

Predicting Health Insurance Costs

Zara Nip and Nehal Linganur

Dependencies

```
library(NbClust)
library(mclust)
```

```
## Warning: package 'mclust' was built under R version 4.2.3
```

```
## Package 'mclust' version 6.0.0
## Type 'citation("mclust")' for citing this R package in publications.
```

```
library(flexclust)
```

```
## Warning: package 'flexclust' was built under R version 4.2.3
```

```
## Loading required package: grid
```

```
## Loading required package: lattice
```

```
## Loading required package: modeltools
```

```
## Loading required package: stats4
```

```
library(cluster)
library(fpc)
```

```
## Warning: package 'fpc' was built under R version 4.2.3
```

Data Preparation

Import data and remove missing rows.

```
insurance <- read.csv("insurance.csv")
insurance <- na.omit(insurance)
insurance$age <- as.numeric(insurance$age)
insurance$bmi <- as.numeric(insurance$bmi)
```

```
####Linear Regression Models
```

```
mod<-lm(charges~., data=insurance)
summary(mod)
```

```
##
## Call:
## lm(formula = charges ~ ., data = insurance)
##
## 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 ***
## age             256.9       11.9   21.587 < 2e-16 ***
## sexmale        -131.3      332.9   -0.394 0.693348
## bmi             339.2       28.6   11.860 < 2e-16 ***
## children        475.5      137.8    3.451 0.000577 ***
## smokeryes      23848.5     413.1   57.723 < 2e-16 ***
## 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
```

Using only statistically significant factors: age, bmi, children, smoker (yes).

```
sigFac<-lm(charges~age + bmi + children + smoker, data=insurance)
summary(sigFac)
```

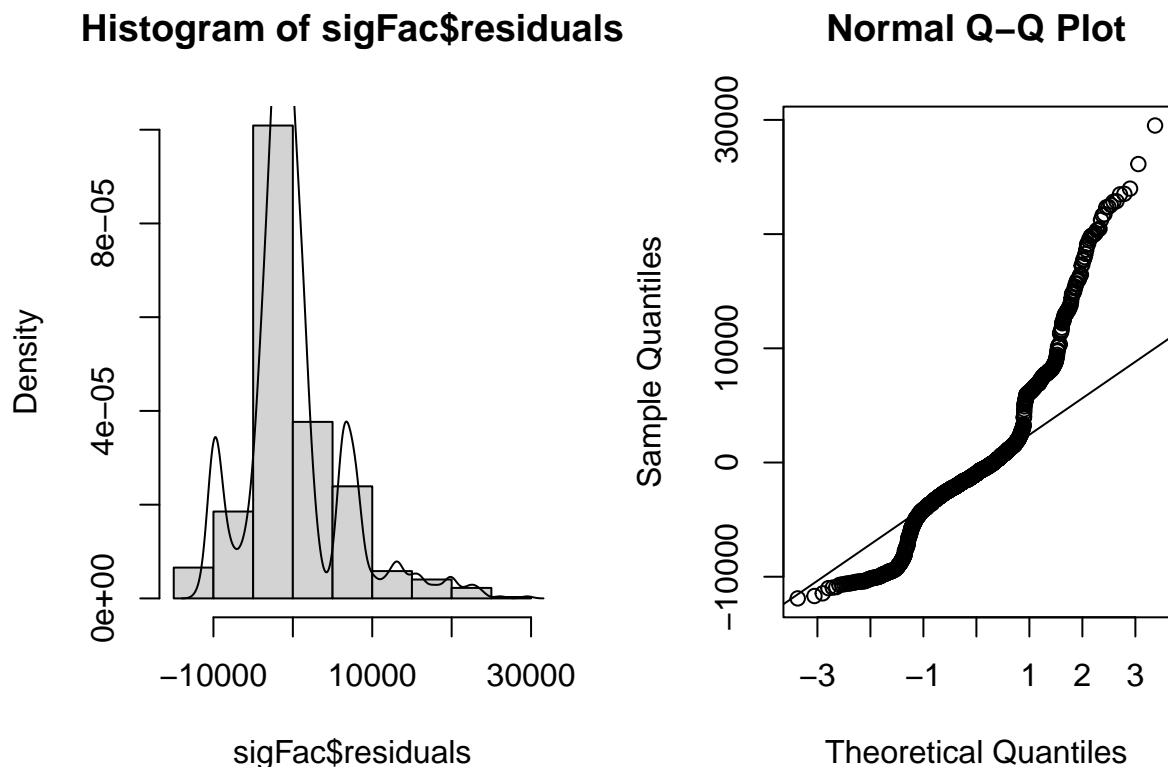
```
##
## Call:
## lm(formula = charges ~ age + bmi + children + smoker, data = insurance)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -11897.9  -2920.8   -986.6   1392.2  29509.6
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -12102.77     941.98  -12.848 < 2e-16 ***
## age           257.85      11.90   21.675 < 2e-16 ***
## bmi           321.85      27.38   11.756 < 2e-16 ***
## children      473.50     137.79    3.436 0.000608 ***
## smokeryes     23811.40     411.22   57.904 < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 6068 on 1333 degrees of freedom
## Multiple R-squared:  0.7497, Adjusted R-squared:  0.7489
## F-statistic: 998.1 on 4 and 1333 DF,  p-value: < 2.2e-16
```

