machine learning approach

Julia Sheriff

10/28/2018

GENERAL IDEA OF DATASET

```
hc1 <- read.csv("health_charges_clean.csv")
colnames(hc1)
```

```
##  [1] "X"              "age"            "sex"            "bmi"
##  [5] "bmi_factor"     "children"       "smoker"         "region"
##  [9] "charges"        "charges_factor" "age_factor"
```

```
hc1 <- hc1[ c(-1)]
str(hc1)
```

```
## 'data.frame':    1338 obs. of  10 variables:
##  $ age           : int  19 18 28 33 32 31 46 37 37 60 ...
##  $ sex           : Factor w/ 2 levels "female","male": 1 2 2 2 2 1 1 1 2 1 ...
##  $ bmi           : num  27.9 33.8 33 22.7 28.9 ...
##  $ bmi_factor    : Factor w/ 6 levels "healthy_weight",..: 5 2 2 1 5 5 2 5 5 5 ...
##  $ children      : int  0 1 3 0 0 0 1 3 2 0 ...
##  $ smoker        : Factor w/ 2 levels "no","yes": 2 1 1 1 1 1 1 1 1 1 ...
##  $ region        : Factor w/ 4 levels "northeast","northwest",..: 4 3 3 2 2 3 3 2 1 2 ...
##  $ charges       : num  16885 1726 4449 21984 3867 ...
##  $ charges_factor: Factor w/ 2 levels "high","low": 1 2 2 1 2 2 2 2 2 1 ...
##  $ age_factor    : Factor w/ 6 levels "10s","20s","30s",..: 1 1 2 3 3 3 4 3 3 6 ...
```

```
head(hc1)
```

```
##   age    sex    bmi     bmi_factor children smoker    region   charges
## 1  19 female 27.900     overweight        0    yes southwest 16884.924
## 2  18   male 33.770        obese_1        1     no southeast  1725.552
## 3  28   male 33.000        obese_1        3     no southeast  4449.462
## 4  33   male 22.705 healthy_weight        0     no northwest 21984.471
## 5  32   male 28.880     overweight        0     no northwest  3866.855
## 6  31 female 25.740     overweight        0     no southeast  3756.622
##   charges_factor age_factor
## 1           high        10s
## 2            low        10s
## 3            low        20s
## 4           high        30s
## 5            low        30s
## 6            low        30s
```

```
hc2 <- read.csv("binary_charges.csv")
colnames(hc2)
```

```
##  [1] "X"                         "charges_factor_high"
##  [3] "charges_factor_low"        "bmi_factor_overweight"
##  [5] "bmi_factor_obese_1"        "bmi_factor_healthy_weight"
##  [7] "bmi_factor_obese_2"        "bmi_factor_obese_3"
##  [9] "bmi_factor_underweight"    "children_0"
## [11] "children_1"                 "children_3"
## [13] "children_2"                 "children_5"
## [15] "children_4"                 "smoker_yes"
## [17] "smoker_no"                  "region_southwest"
## [19] "region_southeast"           "region_northwest"
## [21] "region_northeast"           "sex_female"
## [23] "sex_male"
```

```r
hc2 <- hc2[ c(-1)]
str(hc2)
```

```
## 'data.frame':    1338 obs. of  22 variables:
##  $ charges_factor_high      : int  1 0 0 1 0 0 0 0 0 1 ...
##  $ charges_factor_low       : int  0 1 1 0 1 1 1 1 1 0 ...
##  $ bmi_factor_overweight    : int  1 0 0 0 1 1 0 1 1 1 ...
##  $ bmi_factor_obese_1       : int  0 1 1 0 0 0 1 0 0 0 ...
##  $ bmi_factor_healthy_weight: int  0 0 0 1 0 0 0 0 0 0 ...
##  $ bmi_factor_obese_2       : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ bmi_factor_obese_3       : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ bmi_factor_underweight   : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ children_0               : int  1 0 0 1 1 1 0 0 0 1 ...
##  $ children_1               : int  0 1 0 0 0 0 1 0 0 0 ...
##  $ children_3               : int  0 0 1 0 0 0 0 1 0 0 ...
##  $ children_2               : int  0 0 0 0 0 0 0 0 1 0 ...
##  $ children_5               : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ children_4               : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ smoker_yes               : int  1 0 0 0 0 0 0 0 0 0 ...
##  $ smoker_no                : int  0 1 1 1 1 1 1 1 1 1 ...
##  $ region_southwest         : int  1 0 0 0 0 0 0 0 0 0 ...
##  $ region_southeast         : int  0 1 1 0 0 1 1 0 0 0 ...
##  $ region_northwest         : int  0 0 0 1 1 0 0 1 0 1 ...
##  $ region_northeast         : int  0 0 0 0 0 0 0 0 1 0 ...
##  $ sex_female               : int  1 0 0 0 0 1 1 1 0 1 ...
##  $ sex_male                 : int  0 1 1 1 1 0 0 0 1 0 ...
```

```r
head(hc2)
```

```
##   charges_factor_high charges_factor_low bmi_factor_overweight
## 1                   1                  0                     1
## 2                   0                  1                     0
## 3                   0                  1                     0
## 4                   1                  0                     0
## 5                   0                  1                     1
## 6                   0                  1                     1
##   bmi_factor_obese_1 bmi_factor_healthy_weight bmi_factor_obese_2
## 1                  0                         0                  0
## 2                  1                         0                  0
## 3                  1                         0                  0
## 4                  0                         1                  0
## 5                  0                         0                  0
## 6                  0                         0                  0
##   bmi_factor_obese_3 bmi_factor_underweight children_0 children_1
## 1                  0                      0          1          0
## 2                  0                      0          0          1
## 3                  0                      0          0          0
## 4                  0                      0          1          0
## 5                  0                      0          1          0
## 6                  0                      0          1          0
##   children_3 children_2 children_5 children_4 smoker_yes smoker_no
## 1          0          0          0          0          1         0
## 2          0          0          0          0          0         1
## 3          1          0          0          0          0         1
## 4          0          0          0          0          0         1
## 5          0          0          0          0          0         1
## 6          0          0          0          0          0         1
##   region_southwest region_southeast region_northwest region_northeast
## 1                1                0                0                0
## 2                0                1                0                0
## 3                0                1                0                0
## 4                0                0                1                0
## 5                0                0                1                0
## 6                0                1                0                0
##   sex_female sex_male
## 1          1        0
## 2          0        1
## 3          0        1
## 4          0        1
## 5          0        1
## 6          1        0
```

