Young People Survey

Statistical Models

Dario Fabiani, Beatrice Marsili

Abstract

The purpose of this project is to discover if we can predict spending habits, and in case other preferences, based on possible types of personalities. We use the Young People Survey dataset (available at Kaggle).
This dataset includes 1010 observations for 150 different varibles. It has been built over a questionary answered by Slovakian people aged between 15 and 30. These people were asked to respond questions regarding their Music and Movies preferences, their Hobbies and Phobias, their Spending and Health Habits, some of their Demographic characteristics and their thoughts about life.
All variables in the survey are Likert Scales, which mean that the values should be considered as ordered factors, given that we cannot consider the various levels as distant as two integer could be. Therefore we decided to transform the variables into the correct type (ordered factors).

57 over 150 questions regard personality traits and views about life, thus we decided to find some patterns between these features, in order to utilize them to predicte spending habits and other preferences expressed in the survey.
We're in the scenario of *unsupervised learning*, as we have a set of $X_1, X_2, \ldots, X_{57}$ features measured over 1010 observations, and we want to discover unknown subgroups within these features. We first had a look to these 57 questions and tried to understand which characteristics of personality were observed by each variable, simply using the Myers-Briggs indicator, often used in psichology. Obviously this method was very rough as it is subject to our personal thoughts and could misleading the results.

We then decided to go for a cluster analysis, a set of techniques that help discover hidden structures on the basis of a dataset.
To perform the Cluster analysis, we chose a package called "ClusOfVar", which allow to construct a synthetic variable based on the combination of the original ones; the technique applied by this package computes this variable via a PCA performed in each cluster and then retains the first principal component as the synthetic variable.

In order to identify the clusters we used the "stability" method in the same package, which is based on boostraps. Its output is not really helpful, thus we opted for a visual interpretation and selected 5 clusters.
Afterwards, we assigned the scores for the firsts principal components of PCAMIX applied to the K clusters to the n observations.
Eventually, we performed a *Ordered Logistic Regression* to answer our research question.

Data Preparation

Our data was already technically correct. We renamed columns to avoid problems in handling them.

We had to decide how to deal with NAs, that were the only special values in our dataset.

```
omit <- na.omit(df.responses)
dim(df.responses)[1] - dim(omit)[1]
```

```
## [1] 324
```

Having 324 rows with NA values we decided that it wasn't a good idea to omit them, we then decided to fill those missing values with the median of the column. We choose to use the median as the variables are of ordered categorical type and choosing the mean would have changed the levels of our parameters.
Looking at the data referred to personality questions we preferred to substitute missing values with *"3"*, as using the median as for the rest of the dataset would have affected correlation. To handle properly the dataset we divided it into subsets,according to the topic they deal with.