Residual Plots

```
par(mfrow=c(1,2))
hist(sigFac$residuals, prob = TRUE)
lines(density(sigFac$residuals))

qqnorm(y=sigFac$residuals)
qqline(y=sigFac$residuals, datax = FALSE)
```



Testing Correlation With Clustering Our previous model stated that age, bmi, children, smoker (yes) were all significant factors. We are now going to perform a clustering model to see if this still holds true.

```
seed.val<-1234

insurance.scaled <- scale(insurance[1] + insurance[3])
for(i in 1:ncol(insurance.scaled)){
  print(max(insurance.scaled[ , i]))
}
```

```
## [1] 2.336639
```

```
RNGversion("4.1.2")
set.seed(seed.val)
```

```
clusterInsurance <- kmeans(insurance.scaled, 2, nstart = 25)
```

```
clusterInsurance$size
```

```
## [1] 657 681
```

```
clusterInsurance$tot.withinss
```

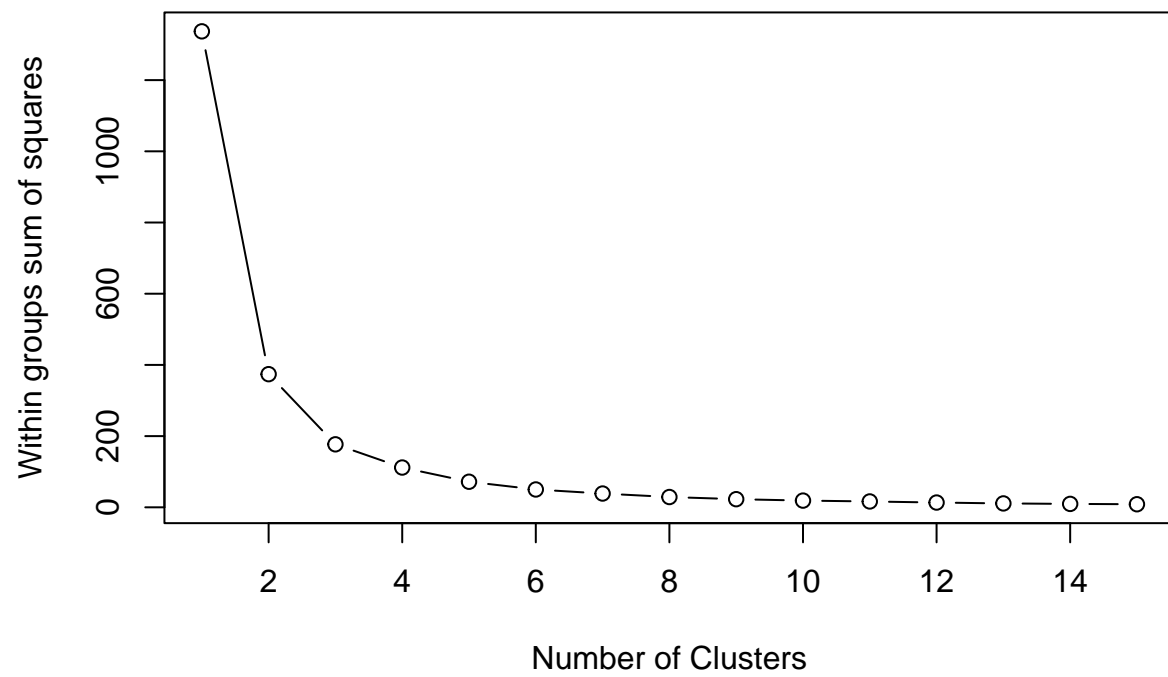
```
## [1] 374.0375
```

```
clusterInsurance$betweenss
```