LINEAR REGRESSIONS:

- Checking class of variables:

```
str(hc1$smoker)
```

```
##  Factor w/ 2 levels "no","yes": 2 1 1 1 1 1 1 1 1 1 ...
```

```
str(hc1$region)
```

```
##  Factor w/ 4 levels "northeast","northwest",..: 4 3 3 2 2 3 3 2 1 2 ...
```

```
str(hc1$sex)
```

```
##  Factor w/ 2 levels "female","male": 1 2 2 2 2 1 1 1 2 1 ...
```

```
str(hc1$children)
```

```
##  int [1:1338] 0 1 3 0 0 0 1 3 2 0 ...
```

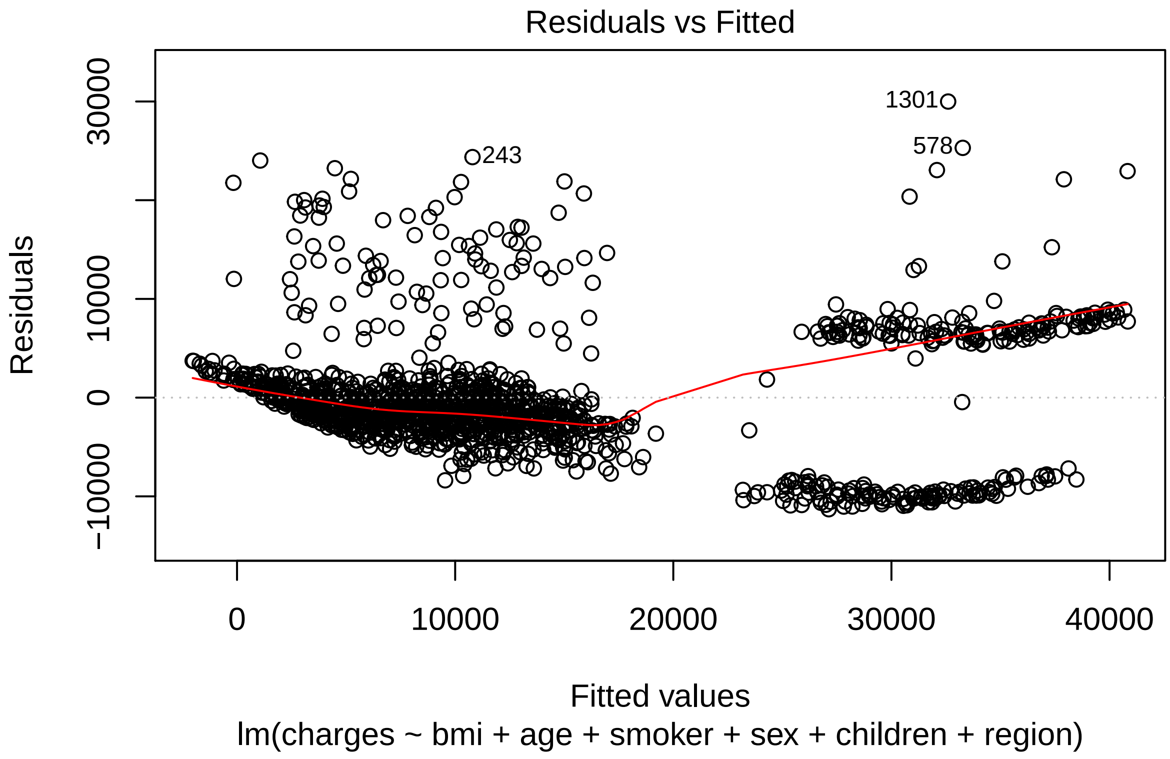
```
#because sex, region, and smoker are factors, the regression output will give a comparative interpretio
#children is integer, so children will be treated as a variable in itself.
```

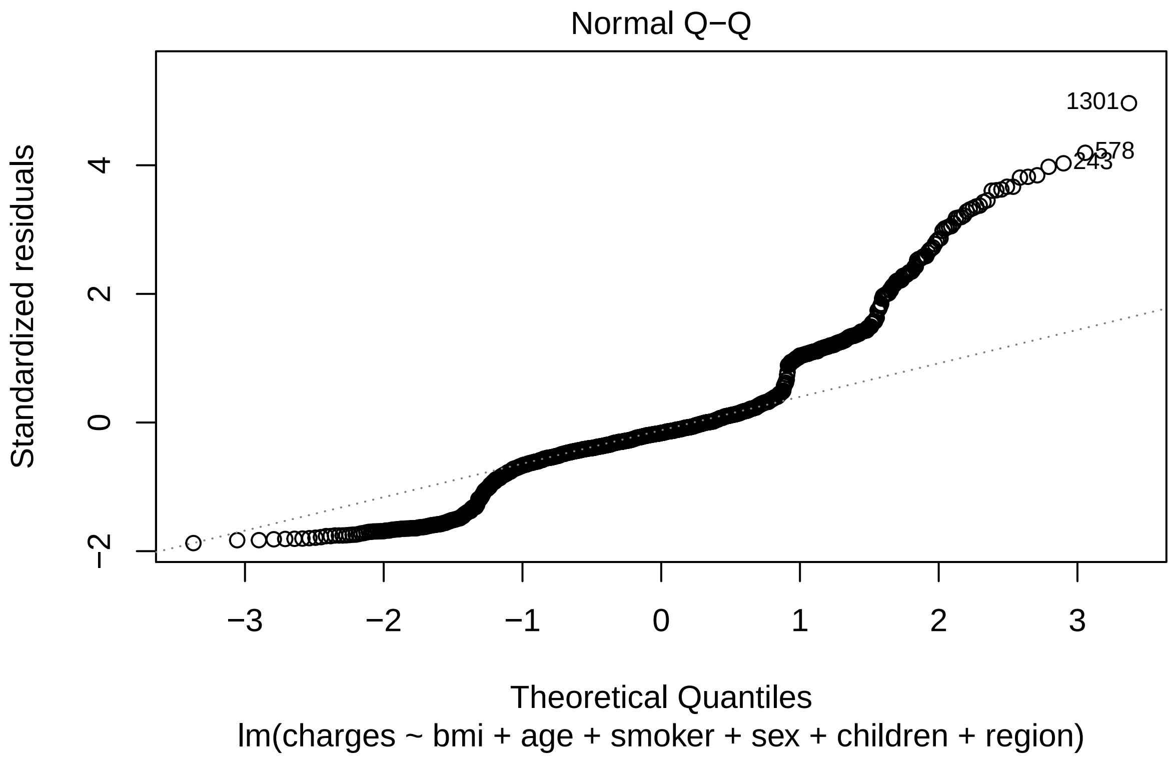
- Linear model:
- Adjusted R-squared is .7494, so the 75% of the value of the charges can be attributed to these variables.
- Significant variables:
  - being a smoker increases charges by $23848.9
  - having higher numbers of children increases charges by $475.5
  - higher bmi increases charges by $339.2
  - higher age increases charges by $256.9

