Statistical models

Precious Ogunbekun

24/07/2021

```
Health <- read.table(file.choose(),sep=",",header=TRUE)
str(Health)
```

```
## '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 ...
```

```
summary(Health) #displays the mean, median, min and max for each variable
```

```
##       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
##
```

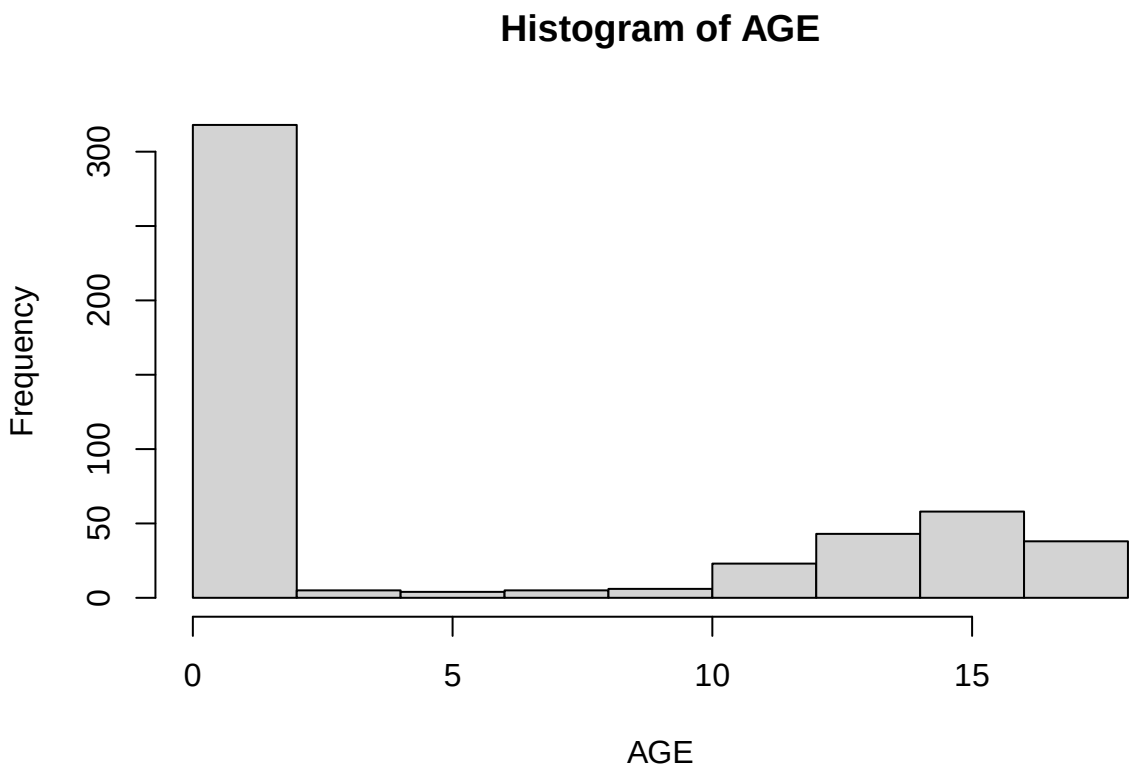
```
attach(Health)
```

```
dim(Health) #displays the no of rows and columns
```

```
## [1] 500   6
```

To find the age category that has the highest frequency of hospital visit

```
hist(Health$AGE, main = "Histogram of AGE", xlab ="AGE")
```




```
AGE <- as.factor(AGE)
max(table(AGE))
```

```
## [1] 307
```

From the graph that is displayed, infants has the maximum frequency of hospital visit, going above 300. After converting the age from numeric to factor. There are 307 entries for those in the range of 0-1 year.