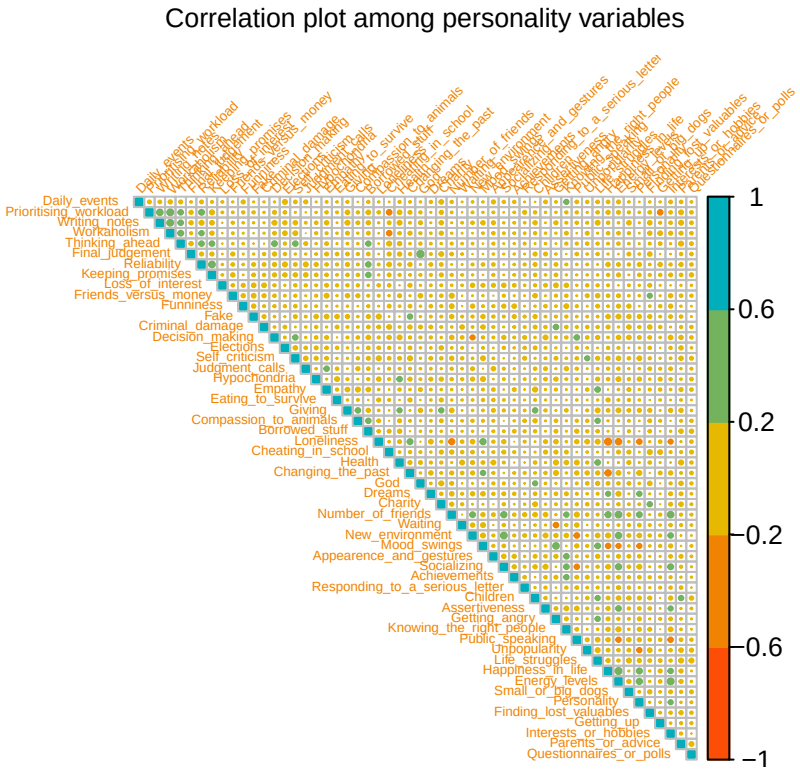
```
pref <- colnames(df.responses[-c(77:133, 141:155, 74, 75)])
for(name in pref){
  df.responses[is.na(df.responses[,name]), name] <- median(df.responses[,name],
                                                           na.rm = TRUE)
}

pers <- colnames(df.responses[77:133])
pers <- pers[-c(108, 109, 133)]
for(name in pers){
  df.responses[is.na(df.responses[,name]), name] <- 3
}

df.music <- data.frame(df.responses[0:19])
df.movies <- data.frame(df.responses[20:31])
df.hobbies <- data.frame(df.responses[32:63])
df.phobias <- data.frame(df.responses[64:73])
df.health <- data.frame(df.responses[74:76])
df.personality <- data.frame(df.responses[77:133])
df.spending <- data.frame(df.responses[134:140])

df.qualitative <- df.responses
for (name in colnames(df.qualitative[-c(141:155, 74,75, 108, 109, 133)])){
  df.qualitative[,name] <- factor(df.qualitative[,name], levels = c("1", "2", "3", "4", "5"), ordered =
}
```

Our aim here is to find possible personality types, we reach the scope using a clustering of the variables. Before doing it, we wanted to visualize possible existing correlation inside this subset; we then performed a correlation matrix and a correlation plot, that confirmed that there are some (positive and negative) patterns among the responses on such variables. This correlation plot has been done to see if there are some anti-correlated variables to be aware of them in the next step of clustering.



Hierarchical Clustering

Even if we had some ideas about the number of clusters (on the basis of the rough analysis done with Myers-Briggs indicator) we decided to use a **Hierarchical Cluster**, without definig *a priori* the number of clusters to be obtained.

The main issue with this approach was to synthesize in a reasonable way the clusters obtained. Once the variables are clustered into groups such that attributes in each group reflect the same aspect, we opted for a synthetic variable realized with the package "ClustOfVar". The dissimilarity among the observations is measured by the correlation ratio for qualitative variables, while the dissimilarity among the cluster is measured taking into account the lost of homogeneity observed when the two clusters are merged.

The quantitative central synthetic variable of each cluster is the first principal component of PCAMIX applied to all the variables in the cluster.

