# Healthcare cost analysis

Andrei Enescu

11/2/2020

# 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 frequently visit the hospital and has the maximum expenditure.

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.

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.

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.

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.

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

---

These are the libraries I used:

```r
library(rio)
library(dplyr)
```

Importing the data using RIO package

```r
HospitalDF <- import("1555054100_hospitalcosts.xlsx")
```

Checking the data

```r
head(HospitalDF)
```

```
##   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
```

```r
str(HospitalDF)
```

```
## ’data.frame':    500 obs. of  6 variables:
##  $ AGE   : num  17 17 17 17 17 17 17 16 16 17 ...
##  $ FEMALE: num  1 0 1 1 1 0 1 1 1 1 ...
##  $ LOS   : num  2 2 7 1 1 0 4 2 1 2 ...
##  $ RACE  : num  1 1 1 1 1 1 1 1 1 1 ...
##  $ TOTCHG: num  2660 1689 20060 736 1194 ...
##  $ APRDRG: num  560 753 930 758 754 347 754 754 753 758 ...
```

```r
summary(HospitalDF)
```

```
##       AGE             FEMALE           LOS              RACE      
##  Min.   : 0.000   Min.   :0.000   Min.   : 0.000   Min.   :1.000  
##  1st Qu.: 0.000   1st Qu.:0.000   1st Qu.: 2.000   1st Qu.:1.000  
##  Median : 0.000   Median :1.000   Median : 2.000   Median :1.000  
##  Mean   : 5.086   Mean   :0.512   Mean   : 2.828   Mean   :1.078  
##  3rd Qu.:13.000   3rd Qu.:1.000   3rd Qu.: 3.000   3rd Qu.:1.000  
##  Max.   :17.000   Max.   :1.000   Max.   :41.000   Max.   :6.000  
##                                                    NA’s   :1      
##      TOTCHG          APRDRG     
##  Min.   :  532   Min.   : 21.0  
##  1st Qu.: 1216   1st Qu.:640.0  
##  Median : 1536   Median :640.0  
##  Mean   : 2774   Mean   :616.4  
##  3rd Qu.: 2530   3rd Qu.:751.0  
##  Max.   :48388   Max.   :952.0  
## 
```

Changing Age, Female, Race variables into factors

```
HospitalDF$AGE <- as.factor(HospitalDF$AGE)
HospitalDF$FEMALE <- as.factor(HospitalDF$FEMALE)
HospitalDF$RACE <- as.factor(HospitalDF$RACE)
HospitalDF$APRDRG <- as.factor(HospitalDF$APRDRG)

str(HospitalDF)
```

```
## ’data.frame’:    500 obs. of  6 variables:
##  $ AGE   : Factor w/ 18 levels "0","1","2","3",..: 18 18 18 18 18 18 18 17 17 18 ...
##  $ FEMALE: Factor w/ 2 levels "0","1": 2 1 2 2 2 1 2 2 2 2 ...
##  $ LOS   : num  2 2 7 1 1 0 4 2 1 2 ...
##  $ RACE  : Factor w/ 6 levels "1","2","3","4",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ TOTCHG: num  2660 1689 20060 736 1194 ...
##  $ APRDRG: Factor w/ 63 levels "21","23","49",..: 32 51 62 55 52 28 52 52 51 55 ...
```

## 1. To record the patient statistics, the agency wants to find the age category of people who frequently visit the hospital and has the maximum expenditure.

```
summary(HospitalDF$AGE)
```

```
##   0   1   2   3   4   5   6   7   8   9  10  11  12  13  14  15  16  17
## 307  10   1   3   2   2   2   3   2   2   4   8  15  18  25  29  29  38
```

**Comments:**

Infants under 1 year group has the most hospital visits - **307**

```
HospitalDF %>%
  group_by(AGE) %>%
  summarise(Expenditure = sum(TOTCHG)) %>%
  arrange(desc(Expenditure))
```

```
## ‘summarise()‘ ungrouping output (override with ‘.groups‘ argument)
```

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

**Comments:**

Infants under 1 year group has the most hospital costs - **678118** Based on the above outputs we can conclude that hospital costs are directly proportional to hospital visits.