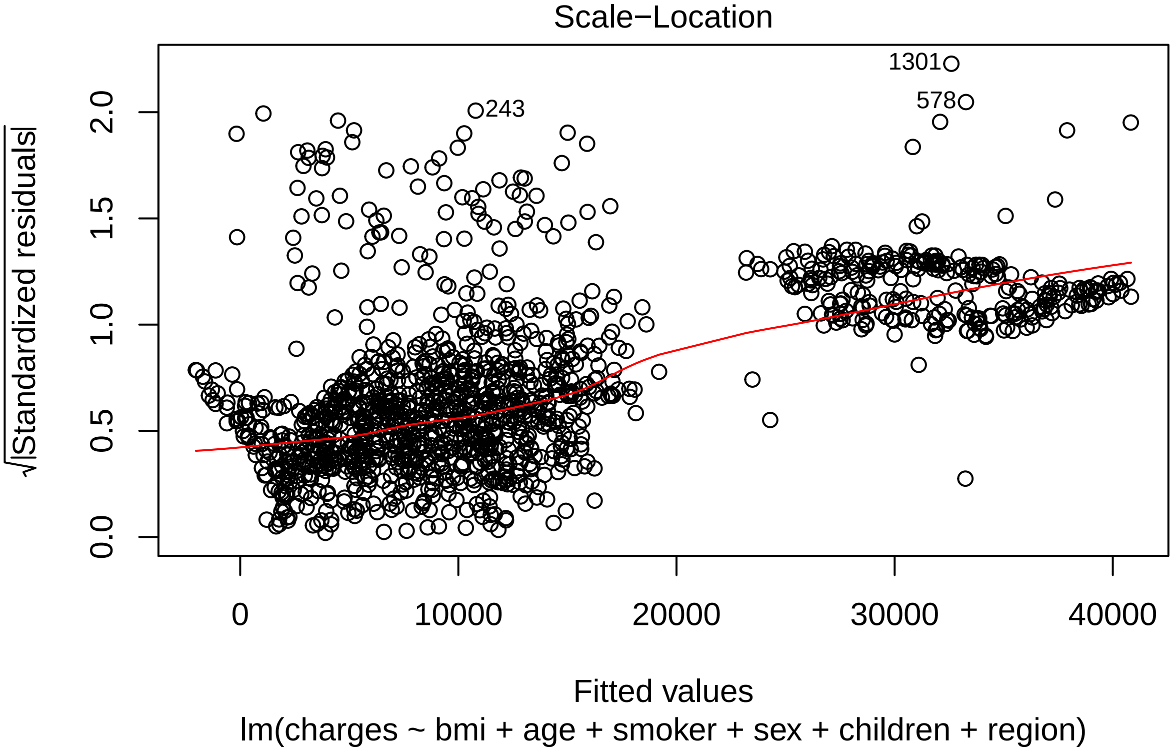
```
lmall <- lm ( charges ~ bmi + age + smoker + sex + children + region, data = hc1)
summary(lmall, method = lm)
```

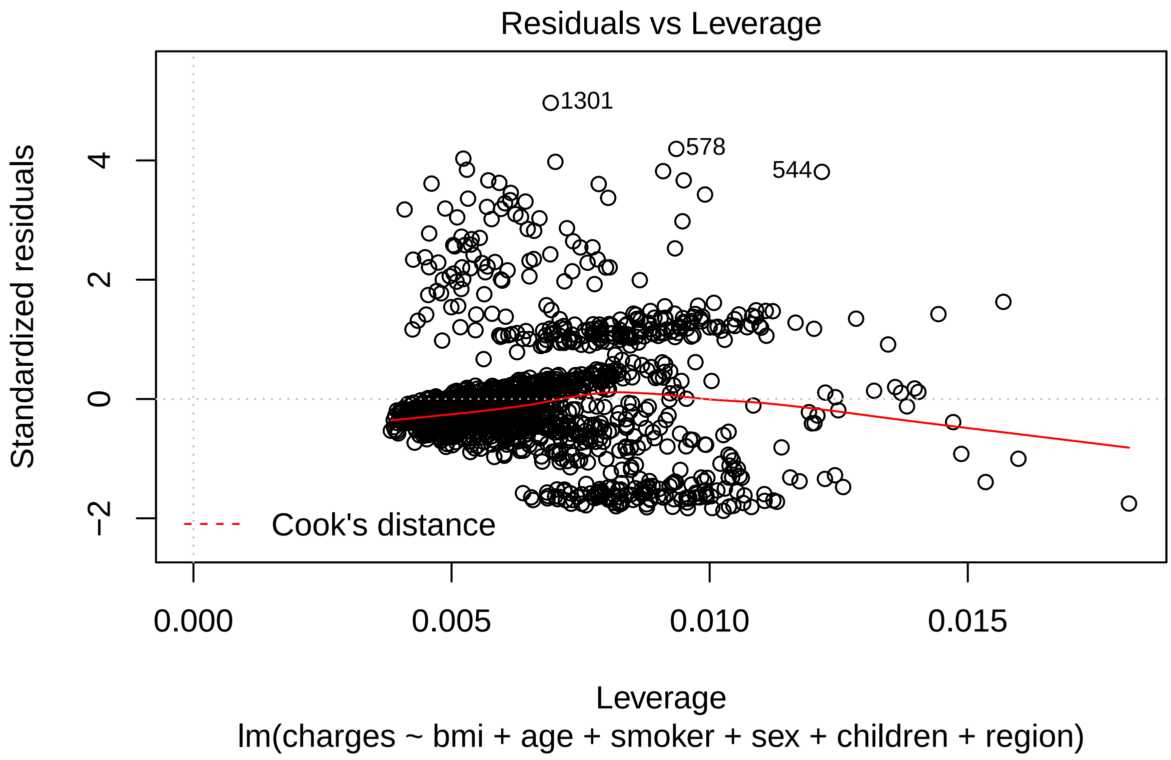
```
##
## Call:
## lm(formula = charges ~ bmi + age + smoker + sex + children +
##     region, data = hc1)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -11304.9  -2848.1   -982.1   1393.9  29992.8
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)     -11938.5      987.8 -12.086  < 2e-16 ***
## bmi                339.2       28.6  11.860  < 2e-16 ***
## age                256.9       11.9  21.587  < 2e-16 ***
## smokeryes        23848.5      413.1  57.723  < 2e-16 ***
## sexmale           -131.3      332.9  -0.394 0.693348
## children           475.5      137.8   3.451 0.000577 ***
## regionnorthwest   -353.0      476.3  -0.741 0.458769
## regionsoutheast  -1035.0      478.7  -2.162 0.030782 *
## regionsouthwest   -960.0      477.9  -2.009 0.044765 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6062 on 1329 degrees of freedom
## Multiple R-squared:  0.7509, Adjusted R-squared:  0.7494
## F-statistic: 500.8 on 8 and 1329 DF,  p-value: < 2.2e-16
```

```r
plot(lmall)
```




Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(charges ~ bmi + age + smoker + sex + children + region)
−2
0
2
4
−3
−2
−1
0
1
2
3
1301
578
243


Scale−Location
√|Standardized residuals|
Fitted values
lm(charges ~ bmi + age + smoker + sex + children + region)
0.0
0.5
1.0
1.5
2.0
0
10000
20000
30000
40000
243
1301
578



LOGISTIC REGRESSIONS:

Create testing set:

```
library(caTools)
set.seed(88)

split  = sample.split(hc1$charges_factor, SplitRatio = .75 )
hc1train = subset(hc1, split == TRUE)
hc1test= subset(hc1, split == FALSE)
```

Logistic regression:

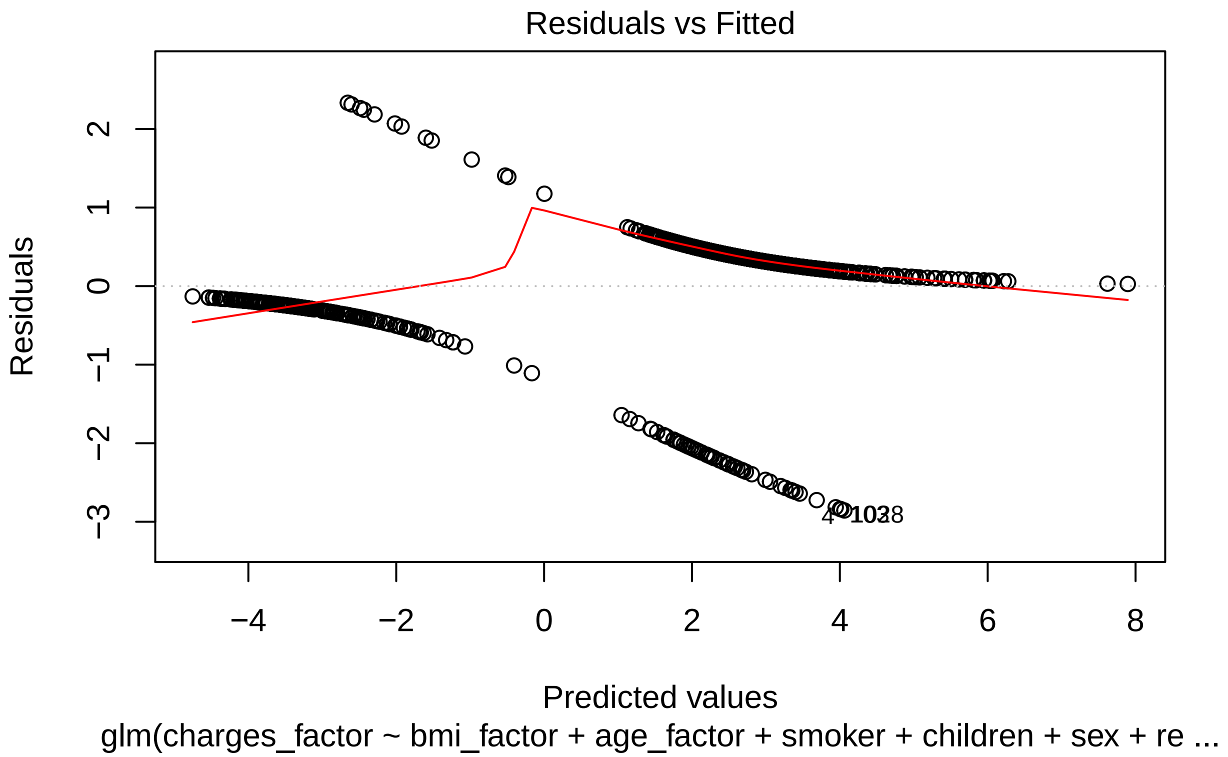
- Important variables:
  - Being a smoker.
  - Age. Being in 40s was a less significant predictor than other age brackets.
  - BMI: overweight, obese1, obese2