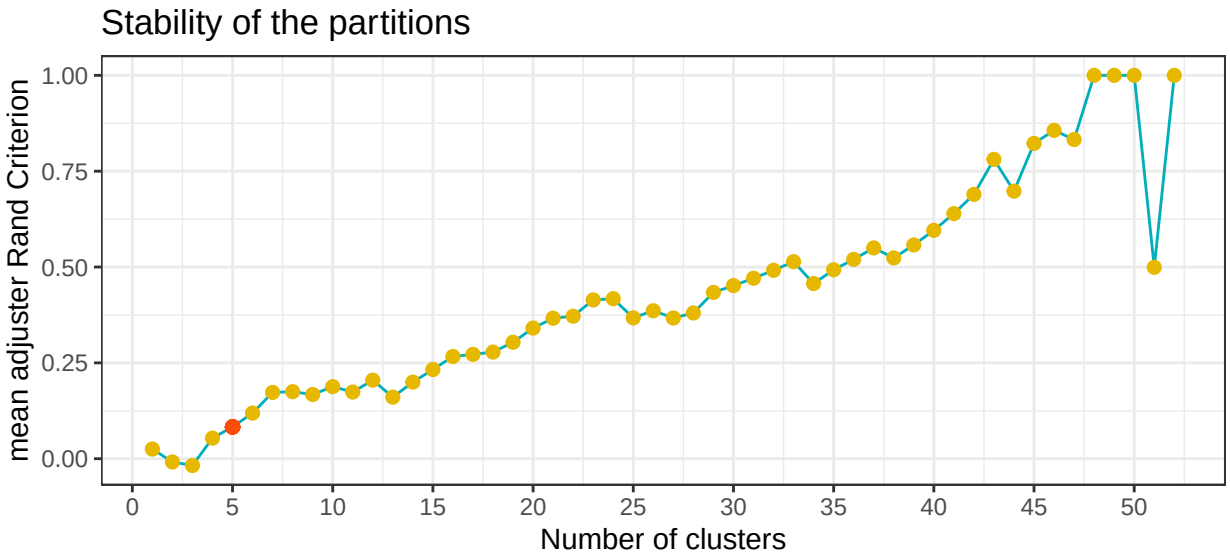
```
library(ClustOfVar)
quali.pers <- df.qualitative[77:133]
hc.clust <- hclustvar(X.quali=quali.pers[-c(32, 33, 57)])
```

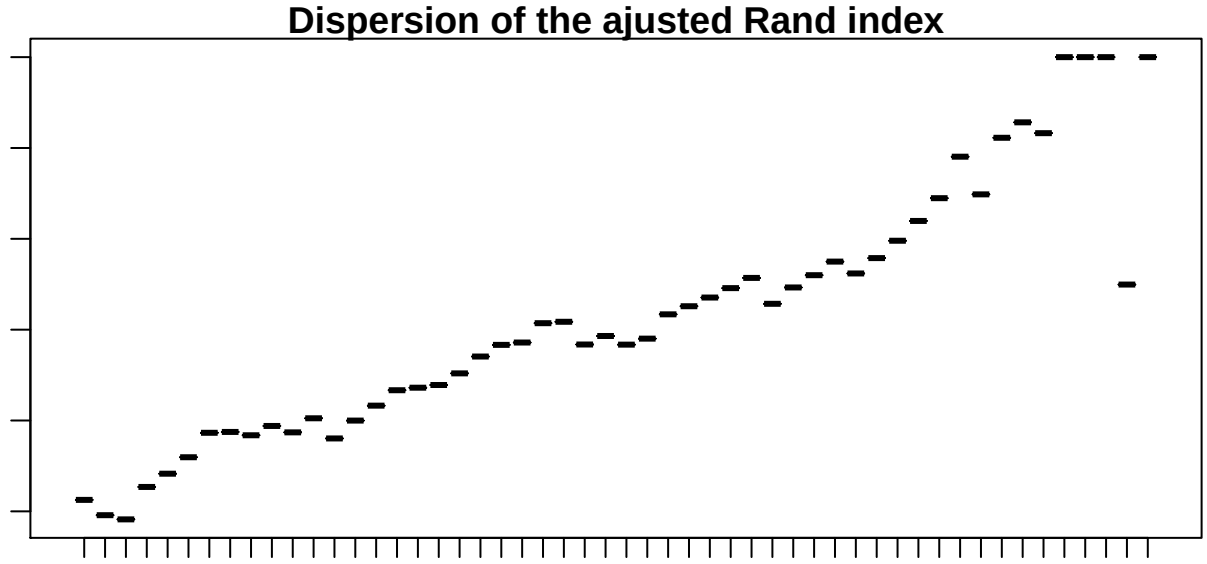
In order to get the optimal number of cluster we tried the *stability* function in the same package, which essentially makes B boostrap samples of the n observations, and B dendrograms with the function hclustvar. The clusters of these B dendrograms are compared with the clusters of the initial hierarchy using the corrected Rand index. The stability of a tree is evaluated by the mean of the B *adjusted Rand indices*. The *Rand index* itself takes into account the similarities and dissimilarities between the different clusters and offers an informative measure that helps deciding if it is the case to merge two clusters or not. The adjusted version of this index, instead, takes into account the expected pairwise similarity among the different clusters.