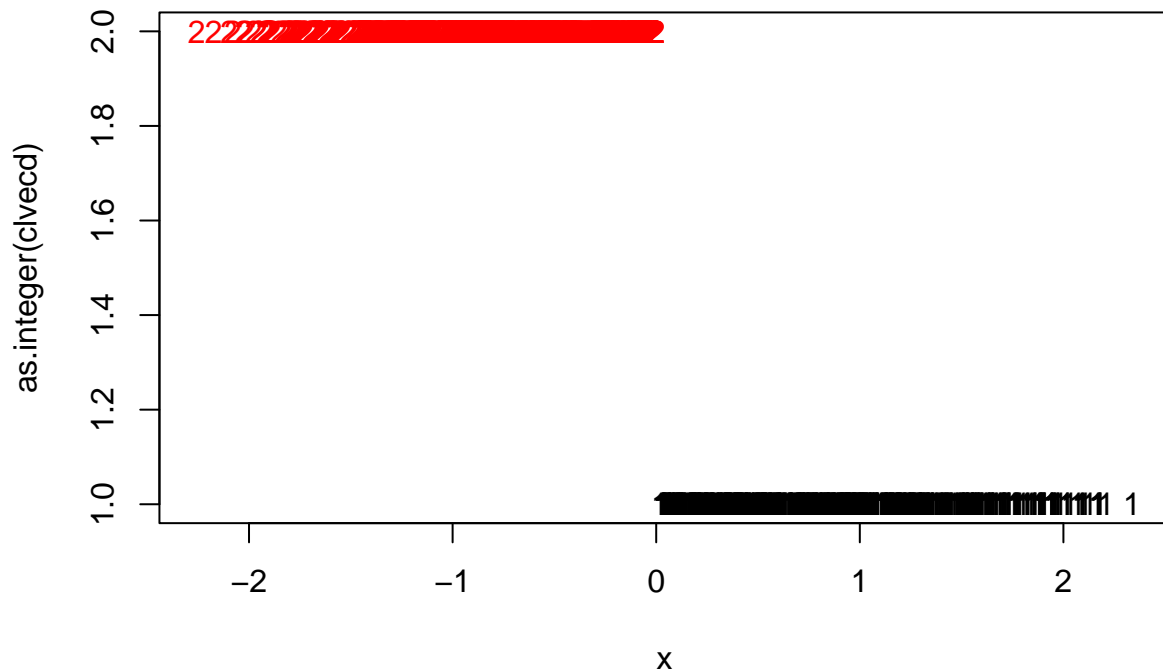
```
## [1] 962.9625
```

```
wssplot <- function(data, nc=15, seed=1234){
  wss <- (nrow(data)-1)*sum(apply(data,2,var))
  for (i in 2:nc){
    set.seed(seed)
    wss[i] <- sum(kmeans(data, centers=i)$withinss)}
  plot(1:nc, wss, type="b", xlab="Number of Clusters",
       ylab="Within groups sum of squares")
}

wssplot(insurance.scaled)
```



```
plotcluster(insurance.scaled, clusterInsurance$cluster)
```