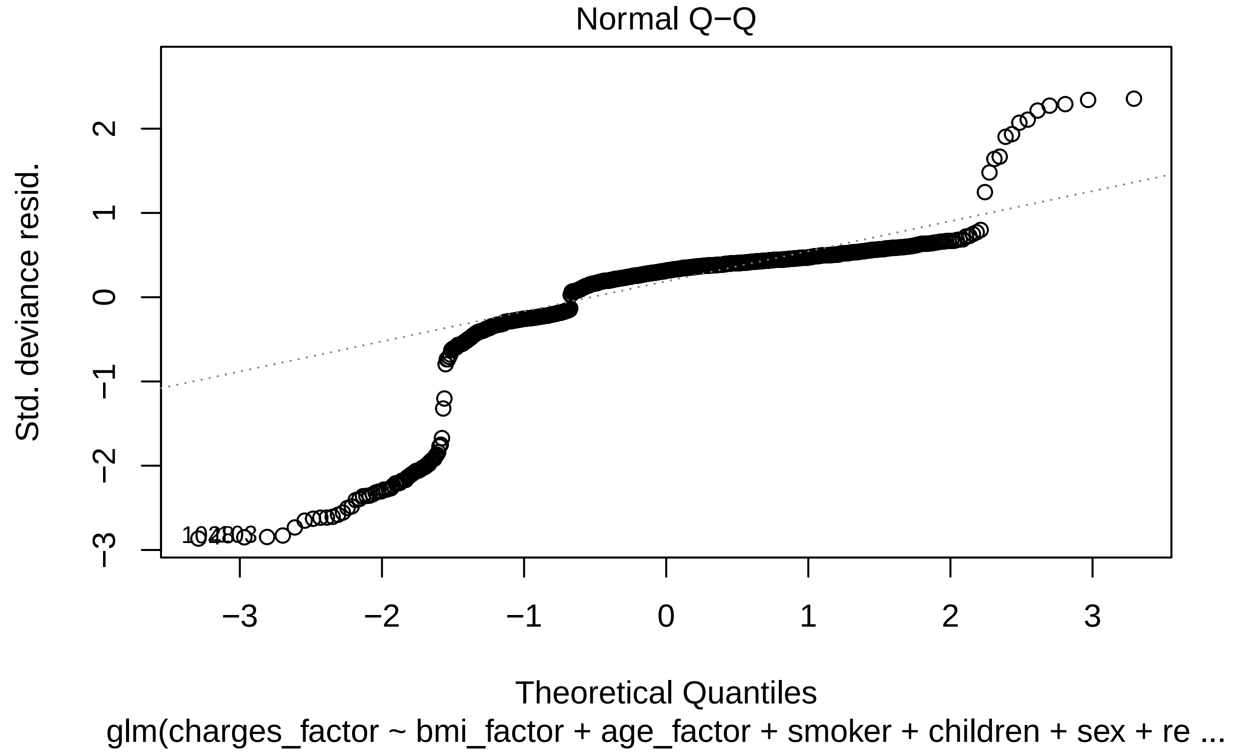
```
lgall= glm(charges_factor ~ bmi_factor + age_factor + smoker + children + sex + region,  data = hc1trai
summary(lgall)
```

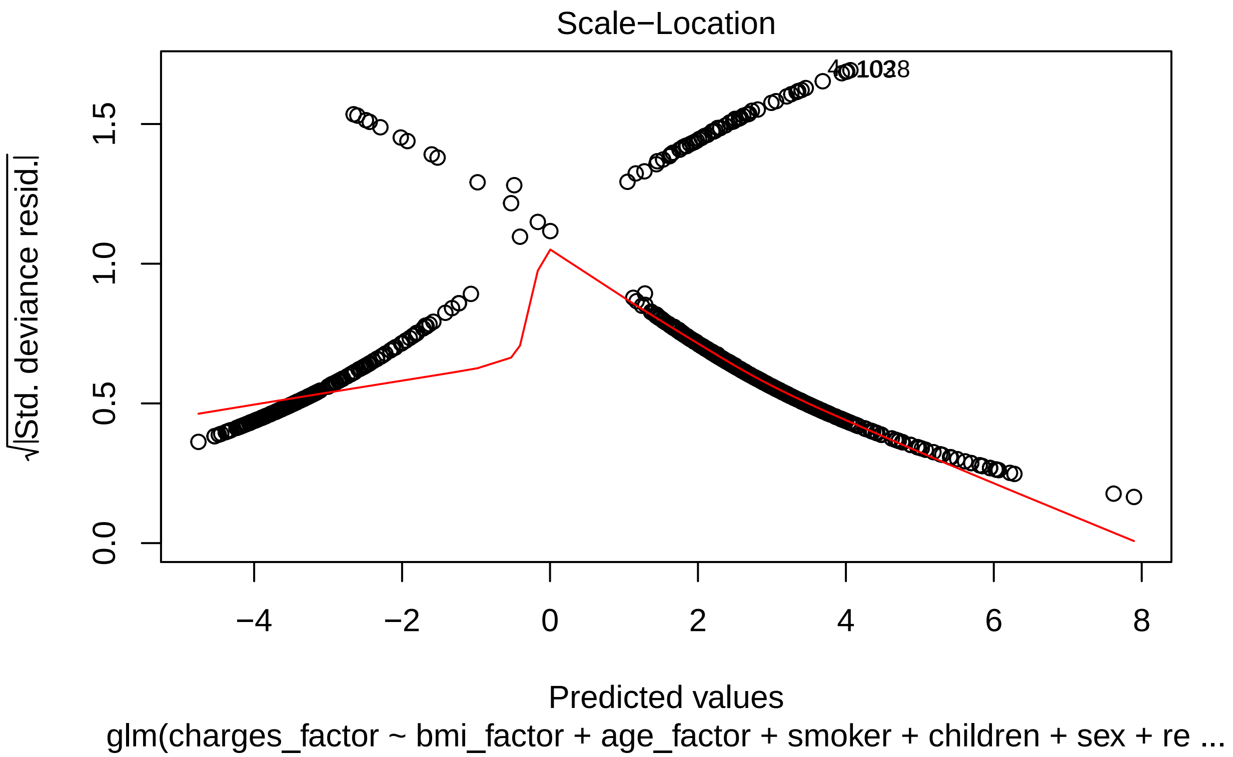
```
##
## Call:
## glm(formula = charges_factor ~ bmi_factor + age_factor + smoker +
##     children + sex + region, family = binomial, data = hc1train)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -2.85612  -0.05187   0.31353   0.42475   2.33385
```