Maximum expenditure for the age group who frequently visited the hospital

```
AGE_EXP <- aggregate(TOTCHG ~ AGE, data = Health, sum)
max(AGE_EXP)
```

```
## [1] 678118
```

To find the diagnosis related group that has maximum hospitalization and expenditure

```
APRDRG_FACTOR<-as.factor(APRDRG)
APRDRG_TABLE <- table(APRDRG_FACTOR)
which.max(APRDRG_TABLE)
```

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

```
APRDRG_TOTCHG<-aggregate(TOTCHG~APRDRG,FUN = sum,data=Health)
APRDRG_TOTCHG
```

```
##    APRDRG TOTCHG
## 1      21  10002
## 2      23  14174
## 3      49  20195
## 4      50   3908
## 5      51   3023
## 6      53  82271
## 7      54    851
## 8      57  14509
## 9      58   2117
## 10     92  12024
## 11     97   9530
## 12    114  10562
## 13    115  25832
## 14    137  15129
## 15    138  13622
## 16    139  17766
## 17    141   2860
## 18    143   1393
## 19    204   8439
## 20    206   9230
## 21    225  25649
## 22    249  16642
## 23    254    615
## 24    308  10585
## 25    313   8159
## 26    317  17524
## 27    344  14802
## 28    347  12597
## 29    420   6357
## 30    421  26356
## 31    422   5177
## 32    560   4877
## 33    561   2296
## 34    566   2129
## 35    580   2825
## 36    581   7453
## 37    602  29188
## 38    614  27531
## 39    626  23289
## 40    633  17591
## 41    634   9952
```

```
## 42    636  23224
## 43    639  12612
## 44    640 437978
## 45    710   8223
## 46    720  14243
## 47    723   5289
## 48    740  11125
## 49    750   1753
## 50    751  21666
## 51    753  79542
## 52    754  59150
## 53    755  11168
## 54    756   1494
## 55    758  34953
## 56    760   8273
## 57    776   1193
## 58    811   3838
## 59    812   9524
## 60    863  13040
## 61    911  48388
## 62    930  26654
## 63    952   4833
```

```
APRDRG_TOTCHG[which.max(APRDRG_TOTCHG$TOTCHG),]
```

```
##    APRDRG TOTCHG
## 44    640 437978
```

To analyze if the race of the patient is related to the hospitalization costs.

```
RACE <- as.factor(RACE)
table(RACE)
```

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

```
lm_Health <- lm((TOTCHG)~RACE)
anova(lm_Health)
```

```
## Analysis of Variance Table
##
## Response: (TOTCHG)
##            Df     Sum Sq  Mean Sq F value Pr(>F)
## RACE        5   18593279  3718656  0.2437 0.9429
## Residuals 493 7523518505 15260687
```

```
summary(lm_Health)
```

```
##
## Call:
## lm(formula = (TOTCHG) ~ RACE)
##
## Residuals:
##    Min     1Q Median     3Q    Max
##  -3049  -1551  -1223   -238  45615
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   2772.7      177.6  15.615   <2e-16 ***
## RACE2         1429.5     1604.7   0.891    0.373
## RACE3          268.3     3910.5   0.069    0.945
## RACE4         -428.0     2262.4  -0.189    0.850
## RACE5         -746.0     2262.4  -0.330    0.742
## RACE6        -1423.7     2768.0  -0.514    0.607
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3906 on 493 degrees of freedom
##   (1 observation deleted due to missingness)
## Multiple R-squared:  0.002465,   Adjusted R-squared:  -0.007652
## F-statistic: 0.2437 on 5 and 493 DF,  p-value: 0.9429
```

- The result of anova shows there is no significance relationship between RACE and TOTCHG
- The result of summary shows There is no significance difference between the different race except for RACE 1