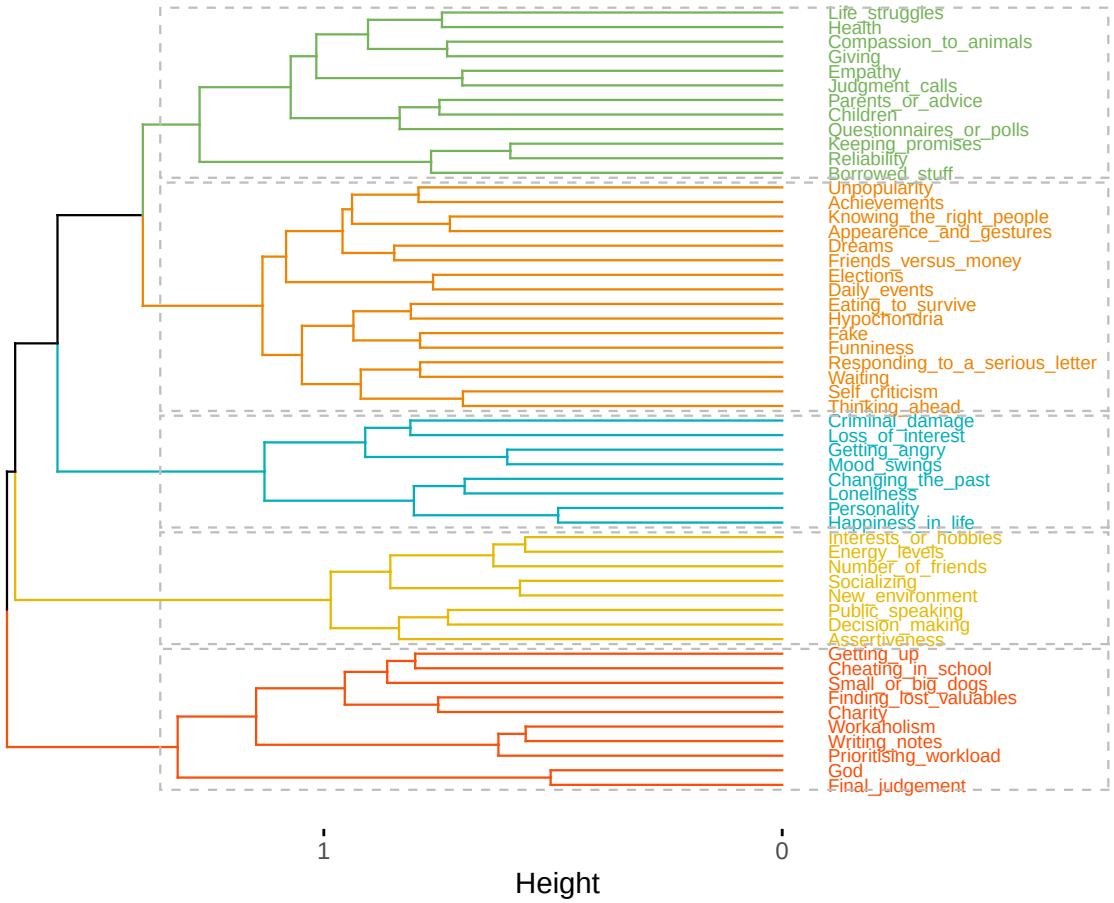
Please note that the following chunk has an high computational cost

```
set.seed(5)
stab <- stability(hc.clust,B=1, graph = F)
```

Looking at the below plots, that do not seem to give a clear response, we decided to select 5 clusters, given that it seems to be a reasonable choice and that it is the same number we thought about when exploring with the Myers-Briggs indicator.







