Healthcare_cost_analysis

BY: KRISHNA ARYAL

8/8/2020

Healthcare cost analysis

Background and Objective: A nationwide survey of hospital costs conducted by the US Agency for Healthcare consists of hospital records of inpatient samples. The given data is restricted to the city of Wisconsin and relates to patients in the age group 0-17 years. The agency wants to analyze the data to research on healthcare costs and their utilization.

Domain: Healthcare

Dataset Description:

Here is a detailed description of the given dataset:

Attribute Description 1. Age Age of the patient discharged

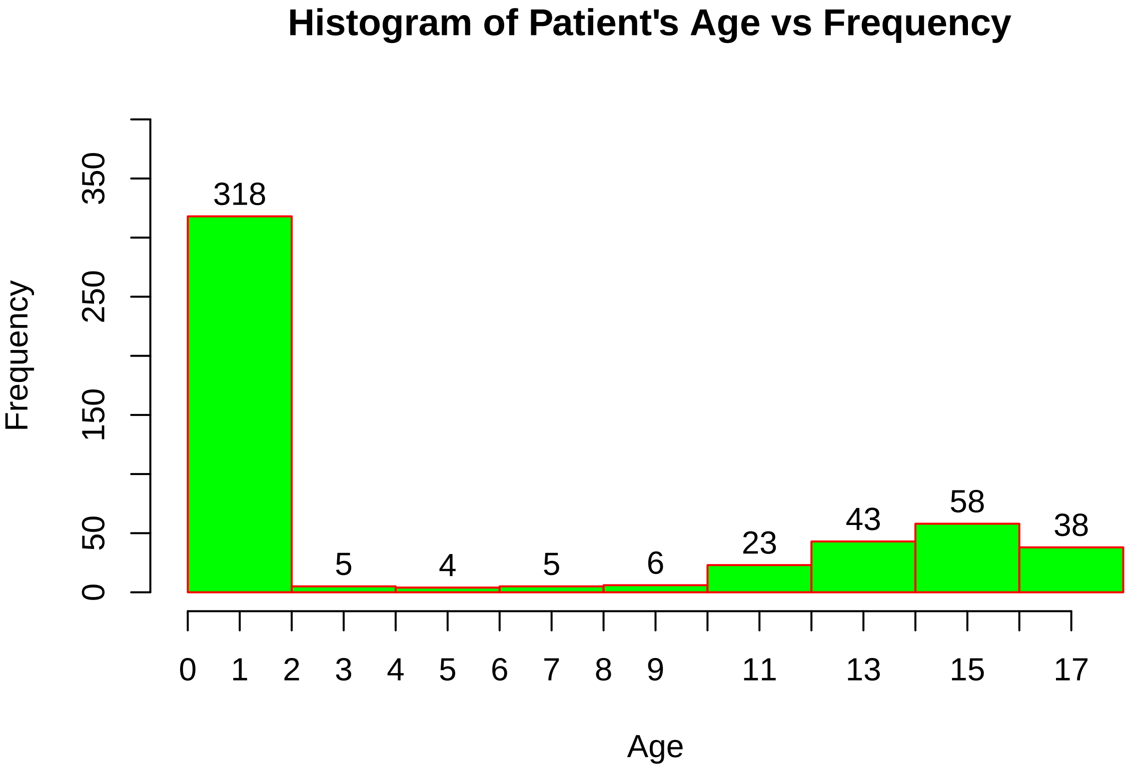
2. Female A binary variable that indicates if the patient is female
3. Los Length of stay in days
4. Race

   Race of the patient (specified numerically)
5. Totchg Hospital discharge costs
6. Aprdrg All Patient Refined Diagnosis Related Groups

ANALYSIS TO BE DONE:

