu.:1.000 3rd Qu.: 3.000

Max. :17.000 Max. :1.000 Max. :41.000

RACE TOTCHG APRDRG

Min. :1.000 Min. : 532 Min. : 21.0

1st Qu.:1.000 1st Qu.: 1216 1st Qu.:640.0

Median :1.000 Median : 1536 Median :640.0

Mean :1.078 Mean : 2774 Mean :616.4

3rd Qu.:1.000 3rd Qu.: 2530 3rd Qu.:751.0

Max. :6.000 Max. :48388 Max. :952.0

NA's :1

# STEP 3 : SPLIT THE DATA INTO TRAINING AND TESTING : ( 70:30)

data1 <- read.csv("/home/kuntalc1\_gmail/1555054100\_hospitalcosts.csv", header= TRUE,sep=",")

install.packages("caTools") #install the “caTools” package

library(caTools) #import package

# dividing data into Training and Testing

# sample the input data with 70% for training and 30% for testing

# NOTE : here dependent variable is TOTCHG ( Hospital discharge costs)

sample <- sample.split(data1$TOTCHG,SplitRatio=0.70)

#split of the data using subset command

train\_data <- subset(data1,sample==TRUE)

test\_data <- subset(data1,sample==FALSE)

# STEP 4 : BUILD THE MODEL USING TRAINING DATA#Here we are doing multiple linear regression model

#building model using train\_data

model <- lm(TOTCHG ~.,data= train\_data )

summary(model)

> summary(model)

Call:

lm(formula = TOTCHG ~ ., data = train\_data)

Residuals:

Min 1Q Median 3Q Max

-6795.3 -493.7 -146.3 146.3 16121.6

Coefficients:

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

(Intercept) 6063.5481 415.0698 14.609 < 2e-16 \*\*\*

AGE 115.6067 14.7142 7.857 5.13e-14 \*\*\*

FEMALE -455.6563 209.1361 -2.179 0.03 \*

LOS 705.7524 26.7288 26.404 < 2e-16 \*\*\*

RACE -143.0735 211.7918 -0.676 0.50

APRDRG -9.0667 0.5431 -16.694 < 2e-16 \*\*\*

---

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

Residual standard error: 1862 on 343 degrees of freedom

(1 observation deleted due to missingness)

Multiple R-squared: 0.749, Adjusted R-squared: 0.7453

F-statistic: 204.7 on 5 and 343 DF, p-value: < 2.2e-16

# STEP 5 : MODEL INTERPRETATION

1) In the above output, look at the P values in column **Pr(>|t|)**2) **Here we can see multiple Significant Variable** AGE,FEMALE,LOS and APRDRG. **So we need to rebuild the model again**

**with these Significant variable**

# STEP 6 : IF MORE THAN ONE SIG VAR, THEN REBUILD THE MODEL AGAIN USING ONLY SIGNIFICANT VAR

final\_model <- lm(TOTCHG ~ AGE+FEMALE+LOS+APRDRG, data = train\_data)summary(final\_model)

> summary(final\_model)

Call:

lm(formula = TOTCHG ~ AGE + FEMALE + LOS + APRDRG, data = train\_data)

Residuals:

Min 1Q Median 3Q Max

-6782.8 -484.9 -148.0 149.6 16132.2

Coefficients:

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

(Intercept) 5909.5978 345.9258 17.083 < 2e-16 \*\*\*

AGE 115.1596 14.6474 7.862 4.88e-14 \*\*\*

FEMALE -449.8285 208.1206 -2.161 0.0314 \*

LOS 705.8836 26.6634 26.474 < 2e-16 \*\*\*

APRDRG -9.0654 0.5419 -16.729 < 2e-16 \*\*\*

---

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

Residual standard error: 1858 on 345 degrees of freedom

Multiple R-squared: 0.7488, Adjusted R-squared: 0.7459

F-statistic: 257.1 on 4 and 345 DF, p-value: < 2.2e-16

# STEP 7 : MODEL INTERPRETATION OF FINAL\_MODEL

1) In the above output of the Coefficient table, look at the P values in column **Pr(>|t|)**

2) Here P values of AGE,FEMALE,LOS & APRDRG are less than 0.05 and which are significant variable 3) Slope of the significant variable AGE & LOS are positive and FEMALE & APRDRG are negative4) So AGE & LOS are positively corelated with TOTCHG and FEMALE & APRDRG are negatively corelated with TOTCHG 5) Here R SQUARE is 75% ( CUT OFF > 70% )

# STEP 8 : PREDICTION - PREDICT THE FINAL\_MODEL USING TEST\_DATA

predtest<- predict(final\_model,test\_data)

head(predtest)

> head(predtest)

1 5 8 10 17 25

3752.6229 1288.0507 1878.7746 1957.6726 942.5718 1533.2958

# attach it with the dataframe

predtest1<- data.frame(predtest)

View(predtest1)

# STEP 9 : COMBIND TEST\_DATA AND PREDICTED DATA

#to combind the predicted data set with original data set by cbind function

final\_data<- cbind(test\_data,predtest1)

# STEP 10 : EXPORT THE FINAL DATA INTO NEW CSV FILE

#export write.csv(final\_data,"Project9\_output.csv" )

# STEP 11 : LOOK AT THE PREDICTED TOTCHG VALUES

Look at the predicted TOTCHG values in the new CSV file which has been exported in the previous step (STEP 10)

# STEP 12 : CALCULATE THE RESIDUAL VALUE (ORIGINAL TOTCHG VALUE - PREDICTED TOTCHG VALUE )

Here in this final output(Project9\_output.csv) in STEP 10 , from the top 10 observation shows that some Residual value (TOTCHG – predtest) are positive and some are negative, which is random pattern also not high, that shows the model provides a decent fit to the data.