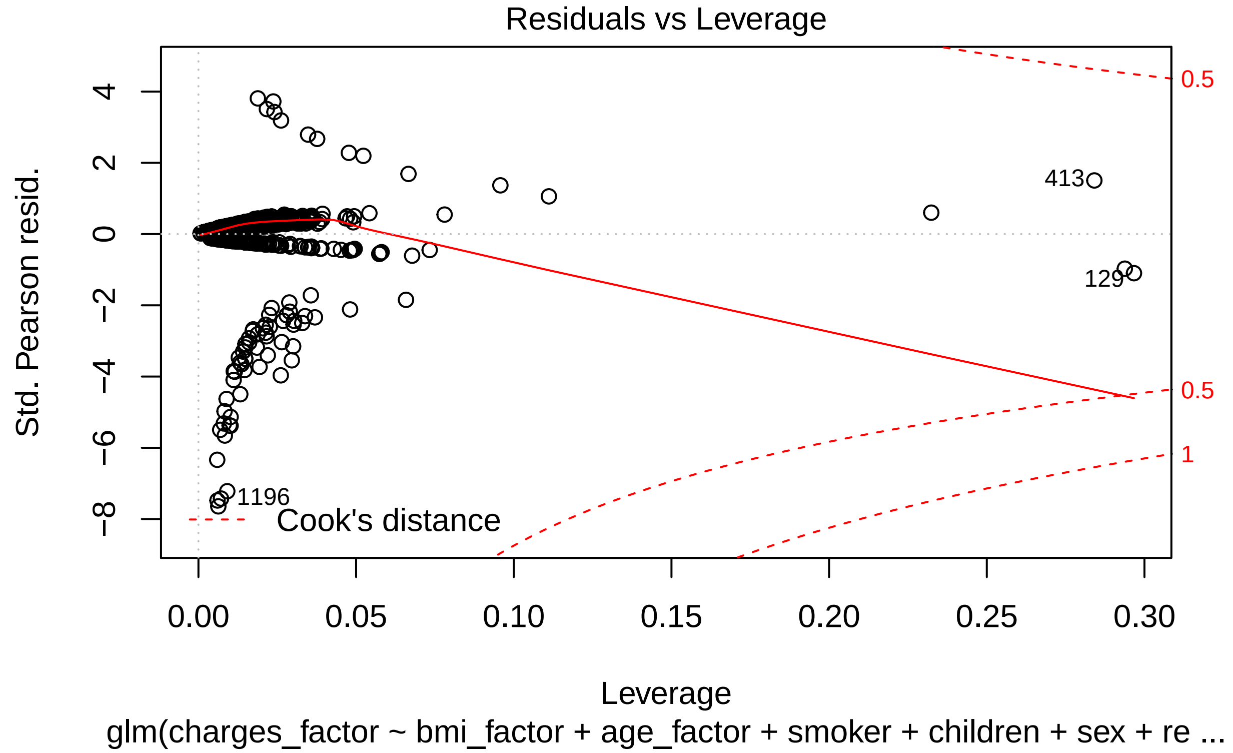
```
##
## Coefficients:
##                       Estimate Std. Error z value Pr(>|z|)    
## (Intercept)             5.3289     0.7712   6.910 4.83e-12 ***
## bmi_factorobese_1      -1.3263     0.4788  -2.770  0.00561 ** 
## bmi_factorobese_2      -1.5547     0.5196  -2.992  0.00277 ** 
## bmi_factorobese_3      -0.9234     0.6802  -1.358  0.17460    
## bmi_factoroverweight   -1.0735     0.4857  -2.210  0.02711 *  
## bmi_factorunderweight   2.2005     1.1518   1.910  0.05608 .  
## age_factor20s          -1.6764     0.6081  -2.757  0.00584 ** 
## age_factor30s          -1.6330     0.6246  -2.615  0.00893 ** 
## age_factor40s          -1.2796     0.6123  -2.090  0.03663 *  
## age_factor50s          -1.9955     0.6104  -3.269  0.00108 ** 
## age_factor60s          -1.9176     0.6859  -2.796  0.00518 ** 
## smokeryes              -5.9458     0.4111 -14.462  < 2e-16 ***
## children               -0.1954     0.1043  -1.874  0.06092 .  
## sexmale                 0.0905     0.2540   0.356  0.72164    
## regionnorthwest         0.2752     0.3731   0.738  0.46081    
## regionsoutheast        -0.1448     0.3440  -0.421  0.67375    
## regionsouthwest         0.6207     0.3825   1.623  0.10465    
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1128.59  on 1002  degrees of freedom
## Residual deviance:  475.29  on  986  degrees of freedom
## AIC: 509.29
##
## Number of Fisher Scoring iterations: 6
```

```r
plot(lgall)
```


Residuals vs Fitted
Residuals
2
1
0
−1
−2
−3
−4
−2
0
2
4
6
8
Predicted values
glm(charges_factor ~ bmi_factor + age_factor + smoker + children + sex + re ...
4
1028


Normal Q−Q
Std. deviance resid.
2
1
0
−1
−2
−3
−3
−2
−1
0
1
2
3
Theoretical Quantiles
glm(charges_factor ~ bmi_factor + age_factor + smoker + children + sex + re ...





- Prediction:
  - 25% ability to prdict high charges, and 92% ability to predict low charges?

```
predicttrain = predict(lgall, type = "response")
tapply(predicttrain, hc1train$charges_factor, mean)
```

```
##      high      low
## 0.2510013 0.9162216
```

- Confusion matrix on training set with .5 threshhold:
  - 93.02% accuracy of predicting high health charges

```
table(hc1train$charges_factor, predicttrain > .5)
```

```
##
##         FALSE TRUE
##    high   193   58
##    low     12  740
```

```
(193 + 740) / ( 193 + 58 + 12 + 740)
```

```
## [1] 0.9302094
```

- Confusion matrix on testing set with .5 threshhold:
  - 91.94% accuracy of predicting high health charges

```
predicttest = predict(lgall, type = "response", newdata = hc1test)
table(hc1test$charges_factor, predicttest > .5)
```

```
##
##         FALSE TRUE
##    high    62   22
##    low      5  246
```

```
(62 + 246) / ( 62 + 22 + 5 + 246)
```

```
## [1] 0.919403
```

CLUSTERING:

- Heigharchial clustering with dendrogram of all variables aside from charge.