1. TO RECORD THE PATIENT STATISTICS, THE AGENCY WANTS TO FIND THE AGE CATEGORY OF PEOPLE WHO FREQUENT THE HOSPITAL AND HAS THE MAXIMUM EXPENDITURE.

```
hos_cost <- read.csv("HospitalCosts.csv")
#TO FIND AGE CATEGORY OF PEOPLE WHO VISITED HOSPITAL FREQUENTLY
Age<-hos_cost$AGE
hist(Age,freq = TRUE,right = TRUE,border = "red",
     main=paste("Histogram of Patient's Age vs Frequency "),col="green",
     axes = TRUE,
     labels = TRUE, ylim=c(0,400),
     xlim = c(0, 18),xaxp=c(0,17,17),
     yaxp=c(0,400,8))
```



```
aggregate(TOTCHG ~ AGE, FUN = sum, data = hos_cost)
```

```
##    AGE TOTCHG
## 1     0 678118
## 2     1  37744
## 3     2   7298
## 4     3  30550
## 5     4  15992
## 6     5  18507
## 7     6  17928
## 8     7  10087
## 9     8   4741
## 10    9  21147
## 11   10  24469
## 12   11  14250
## 13   12  54912
## 14   13  31135
## 15   14  64643
## 16   15 111747
## 17   16  69149
## 18   17 174777
```

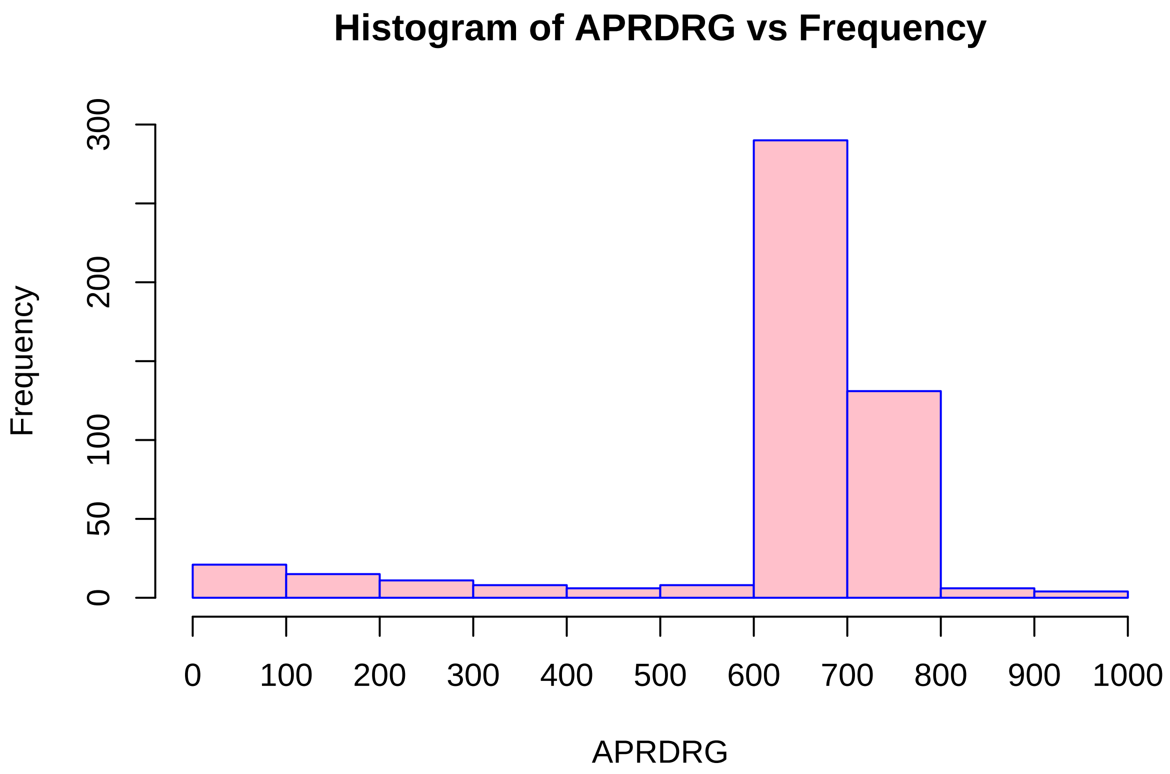
THIS HISTOGRAM SHOWS 0 AGE PEOPLE VISITED MOST FREQUENTLY.

Expenditure: 0 age people is 678118 and 15 years age cost 111747.