Next we added 5 columns to the dataset, one for each cluster, and filled the values with *hc.cut$scores* that is the matrix (1010x5) of the scores for the n observations on the first principal components of PCAMIX applied to the 5 clusters. Each column is then a synthetic variable of a cluster.

```
hc.cut <-cutreevar(hc.clust, 5)

df.qualitative$Cluster_1 <- hc.cut$scores[,1]
df.qualitative$Cluster_2 <- hc.cut$scores[,2]
df.qualitative$Cluster_3 <- hc.cut$scores[,3]
df.qualitative$Cluster_4 <- hc.cut$scores[,4]
df.qualitative$Cluster_5 <- hc.cut$scores[,5]
```

Regression

Now that we have our clusters that try to describe different aspects of personality we want to use regression methods to respond to our main research constraint: *"It is possible to try predicting spending habits on the basis of different types of personalities?"*.

We remind that variables are ordered: level 1 corresponds to "Strongly disagree" and level 5 to "Strongly agree"; we decided to consider the middle level, the third one, as a neutral response. We divided the dataset into training and testing subset and then started performing *Ordinal Logistic Regression* with the variables

referred to Spending Habits as *responses* and the clusters obtained from the features referred to personality traits as *predictors*.

Here in this file, we put the relative most significant regression we made, the one with the variable *Finances*. The variable *Finances* corresponds to the answers given to the question *"I save all the money I can"*. We also performed a best subset selection, in order to select the "best" model, according to the RSS.

```
library(leaps)
library(glmnet)
library(MASS)
library(ggplot2)
library(gridExtra)
library(factoextra)

set.seed(5)
trainingRows <- sample(1:nrow(df.qualitative), 0.7 * nrow(df.qualitative))
trainingData <- df.qualitative[trainingRows, ]
testData <- df.qualitative[-trainingRows, ]

options(contrasts = c("contr.treatment", "contr.poly"))
ologit.fin <- polr(Finances ~ Cluster_1 + Cluster_2+
                     Cluster_3 + Cluster_4 + Cluster_5, data=trainingData)

predicted.fin <- predict(ologit.fin, testData)
table(testData$Finances, predicted.fin)
```

```
##    predicted.fin
##      1  2  3  4  5
##   1  9  0 28  2  0
##   2  2  0 47  8  0
##   3  3  0 85 23  0
##   4  2  0 46 22  0
##   5  0  0 16 10  0
```

```
mean(as.character(testData$Finances) == as.character(predicted.fin))
```