## 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.

```
count(HospitalDF, APRDRG, sort = T)
```

```
##    APRDRG   n
## 1     640 267
## 2     754  37
## 3     753  36
## 4     758  20
## 5     751  14
## 6     755  13
## 7      53  10
## 8     249   6
## 9     626   6
## 10    139   5
## 11    138   4
## 12    633   4
## 13    639   4
## 14    347   3
## 15    422   3
## 16    581   3
## 17    614   3
## 18    636   3
## 19    812   3
## 20     57   2
## 21    115   2
## 22    225   2
## 23    344   2
## 24    420   2
## 25    560   2
## 26    634   2
## 27    723   2
## 28    756   2
## 29    760   2
## 30    811   2
## 31    930   2
## 32     21   1
## 33     23   1
## 34     49   1
## 35     50   1
## 36     51   1
## 37     54   1
## 38     58   1
## 39     92   1
## 40     97   1
## 41    114   1
## 42    137   1
## 43    141   1
## 44    143   1
## 45    204   1
## 46    206   1
## 47    254   1
## 48    308   1
## 49    313   1
## 50    317   1
## 51    421   1
## 52    561   1
## 53    566   1
## 54    580   1
## 55    602   1
## 56    710   1
## 57    720   1
## 58    740   1
## 59    750   1
## 60    776   1
## 61    863   1
## 62    911   1
## 63    952   1
```

```
HospitalDF %>%
  group_by(APRDRG) %>%
  summarise(Expenditure = sum(TOTCHG)) %>%
  arrange(desc(Expenditure))
```

```
## `summarise()` ungrouping output (override with `.groups` argument)
```

```
## # A tibble: 63 x 2
##    APRDRG Expenditure
##    <fct>        <dbl>
##  1 640         437978
##  2 53           82271
##  3 753          79542
##  4 754          59150
##  5 911          48388
##  6 758          34953
##  7 602          29188
##  8 614          27531
##  9 930          26654
## 10 421          26356
## # ... with 53 more rows
```

**Comments:**

Diagnosis-related group 640 has the most hospitalizations - **267** out of **500** - and the highest expenditure - **437978**

## 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.

Removing NA value

```
HospitalDF <- na.omit(HospitalDF)
```

Using ANOVA

```
fit1 <- aov(HospitalDF$TOTCHG ~ HospitalDF$RACE)
summary(fit1)
```

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

```
summary(HospitalDF$RACE)
```

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

**Comments:**

As the p-value is higher than the significance level 0.05, we can conclude that there are no significant differences between the groups in the model summary. So by accepting the Null hypothesis we can say that there is no relationship between race and hospitalization costs.
Furthermore we can observe that we don't have a normal distributed data for Race, where in group 1 we have 484 patients out of 500. we can conclude that we don't have enough information to say if tha race is affecting the costs.

## 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.

```r
fit2 <- glm(TOTCHG ~ AGE + FEMALE, family = gaussian(), HospitalDF)
summary(fit2)
```

```
##
## Call:
## glm(formula = TOTCHG ~ AGE + FEMALE, family = gaussian(), data = HospitalDF)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
##   -4624    -1139     -789      -76    43941
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2410.04     264.25   9.120  < 2e-16 ***
## AGE1         1453.34    1198.38   1.213 0.225818
## AGE2         4887.96    3728.85   1.311 0.190536
## AGE3         7921.60    2158.30   3.670 0.000269 ***
## AGE4         5808.42    2638.72   2.201 0.028195 *
## AGE5         7065.92    2638.72   2.678 0.007665 **
## AGE6         6553.96    2643.31   2.479 0.013501 *
## AGE7          952.29    2163.64   0.440 0.660038
## AGE8          -39.54    2643.31  -0.015 0.988071
## AGE9         8163.46    2643.31   3.088 0.002129 **
## AGE10        3818.44    1873.11   2.039 0.042042 *
## AGE11        -517.56    1333.88  -0.388 0.698180
## AGE12        1517.72     985.15   1.541 0.124075
## AGE13        -334.26     909.75  -0.367 0.713467
## AGE14         549.42     786.19   0.699 0.484988
## AGE15        1734.81     726.50   2.388 0.017330 *
## AGE16         327.28     733.08   0.446 0.655474
## AGE17        2482.07     644.17   3.853 0.000132 ***
## FEMALE1      -444.93     353.03  -1.260 0.208158
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## (Dispersion parameter for gaussian family taken to be 13834479)
##
##     Null deviance: 7542111784  on 498  degrees of freedom
## Residual deviance: 6640550108  on 480  degrees of freedom
## AIC: 9641.6
##
## Number of Fisher Scoring iterations: 2
```

```r
summary(HospitalDF$FEMALE)
```

```
##   0   1
## 244 255
```

**Comments:**

From above analysis we can conclude that the costs are not affected by gender, but are significant affected by age.