To analyze the severity of the hospital costs by age and gender for proper allocation of resources.

```
FEMALE <- as.factor(FEMALE)
table(FEMALE)
```

```
## FEMALE
##   0   1
## 244 256
```

```
lm_health2 <- lm(TOTCHG ~ AGE*FEMALE)
anova(lm_health2)
```

```
## Analysis of Variance Table
##
## Response: TOTCHG
##             Df     Sum Sq  Mean Sq F value    Pr(>F)
## AGE         17  881098421 51829319  3.8442 4.628e-07 ***
## FEMALE       1   22436964 22436964  1.6642   0.19768
## AGE:FEMALE  12  317997801 26499817  1.9655   0.02563 *
## Residuals  469 6323203028 13482309
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
summary(lm_health2)
```

```
##
## Call:
## lm(formula = TOTCHG ~ AGE * FEMALE)
##
## Residuals:
##    Min     1Q Median     3Q    Max
##  -6616  -1079   -712      0  43457
##
## Coefficients: (5 not defined because of singularities)
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)    2198.44     281.62   7.806 3.87e-14 ***
## AGE1           2129.31    1328.38   1.603 0.109621
## AGE2           5099.56    3682.61   1.385 0.166781
## AGE3           8965.06    2611.60   3.433 0.000650 ***
## AGE4           7031.56    3682.61   1.909 0.056821 .
## AGE5           5724.56    3682.61   1.554 0.120743
## AGE6           6765.56    2611.60   2.591 0.009880 **
## AGE7           1163.90    2138.55   0.544 0.586531
## AGE8            172.06    2611.60   0.066 0.947498
## AGE9           8375.06    2611.60   3.207 0.001434 **
## AGE10          5571.23    2138.55   2.605 0.009475 **
## AGE11          -730.44    1525.24  -0.479 0.632234
## AGE12           393.73    1525.24   0.258 0.796409
## AGE13         -1144.44    1857.39  -0.616 0.538092
## AGE14          3542.56    1857.39   1.907 0.057094 .
## AGE15          5024.56    1194.80   4.205 3.12e-05 ***
## AGE16          2431.40    1525.24   1.594 0.111586
## AGE17          1762.72    1056.60   1.668 0.095926 .
## FEMALE1          23.35     421.57   0.055 0.955861
## AGE1:FEMALE1  -2790.10    2933.29  -0.951 0.342001
## AGE2:FEMALE1        NA         NA      NA       NA
## AGE3:FEMALE1  -2963.85    4516.77  -0.656 0.512025
## AGE4:FEMALE1  -2491.35    5209.83  -0.478 0.632730
## AGE5:FEMALE1   2637.65    5209.83   0.506 0.612895
## AGE6:FEMALE1        NA         NA      NA       NA
## AGE7:FEMALE1        NA         NA      NA       NA
## AGE8:FEMALE1        NA         NA      NA       NA
## AGE9:FEMALE1        NA         NA      NA       NA
## AGE10:FEMALE1 -6633.01    4260.77  -1.557 0.120201
## AGE11:FEMALE1  1229.65    3027.53   0.406 0.684812
## AGE12:FEMALE1  1757.71    1980.61   0.887 0.375286
## AGE13:FEMALE1   845.44    2123.99   0.398 0.690779
## AGE14:FEMALE1 -3779.63    2047.02  -1.846 0.065464 .
## AGE15:FEMALE1 -5166.50    1495.17  -3.455 0.000599 ***
## AGE16:FEMALE1 -2854.48    1735.21  -1.645 0.100633
## AGE17:FEMALE1   946.78    1324.43   0.715 0.475052
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3672 on 469 degrees of freedom
## Multiple R-squared:  0.1619, Adjusted R-squared:  0.1083
```

```
## F-statistic:  3.02 on 30 and 469 DF,  p-value: 3.372e-07
```

- The feature Female has no interaction with the hospitalization cost(TOTCHG) but AGE is related to (TOTCHG)
- There is interaction between (AGE & FEMALE) summary show the features and the interactions between the different labels of the factors