Expenditure wise cost:0 age people 678118 and 17 years age people cost 174777

2. IN ORDER OF SEVERITY OF THE DIAGNOSIS AND TREATMENTS AND TO FIND OUT THE EXPENSIVE TREATMENTS, THE AGENCY WANTS TO FIND THE DIAGNOSIS-RELATED GROUP THAT HAS MAXIMUM HOSPITALIZATION AND EXPENDITURE.

```r
#TO FIND THE EXPENDITURE OF by category
hist(hos_cost$APRDRG,freq = TRUE,right = TRUE,
     border = "blue",
  main=paste("Histogram of APRDRG vs Frequency"), col="pink",
  axes = TRUE, xlab="APRDRG",
  ylab = "Frequency", ylim =c(0,300),
     xlim = c(0, 1000),xaxp=c(0,1000,10),
     yaxp=c(0,300,6))
```



In category 600-700 people have visited hospital frequently.

let's find which number related dignosis group visited that.

```
factor_cost <-as.factor(hos_cost$APRDRG)
summary(factor_cost)
```

```
##  21  23  49  50  51  53  54  57  58  92  97 114 115 137 138 139 141 143 204 206
##   1   1   1   1   1  10   1   2   1   1   1   1   2   1   4   5   1   1   1   1
## 225 249 254 308 313 317 344 347 420 421 422 560 561 566 580 581 602 614 626 633
##   2   6   1   1   1   1   2   3   2   1   3   2   1   1   1   3   1   3   6   4
## 634 636 639 640 710 720 723 740 750 751 753 754 755 756 758 760 776 811 812 863
##   2   3   4 267   1   1   2   1   1  14  36  37  13   2  20   2   1   2   3   1
## 911 930 952
##   1   2   1
```

```
which.max(summary(factor_cost))
```

```
## 640
##  44
```

THIS SHOW 640 DIAGNISTIC GROUP HAS MAXIMUM EXPENDITURE AND MAXIMUM HOSPITALIZATION

3. TO MAKE SURE THAT THERE IS NO MALPRACTICE, THE AGENCY NEEDS TO ANALYZE IF THE RACE OF THE PATIENT IS RELATED TO THE HOSPITALIZATION COSTS.

We are interested in anova analysis. H0:All means are equal among races. There is no malpractice. H1:There is some biasness and people are treated according with race. Known fact:if p value<0.05 : H0 is rejected, H1 is accepted P value>0.05 , H0 is accepted Degrees of freedom: df(num)=k-1 # k= number of groups has been tested here

```
hos_cost2<-na.omit(hos_cost)
factor_race <-as.factor(hos_cost2$RACE)
summary(factor_race)
```

```
##   1   2   3   4   5   6
## 484   6   1   3   3   2
```

```
# KEY IDEAS
# 6 RACES SO 5 DEGREES OF FREEDOM
#Calculate the test statistic: F=2.23
#Define probability statement: p-value=P(F>2.23)=0.1241
#Decide: IF ALPHA < P, ACCEPT H0
ano_model <- aov(TOTCHG~factor_race,data=hos_cost2)
summary(ano_model)
```

```
##              Df    Sum Sq  Mean Sq F value Pr(>F)
## factor_race   5 1.859e+07  3718656   0.244  0.943
## Residuals   493 7.524e+09 15260687
```

p value (0.943) > 0.05 so H0 is accepted. The Residual Value (deviation of the observed value) is very high specifying that there is no relation between the race of patient and the hospital cost.

From the summary we can also see that the data has 484 patients of Race 1 out of the 500 entries. This will affect the results of ANOVA as well, since the number of observations is very much skewed.

Hence we can conclude that there is no race wise cost bias in the observed data.

4.TO PROPERLY UTILIZE THE COSTS, THE AGENCY HAS TO ANALYZE THE SEVERITY OF THE HOSPITAL COSTS BY AGE AND GENDER FOR THE PROPER ALLOCATION OF RESOURCES.