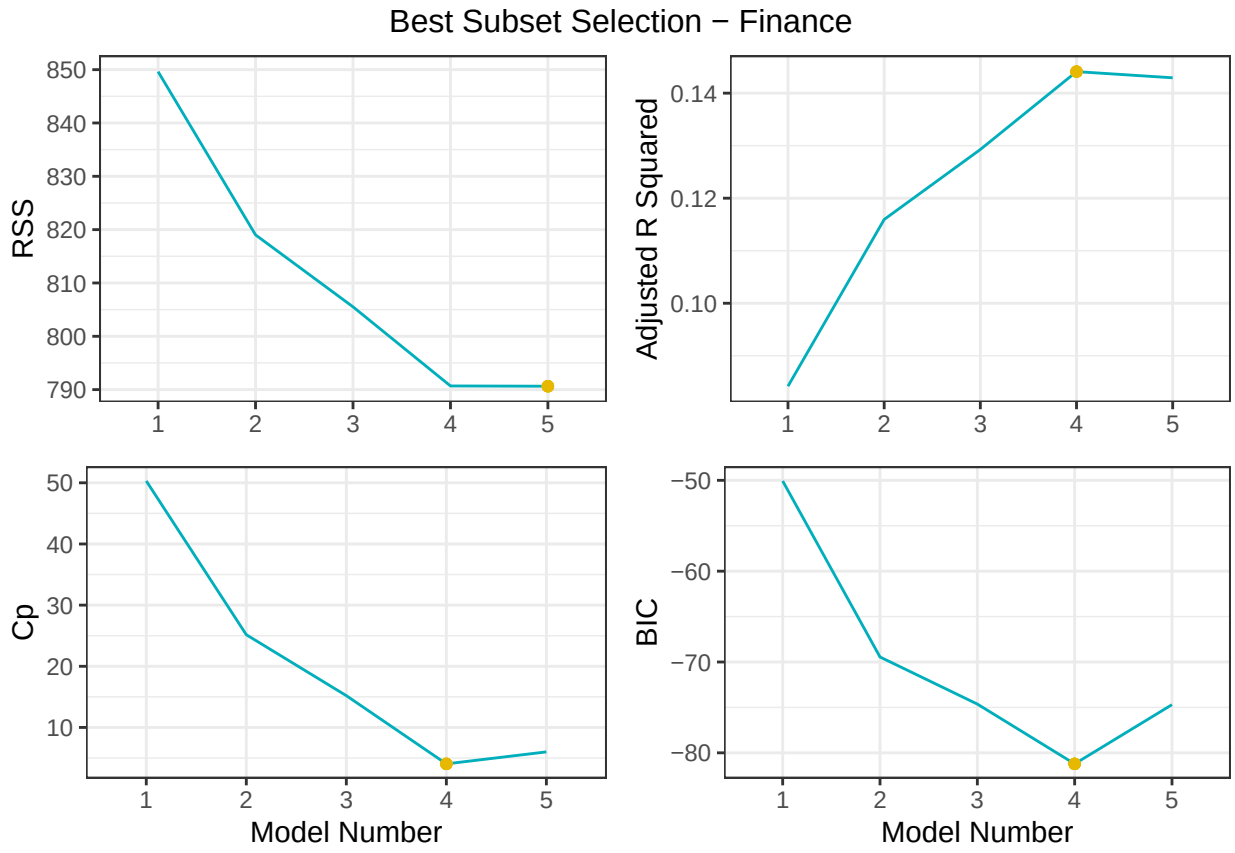
```
## [1] 0.3828383
```

```
regfit.finance <- regsubsets(Finances ~ Cluster_1+Cluster_2+
                             Cluster_3+Cluster_4+
                             Cluster_5, data=trainingData)

finance.sum<- summary(regfit.finance)
finance.sum
```

```
## Subset selection object
## Call: regsubsets.formula(Finances ~ Cluster_1 + Cluster_2 + Cluster_3 +
##     Cluster_4 + Cluster_5, data = trainingData)
## 5 Variables  (and intercept)
##           Forced in Forced out
## Cluster_1     FALSE      FALSE
## Cluster_2     FALSE      FALSE
## Cluster_3     FALSE      FALSE
```

```
## Cluster_4     FALSE      FALSE
## Cluster_5     FALSE      FALSE
## 1 subsets of each size up to 5
## Selection Algorithm: exhaustive
##          Cluster_1 Cluster_2 Cluster_3 Cluster_4 Cluster_5
## 1  ( 1 ) " "       "*"       " "       " "       " "
## 2  ( 1 ) " "       "*"       " "       " "       "*"
## 3  ( 1 ) " "       "*"       "*"       " "       "*"
## 4  ( 1 ) " "       "*"       "*"       "*"       "*"
## 5  ( 1 ) "*"       "*"       "*"       "*"       "*"
```



The model identifies as the "best" the model one with four variables, as stated from both Cp and Adjusted R squared. The BIC is negative since we are in a discrete distribution. We use the model with 4 variables to compute again the *Ordinal Logistic Regression*. However the predict function does not improve, the only improvement we reach in this way is to increase the proportion of variance explained, which still remain close to 0. This is probably to be addressed to the difficulty of predicting a discrete variable on the basis of summarized variables as the ones in the cluster scores are.

```
lm.bestfin <- polr(Finances ~ Cluster_2+Cluster_3+Cluster_4+Cluster_5,data=trainingData,)
predictedbestFin <- predict(lm.bestfin, testData)  # predict the classes directly

table(testData$Finances, predictedbestFin)
```

```
##    predictedbestFin
##      1  2  3  4  5
```

```
##   1  9  0 28  2  0
##   2  2  0 47  8  0
##   3  3  0 86 22  0
##   4  2  0 46 22  0
##   5  0  0 16 10  0
```

```
mean(as.character(testData$Finances) == as.character(predictedbestFin))
```

```
## [1] 0.3861386
```