**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.**

```
fit3 <- glm(LOS ~ AGE + FEMALE + RACE, family = gaussian(), HospitalDF)
summary(fit3)
```

```
##
## Call:
## glm(formula = LOS ~ AGE + FEMALE + RACE, family = gaussian(),
##     data = HospitalDF)
##
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
## -3.262  -1.224  -0.892   0.045  37.776
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.95535    0.24457  12.084   <2e-16 ***
## AGE1        -1.20910    1.09842  -1.101   0.2716
## AGE2        -0.95535    3.41674  -0.280   0.7799
## AGE3         0.28840    1.97773   0.146   0.8841
## AGE4        -1.08973    2.41786  -0.451   0.6524
## AGE5        -0.58973    2.41786  -0.244   0.8074
## AGE6        -0.45535    2.42218  -0.188   0.8510
## AGE7        -2.62201    1.98274  -1.322   0.1867
## AGE8        -1.49810    2.53185  -0.592   0.5543
## AGE9        -0.95535    2.42218  -0.394   0.6935
## AGE10       -0.27254    1.71648  -0.159   0.8739
## AGE11       -1.65823    1.23557  -1.342   0.1802
## AGE12       -0.71661    0.90295  -0.794   0.4278
## AGE13       -0.86106    0.84041  -1.025   0.3061
## AGE14       -0.16271    0.72444  -0.225   0.8224
## AGE15        0.03803    0.66785   0.057   0.9546
## AGE16       -1.33221    0.68452  -1.946   0.0522 .
## AGE17       -0.50059    0.59066  -0.848   0.3971
## FEMALE1      0.26877    0.32509   0.827   0.4088
## RACE2        0.08552    1.49616   0.057   0.9544
## RACE3        0.77589    3.41835   0.227   0.8205
## RACE4        0.54007    2.00086   0.270   0.7873
## RACE5       -0.95535    1.98274  -0.482   0.6301
## RACE6       -0.42362    2.43389  -0.174   0.8619
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## (Dispersion parameter for gaussian family taken to be 11.61431)
##
##     Null deviance: 5644.5  on 498  degrees of freedom
## Residual deviance: 5516.8  on 475  degrees of freedom
## AIC: 2665.2
```

```
##
## Number of Fisher Scoring iterations: 2
```

**Comments:**

The p-values for all independent variables are high, so we can say that there is no relationship between the variables. We can conclude that based on the given data we can not predict the length of stay based on age, gender or race.