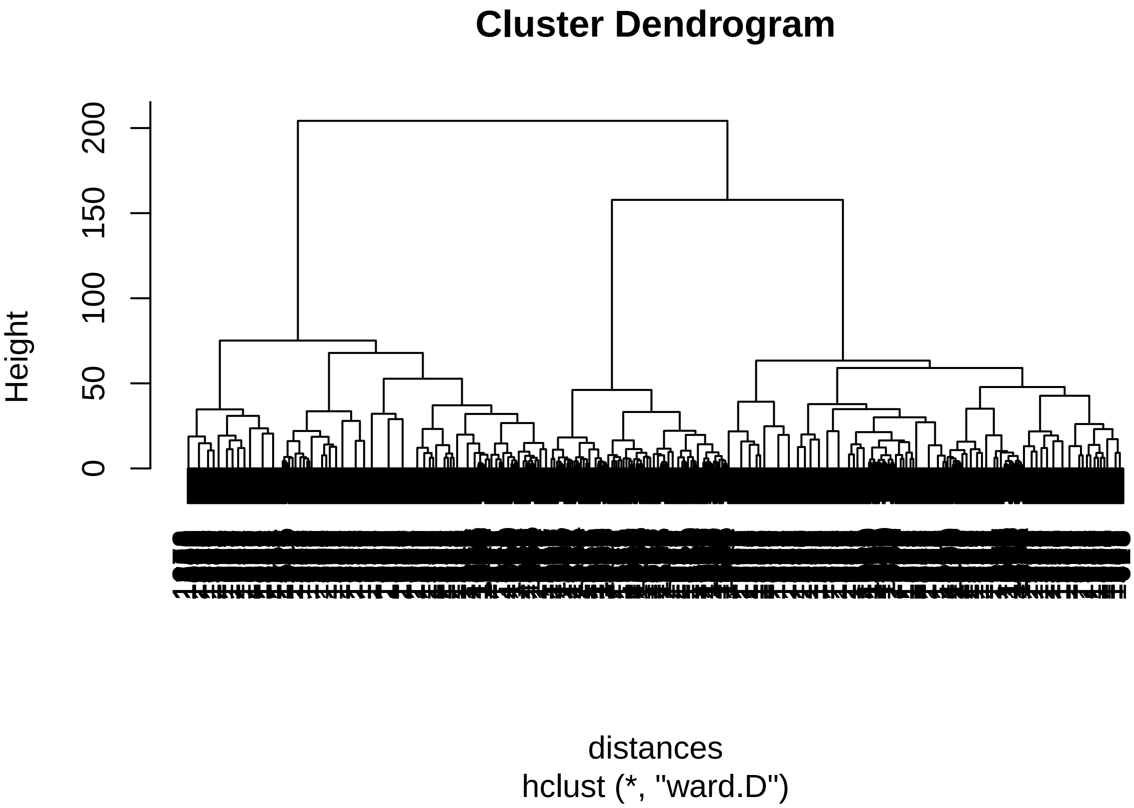
```
colnames(hc2)
```

```
##  [1] "charges_factor_high"        "charges_factor_low"
##  [3] "bmi_factor_overweight"      "bmi_factor_obese_1"
##  [5] "bmi_factor_healthy_weight" "bmi_factor_obese_2"
##  [7] "bmi_factor_obese_3"         "bmi_factor_underweight"
##  [9] "children_0"                 "children_1"
## [11] "children_3"                 "children_2"
## [13] "children_5"                 "children_4"
## [15] "smoker_yes"                 "smoker_no"
## [17] "region_southwest"           "region_southeast"
## [19] "region_northwest"           "region_northeast"
## [21] "sex_female"                 "sex_male"
```

```
distances = dist(hc2[c(-1, -2)], method = "euclidian")
cluster1 = hclust(distances, method = "ward")
```

```
## The "ward" method has been renamed to "ward.D"; note new "ward.D2"
```

```
plot(cluster1)
```



*Eight clusters: groups with high charges: + Group 1 was a predictor of high charges at 91.7% + The percentage of high charges and the percent of smokers within each cluster were equal across all clusters. + Smoking and high charges were most powerful in the clustering algorithm.

```
clustergroups = cutree(cluster1, k = 8)
str(clustergroups)
```

```
##  int [1:1338] 1 2 2 3 3 4 4 3 3 5 ...
```

```
highv <- tapply(hc2$charges_factor_high, clustergroups, mean)
highv <- as.vector(highv)
smokev <- tapply(hc2$smoker_yes, clustergroups, mean)
smokev <- as.vector(highv)
nev <- tapply(hc2$region_northeast, clustergroups, mean)
nev <- as.vector(nev)
nwv <- tapply(hc2$region_northwest, clustergroups, mean)
nwv <- as.vector(nwv)
swv <- tapply(hc2$region_southwest, clustergroups, mean)
swv <- as.vector(swv)
sev <- tapply(hc2$region_southeast, clustergroups, mean)
sev <- as.vector(sev)
sexfv <- tapply(hc2$sex_female, clustergroups, mean)
sexfv <- as.vector(sexfv)
ch0v <- tapply(hc2$children_0, clustergroups, mean)
ch0v <- as.vector(ch0v)
ch1v <- tapply(hc2$children_1, clustergroups, mean)
ch1v <- as.vector(ch1v)
ch2v <- tapply(hc2$children_2, clustergroups, mean)
ch2v <- as.vector(ch2v)
```

```
ch3v <- tapply(hc2$children_3, clustergroups, mean)
ch3v <- as.vector(ch3v)
ch4v <- tapply(hc2$children_4, clustergroups, mean)
ch4v <- as.vector(ch4v)
ch5v <- tapply(hc2$children_5, clustergroups, mean)
ch5v <- as.vector(ch5v)
bmiuv <- tapply(hc2$bmi_factor_underweight, clustergroups, mean)
bmiuv <- as.vector(bmiuv)
bmihv <- tapply(hc2$bmi_factor_healthy_weight, clustergroups, mean)
bmihv <- as.vector(bmihv)
bmiov <- tapply(hc2$bmi_factor_overweight, clustergroups, mean)
bmiov <- as.vector(bmiov)
bmio1v <- tapply(hc2$bmi_factor_obese_1, clustergroups, mean)
bmio1v <- as.vector(bmio1v)
bmio2v <- tapply(hc2$bmi_factor_obese_2, clustergroups, mean)
bmio2v <- as.vector(bmio2v)
bmio3v <- tapply(hc2$bmi_factor_obese_3, clustergroups, mean)
bmio3v <- as.vector(bmio3v)

clusterframe <- cbind(highv, smokev, sexfv, bmiuv, bmihv, bmiov, bmio1v, bmio2v, bmio3v, ch0v, ch1v, ch

View(clusterframe)
```

“‘ ###CLUSTERING WITHOUT SMOKERS DATA *more symmetrical dendrogram; high charges distributed relatively evenly between clusters.

```
colnames(hc2)
```

```
##  [1] "charges_factor_high"       "charges_factor_low"
##  [3] "bmi_factor_overweight"     "bmi_factor_obese_1"
##  [5] "bmi_factor_healthy_weight" "bmi_factor_obese_2"
##  [7] "bmi_factor_obese_3"        "bmi_factor_underweight"
##  [9] "children_0"                "children_1"
## [11] "children_3"                "children_2"
## [13] "children_5"                "children_4"
## [15] "smoker_yes"                "smoker_no"
## [17] "region_southwest"          "region_southeast"
## [19] "region_northwest"          "region_northeast"
## [21] "sex_female"                "sex_male"
```

