DESCRIPTION

**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** |
| 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 |

=====================================================================

**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.

read\_xlsx("1555054100\_hospitalcosts.xlsx")->h

h

View(h)

colSums(is.na(health))

h$RACE[is.na(h$RACE)]=1

barplot(table(h$AGE), main = "Age Vs frequency",

xlab = "Age",

ylab = "Frequency", ylim = c(0, 400))

barplot(tapply(X=h$TOTCHG, INDEX = list(h$AGE), FUN = sum))

> barplot(table(h$AGE), main = "Age Vs frequency",

+ xlab = "Age",

+ ylab = "Frequency", ylim = c(0, 400))



From the graph, Age ‘0’ or new born babies have highest frequency visiting the hospital.

barplot(tapply(X=h$TOTCHG, INDEX = list(h$AGE), FUN = sum),

main = "Age-Group Vs TOTCHG",

xlab = "Age-Group",

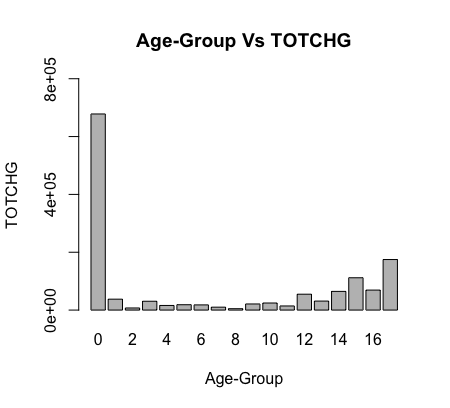
ylab = "TOTCHG", ylim = c(0, 8 \* (10)^5))

> barplot(tapply(X=h$TOTCHG, INDEX = list(h$AGE), FUN = sum),

+ main = "Age-Group Vs TOTCHG",

+ xlab = "Age-Group",

+ ylab = "TOTCHG", ylim = c(0, 8 \* (10)^5))



Comments: From the graph, we can say that Age “0” or new born babies have the peak TOTCHG.

Age 2 - 11 shows very low TOTCHG.

While Age 14-16 shows moderate TTCHG.

=====================================================================

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.

aggregate(h[, c("LOS", "TOTCHG")], by=list(h$APRDRG), FUN = sum)->tab

max(tab)

> aggregate(h[, c("LOS", "TOTCHG")], by=list(h$APRDRG), FUN = sum)->tab

> max(tab)

Output: 640 652 437978

The APRDRG “640” has maxi. Aggregate value of LOS and TOTCHG.

Group.1 LOS TOTCHG

1 21 2 10002

2 23 2 14174

3 49 6 20195

4 50 2 3908

5 51 3 3023

6 53 29 82271

7 54 1 851

8 57 2 14509

9 58 1 2117

10 92 1 12024

11 97 3 9530

12 114 3 10562

13 115 8 25832

14 137 12 15129

15 138 6 13622

16 139 7 17766

17 141 2 2860

18 143 2 1393

19 204 3 8439

20 206 3 9230

21 225 11 25649

22 249 8 16642

23 254 0 615

24 308 1 10585

25 313 1 8159

26 317 7 17524

27 344 5 14802

28 347 1 12597

29 420 3 6357

30 421 39 26356

31 422 5 5177

32 560 4 4877

33 561 4 2296

34 566 2 2129

35 580 1 2825

36 581 1 7453

37 602 41 29188

38 614 45 27531

39 626 36 23289

40 633 25 17591

41 634 9 9952

42 636 20 23224

43 639 18 12612

44 640 652 437978

45 710 4 8223

46 720 5 14243

47 723 2 5289

48 740 5 11125

49 750 3 1753

50 751 32 21666

51 753 109 79542

52 754 85 59150

53 755 13 11168

54 756 2 1494

55 758 53 34953

56 760 14 8273

57 776 2 1193

58 811 1 3838

59 812 5 9524

60 863 24 13040

61 911 7 48388

62 930 10 26654

63 952 1 4833

=====================================================================

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.

table(h$RACE)

barplot(tapply(X= h$TOTCHG, INDEX = h$RACE, FUN = mean),

main = "RACE Vs TOTCHG",

xlab = "RACE",

ylab = "TOTCHG", ylim = c(0, 5000))

aov(formula = TOTCHG~RACE, h)->model

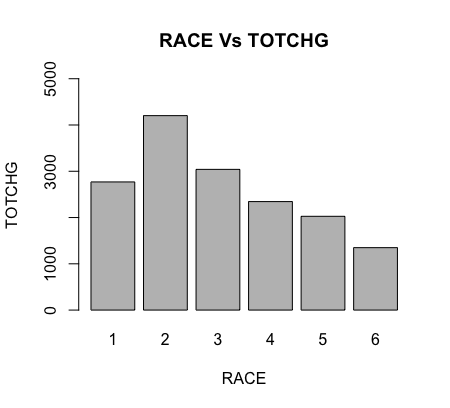
summary(model)