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

```r
fit4 <- lm(TOTCHG ~ ., HospitalDF)
summary(fit4)
```

```
##
## Call:
## lm(formula = TOTCHG ~ ., data = HospitalDF)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5431.1  -199.9   -54.4    91.1  5431.1
##
## Coefficients: (1 not defined because of singularities)
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  7136.47     967.14   7.379 8.84e-13 ***
## AGE1         -549.51     463.85  -1.185 0.236831
## AGE2          220.77     850.69   0.260 0.795365
## AGE3          937.26    1252.75   0.748 0.454791
## AGE4         3746.32    1022.77   3.663 0.000282 ***
## AGE5         1946.56     678.81   2.868 0.004348 **
## AGE6        -5066.18    1136.34  -4.458 1.07e-05 ***
## AGE7         -721.74     728.95  -0.990 0.322700
## AGE8          542.48     691.54   0.784 0.433226
## AGE9         2797.57    1205.25   2.321 0.020765 *
## AGE10        -798.55     747.50  -1.068 0.286013
## AGE11        1160.76     527.96   2.199 0.028462 *
## AGE12        1557.52     483.74   3.220 0.001384 **
## AGE13        1298.97     485.60   2.675 0.007770 **
## AGE14        1186.37     470.69   2.520 0.012096 *
## AGE15        1191.96     460.07   2.591 0.009913 **
## AGE16        1310.30     465.54   2.815 0.005117 **
## AGE17        1444.21     467.46   3.089 0.002141 **
## FEMALE1      -191.55      74.02  -2.588 0.009998 **
## LOS           650.96      19.87  32.767  < 2e-16 ***
## RACE2         253.31     409.40   0.619 0.536437
## RACE3         630.32     791.17   0.797 0.426086
## RACE4          84.36     426.49   0.198 0.843307
## RACE5        1531.61     833.44   1.838 0.066826 .
## RACE6         -52.82     526.05  -0.100 0.920073
## APRDRG23     4291.40    1117.40   3.841 0.000142 ***
## APRDRG49     7960.81    1124.89   7.077 6.36e-12 ***
## APRDRG50    -5499.60    1185.54  -4.639 4.71e-06 ***
## APRDRG51    -7319.01    1121.37  -6.527 1.97e-10 ***
## APRDRG53    -1361.16     909.62  -1.496 0.135310    
## APRDRG54    -8189.09    1121.33  -7.303 1.46e-12 ***
## APRDRG57     -640.69    1300.22  -0.493 0.622449    
## APRDRG58    -5120.92    1175.63  -4.356 1.67e-05 ***
## APRDRG92     3044.61    1123.52   2.710 0.007011 ** 
## APRDRG97     5506.83    1525.08   3.611 0.000343 ***
## APRDRG114   -1324.92    1019.61  -1.299 0.194517    
## APRDRG115    5025.79    1133.24   4.435 1.18e-05 ***
## APRDRG137     181.01    1218.80   0.149 0.882009    
## APRDRG138   -4659.53    1030.91  -4.520 8.09e-06 ***
## APRDRG139   -4587.25     968.30  -4.737 2.98e-06 ***
## APRDRG141   -4779.85    1259.82  -3.794 0.000170 ***
## APRDRG143   -8577.00    1464.58  -5.856 9.67e-09 ***
## APRDRG204   -2094.56    1117.59  -1.874 0.061613 .  
## APRDRG206   -3605.67    1463.83  -2.463 0.014178 *  
## APRDRG225     918.58     994.53   0.924 0.356213    
## APRDRG249   -5060.15     945.69  -5.351 1.46e-07 ***
## APRDRG254   -8053.08    1465.14  -5.496 6.80e-08 ***
## APRDRG308         NA         NA      NA       NA    
## APRDRG313   -1192.04    1019.03  -1.170 0.242762    
## APRDRG317    6629.35    1267.54   5.230 2.70e-07 ***
## APRDRG344   -1646.59    1039.03  -1.585 0.113792    
## APRDRG347   -4580.20    1050.32  -4.361 1.64e-05 ***
## APRDRG420   -6169.94    1003.85  -6.146 1.87e-09 ***
## APRDRG421   -6167.90    1403.94  -4.393 1.42e-05 ***
## APRDRG422   -7152.28     979.09  -7.305 1.44e-12 ***
## APRDRG560   -7252.55     998.76  -7.262 1.92e-12 ***
## APRDRG561   -8439.13    1122.68  -7.517 3.52e-13 ***
## APRDRG566   -7562.05    1121.17  -6.745 5.19e-11 ***
## APRDRG580   -4962.43    1205.25  -4.117 4.63e-05 ***
## APRDRG581   -4741.43    1053.18  -4.502 8.77e-06 ***
## APRDRG602   -4446.28    1426.29  -3.117 0.001952 ** 
## APRDRG614   -7596.17    1080.08  -7.033 8.44e-12 ***
## APRDRG626   -7138.09    1021.16  -6.990 1.11e-11 ***
## APRDRG633   -6711.45    1033.35  -6.495 2.39e-10 ***
## APRDRG634   -5089.79    1092.34  -4.660 4.28e-06 ***
## APRDRG636   -3607.17    1055.06  -3.419 0.000691 ***
## APRDRG639   -7199.92    1052.84  -6.839 2.89e-11 ***
## APRDRG640   -7002.51     966.22  -7.247 2.11e-12 ***
## APRDRG710   -2263.02    1655.49  -1.367 0.172375    
## APRDRG720    2914.47    1653.69   1.762 0.078740 .  
## APRDRG723   -5427.55     998.91  -5.433 9.46e-08 ***
## APRDRG740    -266.69    1127.60  -0.237 0.813156    
## APRDRG750   -8780.56    1117.59  -7.857 3.45e-14 ***
## APRDRG751   -8189.39     871.22  -9.400  < 2e-16 ***
## APRDRG753   -8054.64     851.98  -9.454  < 2e-16 ***
## APRDRG754   -8183.18     848.95  -9.639  < 2e-16 ***
## APRDRG755   -8132.31     863.50  -9.418  < 2e-16 ***
## APRDRG756   -8118.22    1017.01  -7.982 1.43e-14 ***
## APRDRG758   -8235.83     849.29  -9.697  < 2e-16 ***
## APRDRG760   -8551.51    1004.34  -8.515 3.14e-16 ***
## APRDRG776   -8689.60    1117.40  -7.777 6.00e-14 ***
## APRDRG811   -6602.54     985.38  -6.701 6.82e-11 ***
## APRDRG812   -6307.97     937.94  -6.725 5.85e-11 ***
## APRDRG863   -9527.96    1278.48  -7.453 5.42e-13 ***
## APRDRG911   35442.15    1125.66  31.486  < 2e-16 ***
## APRDRG930    1683.07    1000.60   1.682 0.093313 .
## APRDRG952   -4398.64    1117.56  -3.936 9.72e-05 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 720.4 on 413 degrees of freedom
## Multiple R-squared:  0.9716, Adjusted R-squared:  0.9657
## F-statistic: 166.1 on 85 and 413 DF,  p-value: < 2.2e-16
```

**Comments:**

Based on the above analysis we concluded that Total Charge is highly affected by: age,length of stay and , Diagnosis-related groups.
Gender is moderately affecting the Total Charge.
Race has no impact on Total Charge