Since we want to use our cluster to try to predict also the phobias, we make the same thing done with Finance, with the variable Public Speaking in the Phobias subset.

```
options(contrasts = c("contr.treatment", "contr.poly"))
ologit.speak <- polr(Fear_of_public_speaking ~ Cluster_1 + Cluster_2+
                     Cluster_3 + Cluster_4 + Cluster_5, data=trainingData)

predicted.speak <- predict(ologit.speak, testData)
table(testData$Fear_of_public_speaking, predicted.speak)
```

```
##    predicted.speak
##      1  2  3  4  5
##   1 21  6 15  1  2
##   2 12 13 43 10  1
##   3  4 15 68  5  3
##   4  5  5 32  9  6
##   5  1  2 12  5  7
```

```
mean(as.character(testData$Fear_of_public_speaking) == as.character(predicted.speak))
```

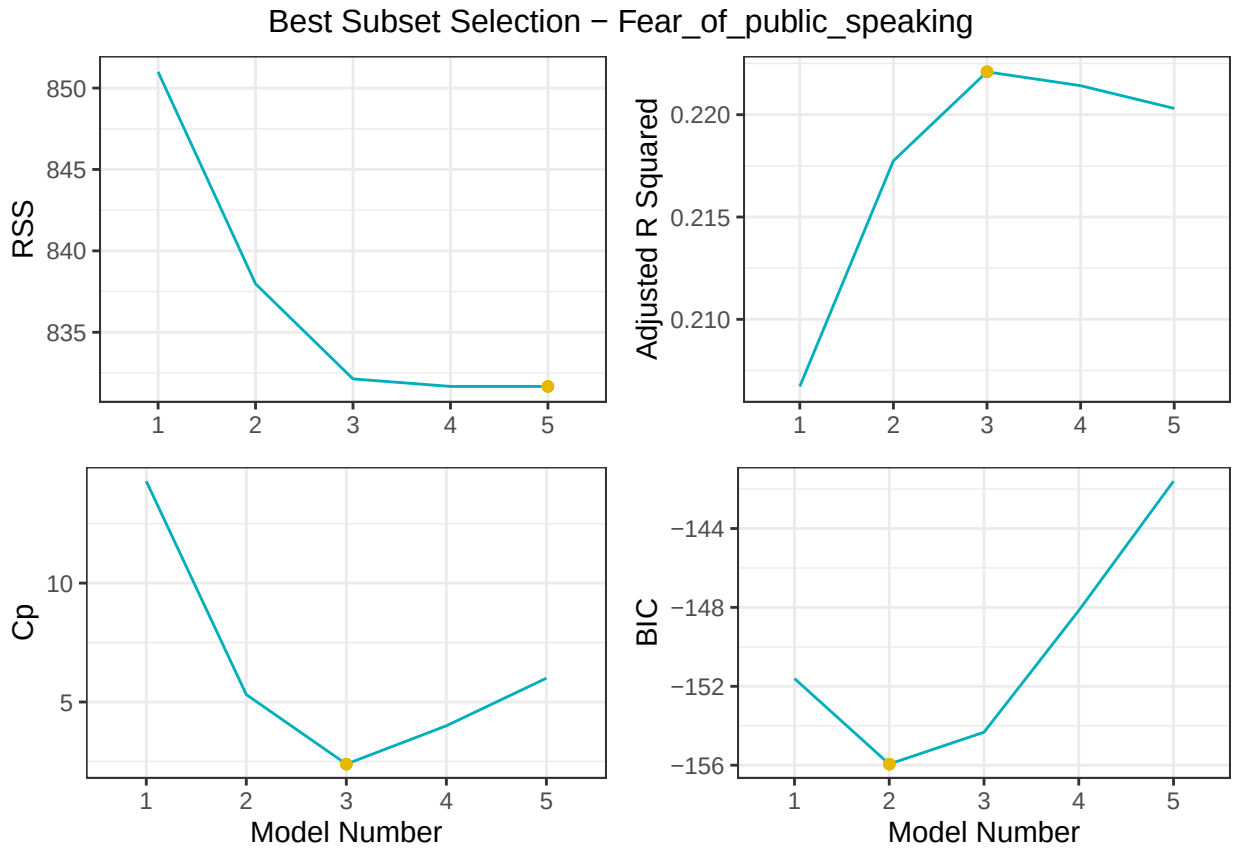
```
## [1] 0.3894389
```

```
regfit.speak<- regsubsets(Fear_of_public_speaking ~ Cluster_1+Cluster_2+
                          Cluster_3+Cluster_4+
                          Cluster_5, data=trainingData)

speak.sum<- summary(regfit.speak)
speak.sum
```

```
## Subset selection object
## Call: regsubsets.formula(Fear_of_public_speaking ~ Cluster_1 + Cluster_2 +
##     Cluster_3 + Cluster_4 + Cluster_5, data = trainingData)
## 5 Variables  (and intercept)
##           Forced in Forced out
## Cluster_1     FALSE      FALSE
## Cluster_2     FALSE      FALSE
## Cluster_3     FALSE      FALSE
## Cluster_4     FALSE      FALSE
## Cluster_5     FALSE      FALSE
## 1 subsets of each size up to 5
## Selection Algorithm: exhaustive
```

```
##          Cluster_1 Cluster_2 Cluster_3 Cluster_4 Cluster_5
## 1  ( 1 ) " "       " "       " "       " "       "*"
## 2  ( 1 ) " "       " "       " "       "*"       "*"
## 3  ( 1 ) " "       "*"       " "       "*"       "*"
## 4  ( 1 ) " "       "*"       "*"       "*"       "*"
## 5  ( 1 ) "*"       "*"       "*"       "*"       "*"
```




```
lm.bestspeak <- polr(Fear_of_public_speaking ~ Cluster_2+Cluster_4+Cluster_5,data=trainingData,)