Commented out because of knitting issues. Returned $k = 2$ as optimal amount of clusters. `#nc <- NbClust(insurance.scaled, min.nc=2, max.nc=15, method="kmeans") #table(nc$Best.n[1,])`

Optimal number of clusters using PAM

```
dist.mat<-daisy(insurance.scaled, metric="euclidean")
pk <- pamk(dist.mat, krange=2:15, usepam=TRUE, diss=TRUE)
pk$nc
```

```
## [1] 2
```

```
fit.pam = pam(dist.mat,2)
plot(fit.pam)
```

Silhouette plot of pam(x = dist.mat, k = 2)

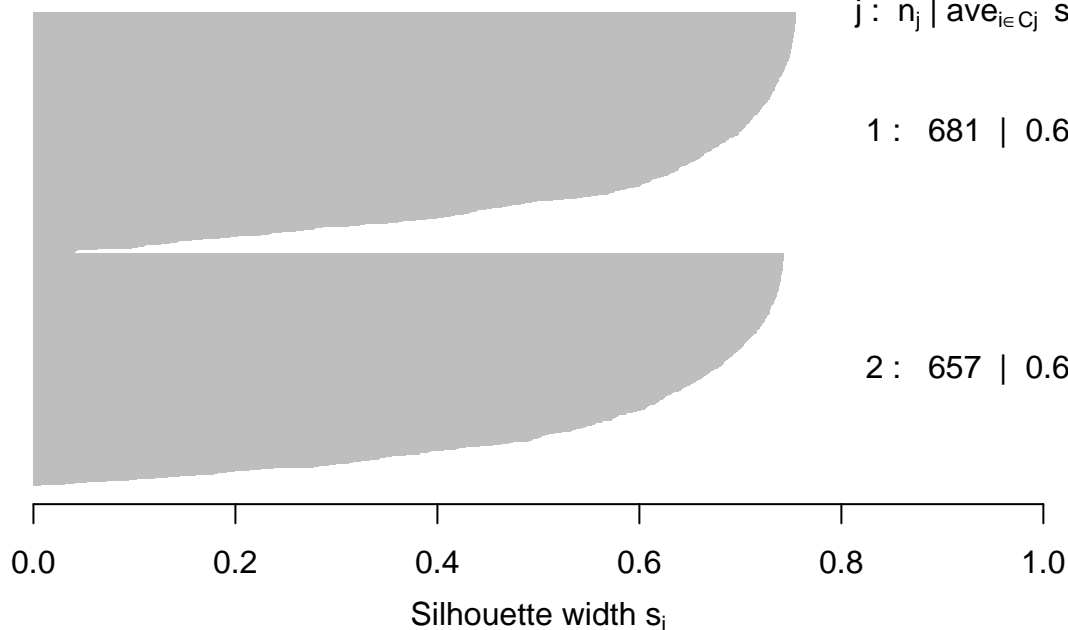
n = 1338

2 clusters C_j

$j: n_j \mid \text{ave}_{i \in C_j} s_i$

1 : 681 | 0.61

2 : 657 | 0.60



```
jpeg("MYPLOT.jpg")
plot(fit.pam)
dev.off()
```

```
## pdf
## 2
```

Chi-Squared Test

```
cont.table <- table(insurance$charges, clusterInsurance$cluster)
print(chisq.test(cont.table))
```

```
## Warning in chisq.test(cont.table): Chi-squared approximation may be incorrect
```

```
##
## Pearson's Chi-squared test
##
## data: cont.table
## X-squared = 1338, df = 1336, p-value = 0.4794
```

How different are the clusters?

```
randIndex(cont.table)
```

```
##          ARI  
## 2.237909e-06
```