To find if the length of stay can be predicted from age, gender, and race.

```
lm_health3 <- lm(LOS ~ AGE+FEMALE+RACE)
anova(lm_health3)
```

```
## Analysis of Variance Table
##
## Response: LOS
##            Df Sum Sq Mean Sq F value Pr(>F)
## AGE        17  114.2  6.7192  0.5785 0.9083
## FEMALE      1    9.0  8.9539  0.7709 0.3804
## RACE        5    4.5  0.9091  0.0783 0.9955
## Residuals 475 5516.8 11.6143
```

```
summary(lm_health3)
```

```
##
## Call:
## lm(formula = LOS ~ AGE + FEMALE + RACE)
##
## 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
##
## Residual standard error: 3.408 on 475 degrees of freedom
##   (1 observation deleted due to missingness)
## Multiple R-squared:  0.02263,    Adjusted R-squared:  -0.0247
## F-statistic: 0.4781 on 23 and 475 DF,  p-value: 0.982
```

There is no interaction between LOS, AGE, FEMALE & RACE (probability value > 0.05)

To find the variable that mainly affects the hospital costs.

```
lm_health4 <- lm(TOTCHG ~ AGE+FEMALE+RACE+LOS+APRDRG)
anova(lm_health4)
```

```
## Analysis of Variance Table
##
## Response: TOTCHG
##            Df     Sum Sq    Mean Sq  F value  Pr(>F)
## AGE        17  879586115   51740360   8.0103 < 2e-16 ***
## FEMALE      1   21975561   21975561   3.4022 0.06573 .
## RACE        5   21969733    4393947   0.6803 0.63859
## LOS         1 3094369121 3094369121 479.0629 < 2e-16 ***
## APRDRG      1  469003475  469003475  72.6100 < 2e-16 ***
## Residuals 473 3055207780    6459213
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
summary(lm_health4)
```

```
##
## Call:
## lm(formula = TOTCHG ~ AGE + FEMALE + RACE + LOS + APRDRG)
##
## Residuals:
##    Min     1Q Median     3Q    Max
##  -6569   -512    -86    141  42413
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4368.5711   532.7100   8.201 2.27e-15 ***
## AGE1         -433.1797   883.4084  -0.490 0.624113
## AGE2         1816.6829  2586.8067   0.702 0.482845
## AGE3         6863.9862  1478.2591   4.643 4.45e-06 ***
## AGE4         4214.6000  1825.6000   2.309 0.021395 *
```

```
## AGE5         3571.3152  1861.5497   1.918 0.055654 .
## AGE6         3476.9284  1850.5402   1.879 0.060877 .
## AGE7          -11.7278  1520.8054  -0.008 0.993850
## AGE8         -681.6356  1891.3461  -0.360 0.718711
## AGE9         6175.0573  1834.3967   3.366 0.000824 ***
## AGE10        2467.6101  1293.1114   1.908 0.056961 .
## AGE11         311.6371   923.4195   0.337 0.735904
## AGE12        2205.9895   674.0471   3.273 0.001143 **
## AGE13        1096.4096   634.2893   1.729 0.084540 .
## AGE14        1039.0420   542.0455   1.917 0.055854 .
## AGE15        1828.2204   498.5187   3.667 0.000273 ***
## AGE16        1521.5701   513.6479   2.962 0.003207 **
## AGE17        3020.1897   441.2231   6.845 2.38e-11 ***
## FEMALE1      -336.2591   245.2515  -1.371 0.171001
## RACE2        1300.8573  1117.3647   1.164 0.244922
## RACE3         347.0865  2549.4288   0.136 0.891766
## RACE4         -66.6115  1492.3905  -0.045 0.964418
## RACE5       -1451.1043  1493.8042  -0.971 0.331838
## RACE6        -321.5116  1815.5013  -0.177 0.859512
## LOS           737.9944    34.2414  21.553  < 2e-16 ***
## APRDRG         -6.8536     0.8043  -8.521  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2541 on 473 degrees of freedom
##   (1 observation deleted due to missingness)
## Multiple R-squared:  0.5949, Adjusted R-squared:  0.5735
## F-statistic: 27.79 on 25 and 473 DF,  p-value: < 2.2e-16
```

There is significance relationship between TOTCHG and (AGE,LOS,APRDRG)