> barplot(tapply(X= h$TOTCHG, INDEX = h$RACE, FUN = mean),

+ main = "RACE Vs TOTCHG",

+ xlab = "RACE",

+ ylab = "TOTCHG", ylim = c(0, 5000))



> aov(formula = TOTCHG~RACE, h)->model

> summary(model)

Df Sum Sq Mean Sq F value Pr(>F)

RACE 5 1.861e+07 3721895 0.244 0.943

Residuals 494 7.526e+09 15235074

Comments: The Pr value is 94.3 % and TOTCHG is not varying significantly w.r.to RACE.

=====================================================================

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.

aov(formula = TOTCHG~AGE+FEMALE, h)->model

summary(model)

> aov(formula = TOTCHG~AGE+FEMALE, h)->model

> summary(model)

Df Sum Sq Mean Sq F value Pr(>F)

AGE 1 1.308e+08 130822234 8.849 0.00308 \*\*

FEMALE 1 6.610e+07 66104210 4.471 0.03497 \*

Residuals 497 7.348e+09 14784325

---

Signif. codes:

0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Age And Gender both are significant for variance in TOTCHG(Hospital charge).

lm(TOTCHG~AGE+FEMALE, h)->model

summary(model)

> lm(TOTCHG~AGE+FEMALE, h)->model

> summary(model)

Call:

lm(formula = TOTCHG ~ AGE + FEMALE, data = h)

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 \*\*\*

AGE 86.28 25.48 3.387 0.000763 \*\*\*

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

**Comments**: The P-value is 0.1%, Adjusted R-squared is 2.218%.

AGE is more significantly affected than GENDER and is between 0 and 0.1% significance , and gender is between 1% and 5%.

As age increases increase by 1 unit, total charge increases by 86.28 units

As female increases from 0 to 1 (from male to female), the total charge decreases by 748 units.

=====================================================================

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.

lm(LOS~AGE + FEMALE + RACE, h)->model

summary(model)

> lm(LOS~AGE + FEMALE + RACE, h)->model

> summary(model)

Call:

lm(formula = LOS ~ AGE + FEMALE + RACE, data = h)

Residuals:

Min 1Q Median 3Q Max

-3.204 -1.204 -0.856 0.144 37.796

Coefficients:

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

(Intercept) 2.85563 0.23137 12.342 <2e-16 \*\*\*

AGE -0.03902 0.02254 -1.731 0.084 .

FEMALE 0.34799 0.31221 1.115 0.266

RACE2 -0.37573 1.39444 -0.269 0.788

RACE3 0.79638 3.38275 0.235 0.814

RACE4 0.59690 1.95542 0.305 0.760

RACE5 -0.85563 1.96098 -0.436 0.663

RACE6 -0.71745 2.39082 -0.300 0.764

---

Signif. codes:

0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.373 on 492 degrees of freedom

Multiple R-squared: 0.008562, Adjusted R-squared: -0.005544

F-statistic: 0.607 on 7 and 492 DF, p-value: 0.7503

Comments: age, gender and race are able to explain only .7% of the variance in los

=====================================================================

6. To perform a complete analysis, the agency wants to find the variable that mainly affects hospital costs.

aov(TOTCHG~., h)->model

summary(model)

> aov(TOTCHG~., h)->model

> summary(model)

Df Sum Sq Mean Sq F value Pr(>F)

AGE 1 1.308e+08 1.308e+08 19.074 1.54e-05 \*\*\*

FEMALE 1 6.610e+07 6.610e+07 9.638 0.00202 \*\*

LOS 1 3.087e+09 3.087e+09 450.075 < 2e-16 \*\*\*

RACE 5 1.325e+07 2.649e+06 0.386 0.85829

APRDRG 1 8.870e+08 8.870e+08 129.333 < 2e-16 \*\*\*

Residuals 490 3.361e+09 6.859e+06

---

Signif. codes:

0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Comments: The TOTCHG is significantly (highly)Varied by “LOS” & “APRDRG” of2e-16 is between 0 and 0.1 % significance.

The TOTCHG is moderately Varied by “AGE” than “FEMLES”.

While RACE has no significance.

=====================================================================