predictedbestspeak <- predict(lm.bestspeak, testData)

table(testData$Fear_of_public_speaking, predictedbestspeak)
```

```
##    predictedbestspeak
##       1  2  3  4  5
##   1 20  7 16  0  2
##   2 12 13 43 10  1
##   3  4 15 68  5  3
##   4  5  5 34  7  6
##   5  1  2 13  4  7
```

```
mean(as.character(testData$Fear_of_public_speaking) == as.character(predictedbestspeak))
```

```
## [1] 0.379538
```

As previously stated, we selected *Finances* and *Public speaking* as responses as the regressions performed on these variables are the more significative found; however the other ones related to all the variables in the subsets *Spending* and *Phobias* can be fount in the secondary submitted markdown.

Conclusions

"It is possible to predict Spending Habits or other characteristics, such as the Phobias, on the basis of possible type of personalities?" our response, after this analysis is negative. On a conceptual point of view the different types of personalities obtained clustering our original 57 features are quite good, the features in each cluster are clearly correlated in real life (see as an example the fact that the variable referred to believing in God is together with the ones addressable to a Christian way of living, such as the propension to give money in Charity, the belief that at the end of everybody life there will be a final judgment...) but it seems that the summarised variables we're using to predict other aspect of behaviour are irrelevant. The accuracy of our prediction is quite high, but this is to be expected, given that having a scale of 5 a model cannot fail too much; Moreover, looking at the Adjusted R squared values we have evidence that the variance explained from our model is really low, then even though our prediction is quite high, we should be aware that it is not due to the relation between the predictors and the responses, but it may to be related to actual structure of the survey, the Likert Scale.