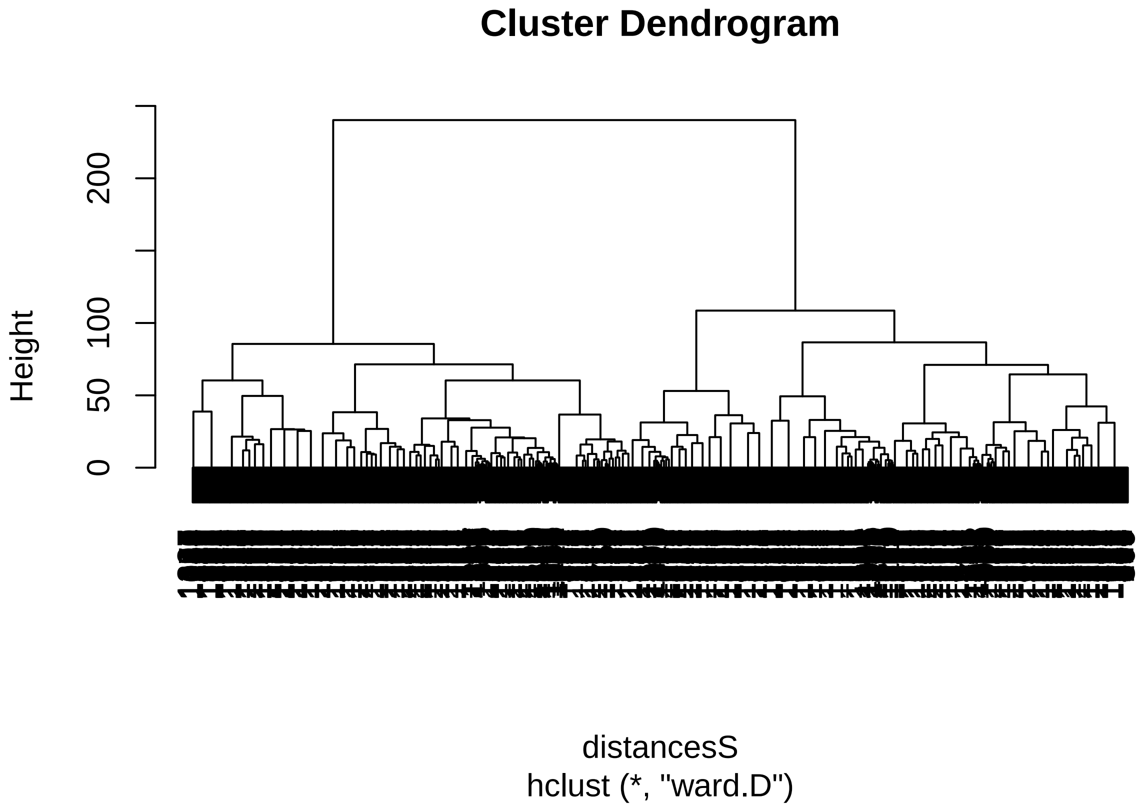
```
hc3 <- hc2[ , c(1:14, 17:22)]
colnames(hc3)
```

```
##  [1] "charges_factor_high"       "charges_factor_low"
##  [3] "bmi_factor_overweight"     "bmi_factor_obese_1"
##  [5] "bmi_factor_healthy_weight" "bmi_factor_obese_2"
##  [7] "bmi_factor_obese_3"        "bmi_factor_underweight"
##  [9] "children_0"                "children_1"
## [11] "children_3"                "children_2"
## [13] "children_5"                "children_4"
## [15] "region_southwest"          "region_southeast"
## [17] "region_northwest"          "region_northeast"
## [19] "sex_female"                "sex_male"
```

```
distancesS = dist(hc3[c(-1, -2)], method = "euclidian")
clustersS = hclust(distancesS, method = "ward")
```

```
## The "ward" method has been renamed to "ward.D"; note new "ward.D2"
```

```
plot(clusterS)
```



```
clustergroupsS = cutree(clusterS, k = 8)
str(clustergroupsS)
```

```
##  int [1:1338] 1 2 3 4 4 2 5 6 7 1 ...
```

```
highvS <- tapply(hc2$charges_factor_high, clustergroupsS, mean)
highvS <- as.vector(highvS)
smokevS <- tapply(hc2$smoker_yes, clustergroupsS, mean)
smokevS <- as.vector(highvS)
nevS <- tapply(hc2$region_northeast, clustergroupsS, mean)
nevS <- as.vector(nevS)
nwvS <- tapply(hc2$region_northwest, clustergroupsS, mean)
nwvS <- as.vector(nwvS)
swvS <- tapply(hc2$region_southwest, clustergroupsS, mean)
swvS <- as.vector(swvS)
sevS <- tapply(hc2$region_southeast, clustergroupsS, mean)
sevS <- as.vector(sevS)
sexfvS <- tapply(hc2$sex_female, clustergroupsS, mean)
sexfvS <- as.vector(sexfvS)
ch0vS <- tapply(hc2$children_0, clustergroupsS, mean)
ch0vS <- as.vector(ch0vS)
ch1vS <- tapply(hc2$children_1, clustergroupsS, mean)
ch1vS <- as.vector(ch1vS)
ch2vS <- tapply(hc2$children_2, clustergroupsS, mean)
ch2vS <- as.vector(ch2vS)
ch3vS <- tapply(hc2$children_3, clustergroupsS, mean)
```

```
ch3vS <- as.vector(ch3vS)
ch4vS <- tapply(hc2$children_4, clustergroupsS, mean)
ch4vS <- as.vector(ch4vS)
ch5vS <- tapply(hc2$children_5, clustergroupsS, mean)
ch5vS <- as.vector(ch5vS)
bmiuvS <- tapply(hc2$bmi_factor_underweight, clustergroupsS, mean)
bmiuvS <- as.vector(bmiuvS)
bmihvS <- tapply(hc2$bmi_factor_healthy_weight, clustergroupsS, mean)
bmihvS <- as.vector(bmihvS)
bmiovS <- tapply(hc2$bmi_factor_overweight, clustergroupsS, mean)
bmiovS <- as.vector(bmiovS)
bmio1vS <- tapply(hc2$bmi_factor_obese_1, clustergroupsS, mean)
bmio1vS <- as.vector(bmio1vS)
bmio2vS <- tapply(hc2$bmi_factor_obese_2, clustergroupsS, mean)
bmio2vS <- as.vector(bmio2vS)
bmio3vS <- tapply(hc2$bmi_factor_obese_3, clustergroupsS, mean)
bmio3vS <- as.vector(bmio3vS)

clusterframeS <- cbind(highvS, smokevS, sexfvS, bmiuvS, bmihvS, bmiovS, bmio1vS, bmio2vS, bmio3vS, ch0vS

View(clusterframeS)
```