To analyse whether gender and age are related with hospital cost. Linear regression model is useful.

```r
lm_model <- lm(hos_cost$TOTCHG ~ hos_cost$AGE+hos_cost$FEMALE)
summary(lm_model)
```

```
##
## Call:
## lm(formula = hos_cost$TOTCHG ~ hos_cost$AGE + hos_cost$FEMALE)
##
## Residuals:
##    Min     1Q Median     3Q    Max
##  -3406  -1443   -869   -152  44951
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)      2718.63     261.14  10.411  < 2e-16 ***
## hos_cost$AGE       86.28      25.48   3.387 0.000763 ***
## hos_cost$FEMALE  -748.19     353.83  -2.115 0.034967 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3845 on 497 degrees of freedom
## Multiple R-squared:  0.0261, Adjusted R-squared:  0.02218
## F-statistic:  6.66 on 2 and 497 DF,  p-value: 0.001399
```

CONCLUSION

P VALUES OF MODEL are LESS THAN 0.05 ARE AGE,gender &CORRESPONDING SLOPES ARE 86.28 ,-748.19

AGE is POSITIVELY RELATED. So older patient cost is more.

Female IS NEGATIVELY RELATED.so being female pay less cost than male.

5. SINCE THE LENGTH OF STAY IS THE CRUCIAL FACTOR FOR INPATIENTS, THE AGENCY WANTS TO FIND IF THE LENGTH OF STAY CAN BE PREDICTED FROM AGE, GENDER, AND RACE.

To analyse whether length of stay is related with age,gender and race. multiple Linear regression model is useful.

```
lm_model2 <- lm(hos_cost$LOS ~ hos_cost$AGE+hos_cost$FEMALE +hos_cost$RACE)
summary(lm_model2)
```

```
##
## Call:
## lm(formula = hos_cost$LOS ~ hos_cost$AGE + hos_cost$FEMALE +
##     hos_cost$RACE)
##
## Residuals:
##    Min     1Q Median     3Q    Max
##  -3.22  -1.22  -0.85   0.15  37.78
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)      2.94377    0.39318   7.487 3.25e-13 ***
## hos_cost$AGE    -0.03960    0.02231  -1.775   0.0766 .
## hos_cost$FEMALE  0.37011    0.31024   1.193   0.2334
## hos_cost$RACE   -0.09408    0.29312  -0.321   0.7484
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.363 on 495 degrees of freedom
##   (1 observation deleted due to missingness)
## Multiple R-squared:  0.007898,   Adjusted R-squared:  0.001886
## F-statistic: 1.314 on 3 and 495 DF,  p-value: 0.2692
```

CONCLUSION

The significance codes are almost null for all the variables, except for the intercept.

The p-value high which signifies that there is no linear relationship between the given variables.

Hence we cannot predict the length of stay of the patients based on the age, gender, and race.

6. TO PERFORM A COMPLETE ANALYSIS, THE AGENCY WANTS TO FIND THE VARIABLE THAT MAINLY AFFECTS HOSPITAL COSTS.

To analyse whether hospital is related with all other variables. multiple Linear regression model is useful.

```
model2<- lm(TOTCHG ~.,data = hos_cost)
summary(model2)
```

```
##
## Call:
## lm(formula = TOTCHG ~ ., data = hos_cost)
##
## Residuals:
##    Min     1Q Median     3Q    Max
##  -6377   -700   -174    122  43378
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 5218.6769   507.6475  10.280  < 2e-16 ***
## AGE          134.6949    17.4711   7.710 7.02e-14 ***
## FEMALE      -390.6924   247.7390  -1.577    0.115
## LOS          743.1521    34.9225  21.280  < 2e-16 ***
## RACE        -212.4291   227.9326  -0.932    0.352
## APRDRG        -7.7909     0.6816 -11.430  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2613 on 493 degrees of freedom
##   (1 observation deleted due to missingness)
## Multiple R-squared:  0.5536, Adjusted R-squared:  0.5491
## F-statistic: 122.3 on 5 and 493 DF,  p-value: < 2.2e-16
```

CONCLUSION

Looking p values hospital cost is mostly depends on age (7.02e-14 < 0.05) and slope is 134.6949. This implies older age people have higher hospital cost.

p value of LOS is also affecting factor. as P VALUE 2e-16$<$0.05. SLOPE IS HIGHER POSITIVE 743.1521. iMPLIES HIGHER LENGTH OF STAY COSTS MORE.

Another important factor affecting cost is All Patient Refined Diagnosis Related Groups. p value is 2e-16 $<$ 0.05 and slope is -7.7909. This group affecting in negative way.